Preparació de dades

Ramon Sangüesa i Solé

P03/05054/01034

UOC
www.uoc.edu

Índex

Introducció

Com ja s'ha comentat en descriure el procés de mineria de dades, el més normal és que les dades necessàries per a dur a terme un projecte de mineria de dades, un cop seleccionades, hagin de ser modificades i preparades per a poder-les aplicar el mètode de construcció del model escollit per a la tasca de què es tracti.

Vegeu la preparació de les dades en el subapartat 2.3 del mòdul "El procés de descobriment de coneixement a partir de dades" d'aquesta assignatura.

En aquest mòdul didàctic es repassen les tècniques de preparació de dades més freqüents:

a) La transformació de dades; en concret, la normalització i la discretització.

b) El tractament dels valors no observats (*missing values*).

c) La reducció de dimensionalitat de les dades; en particular, la selecció d'atributs.

Es fa un repàs tant de les tècniques que tan sols s'adrecen a la transformació dels valors que cal tractar com a altres tècniques una mica més complexes que requereixen prendre decisions o efectuar hipòtesis sobre els valors no presents o que són clarament erronis.

Les tècniques que es presenten a continuació són prou transversals, independents, fins a cert punt, del tipus de model que cal extreure.

Objectius

Amb l'estudi dels materials didàctics associats a aquest mòdul, l'estudiant assolirà els objectius següents:

1. Adonar-se que, tot i la potència del suport de maquinari i programari que poden tenir els mètodes de mineria de dades, la seva complexitat requereix molt sovint introduir simplificacions per a mantenir el cost computacional dins uns nivells adequats sense comprometre la qualitat final dels models obtinguts.

2. Conèixer les principals tècniques de preparació de dades i adonar-se de la seva conveniència per a cada tipus diferent de tasca de construcció de models.

3. Comprendre el concepte de reducció de valors mitjançant la discretització.

4. Comprendre el concepte de reducció de dimensionalitat mitjançant la selecció d'atributs.

5. Tenir una idea de la manera en què es pot reduir el nombre d'observacions d'un conjunt de dades sense afectar negativament la qualitat dels models que se'n poden extreure.

1. Repàs de conceptes estadístics

Com ja sabem, en un projecte de mineria de dades, un cop seleccionades les fonts de dades corresponents –un procés que pot ser prou complicat per ell mateix–, cal adequar tant els valors com els formats de les dades.

Les operacions de preparació de dades...

... permeten que puguin ser tractades pels algorismes corresponents i ens donin el millor model possible amb les dades disponibles.

L'**objectiu principal de la preparació** de dades consisteix a organitzar-les de manera que puguin ser processades pels programes de construcció de models que s'hagin escollit i, al mateix temps, assegurar que les dades estan de manera que s'obtingui el millor model que es pot extreure del conjunt de dades.

La preparació de dades...

... ocupa una mitjana del 70% del temps total d'un projecte de mineria de dades.

Aquest darrer aspecte resulta especialment difícil, ja que no hi ha una mesura de qualitat absoluta ni independent de la tasca que es vulgui fer, o del tipus de model que es vulgui extreure.

Les mesures de qualitat dels models predictius, per exemple, no han de ser necessàriament les mateixes que les que s'utilitzen per a obtenir bons models d'agregació, de descripció o d'explicació. Tot i així, s'han proposat mètodes generals, no supervisats o "cecs" al tipus de tasca que cal dur a terme que tenen una utilitat prou alta.

La qualitat...

... del procés de mineria de dades està en relació amb la tasca a la qual es vulgui aplicar el model.

Dins el procés de preparació de les dades, separem les tasques següents:

- **Transformació de valors**: bàsicament tècniques per a canviar els valors sense perdre informació i fer que puguin ser tractats pel mètode que ens interessi.

- **Reducció de dimensionalitat**: eliminar casos dins el conjunt de dades original, o bé eliminar atributs, o totes dues coses al mateix temps per a obtenir models de la mateixa qualitat amb menys esforç computacional.

Per a poder valorar millor algunes de les tècniques que presentarem en aquest mòdul i en la resta del curs, cal refrescar alguns conceptes que procedeixen de l'estadística. Per a aprofundir més alguns detalls, recomanem la consulta d'un text especialitzat, com la lectura que esmentem al marge.

Lectura recomanada

Aprofundiu els conceptes d'estadística amb la lectura de l'obra següent:

L. Lebart; A. Morineau; J.P. Fenelon (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.

1.1. Conceptes mínims d'estadística

Per a parlar amb una certa uniformitat, adoptarem una bona part de la terminologia emprada en estadística. En efecte, el procés de mineria de dades incor-

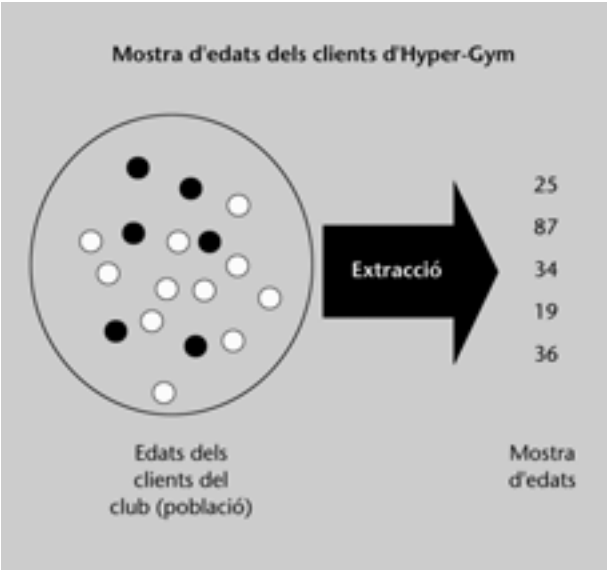
pora molts dels interessos i passos de l'anàlisi de dades, i també de l'obtenció de models a partir de conjunts de dades, que són les tasques pròpies de l'estadística. Cal que recordem una sèrie de conceptes estadístics que ens permetran d'aclarir les coses.

Podem utilitzar tota la sèrie de conceptes i eines aportats durant anys per la inferència estadística, els coneixements d'estadística que permeten de generalitzar a partir de les dades observades les propietats corresponents a casos que no es poden observar directament (en el nostre cas, a casos futurs que s'han de classificar correctament).

El concepte *població* consisteix en tots els objectes que volem estudiar i no solament aquells les característiques dels quals hem pogut observar i recollir directament. Per exemple, els següents: tots els clients de la cadena Hyper-Gym, i no tan sols els que hem utilitzat per a construir el conjunt d'entrenament; o tots els clients (morosos i no morosos) d'un banc i no solament aquells dels quals hem emprat les dades per a construir un model de classificació que predigui qui podria ser morós i qui no.

Vegeu els casos d'exemple dels clients d'un banc i de la cadena Hyper-Gym presentats en el mòdul "El procés de descobriment de coneixement a partir de dades" d'aquesta assignatura.

Una **mostra** és un subconjunt triat aleatòriament a partir d'una població de manera que sigui representatiu de la població. Aquí, la dificultat és garantir que les dades dels casos seleccionats per a formar la mostra (els clients d'Hyper-Gym o els del banc) efectivament seran representatius. Cada conjunt de dades referit a cada cas també s'anomena *observació*.



Aquí teniu una sèrie de conceptes propis de l'estadística descriptiva que ens permetran parlar amb més correcció tan de les propietats de les poblacions com de les mostres. No els donem de manera formal, sinó només per a tenir un vocabulari de discussió. Per a aprofundir-hi més, cal consultar llibres

d'estadística especialitzats. Els conceptes bàsics que ens calen a nosaltres són:

1) Rang: diferència entre els valors mínim i màxim de les observacions de la mostra.

2) Mitjana mostral: suma de totes les observacions de la mostra dividida pel nombre d'observacions fetes. La mitjana mostral (estimada a partir d'un conjunt d'observacions) es denota per $\overline{X}$ i es calcula per a una mostra de *n* observacions segons la fórmula següent:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

on x_i és cadascun dels valors observats per a la variable *X*. La mitjana poblacional es representa mitjançant la lletra grega μ.

3) Variància mostral: mesura de la dispersió dels valors de la mostra respecte a la mitjana. Es denota per s^2 i, per a una mostra de *n* observacions d'una variable x_i, es calcula segons la fórmula següent:

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} (x_i - \overline{X})^2 .$$

La **variància poblacional** es representa per la lletra grega σ^2 i es calcula segons la fórmula següent:

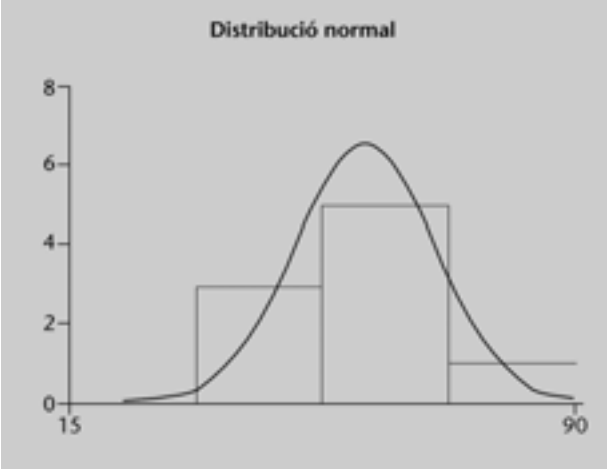
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 .$$

4) Desviació estàndard: arrel quadrada de la variància; hi podem aplicar la mateixa divisió entre mostral i poblacional un altre cop. L'interès de la desviació estàndard està en relació amb les propietats de les distribucions de probabilitat, que tot seguit comentem.

5) Distribució de probabilitat: és una descripció de la manera en què es donen els diversos valors corresponents a una població. Bàsicament, la distribució ens diu amb quina freqüència teòrica apareixen els valors de la població. La mostra se suposa que ha estat extreta d'una població que segueix una distribució determinada. Si la discrepància entre les freqüències observades de cada valor en la mostra respecte a les teòriques de la població és molt forta, cal considerar el següent: a) els valors no són representatius; o b) la població segueix, en realitat, una distribució diferent.

1.2. La distribució normal

Les distribucions es formalitzen com a funcions on l'eix *X* representa els valors i l'eix *Y* representa la freqüència teòrica d'aparició de cada valor (evidentment, això admet extensions a més d'una dimensió). Hi ha molts fenòmens en diversos àmbits que segueixen distribucions semblants. D'aquests fenòmens n'ha sortit la caracterització funcional de les diverses distribucions. Les que ens interessen més són la **distribució normal** o **distribució gaussiana** i la **distribució binomial**. En farem ús i una referència més detallada quan la distribució ho requereixi.



La funció de densitat de probabilitat d'una distribució normal per a una variable *X* té l'aspecte següent (una mica esgarrifós a primer cop d'ull per a qui faci temps que no s'enfronta amb l'anàlisi matemàtica):

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

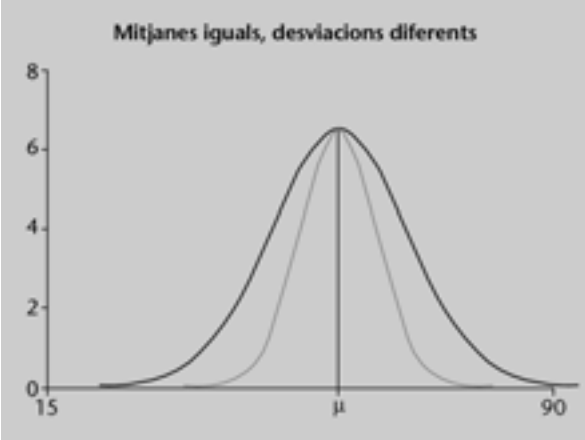
on π és el nombre ben conegut per tots; σ és la variància de la població; μ és la mitjana poblacional i *e* és el nombre 2,718... Aquesta funció no proporciona realment la probabilitat de cada valor de *X*. De fet, la probabilitat és proporcional a la superfície que està compresa entre aquesta funció i l'eix de les abscisses. La suma total de la regió sota la funció situada entre els valors mínim i màxim que pot prendre *X* és 1 i es pot calcular integrant la funció.

No hem d'oblidar les característiques que una funció així dóna a la distribució.

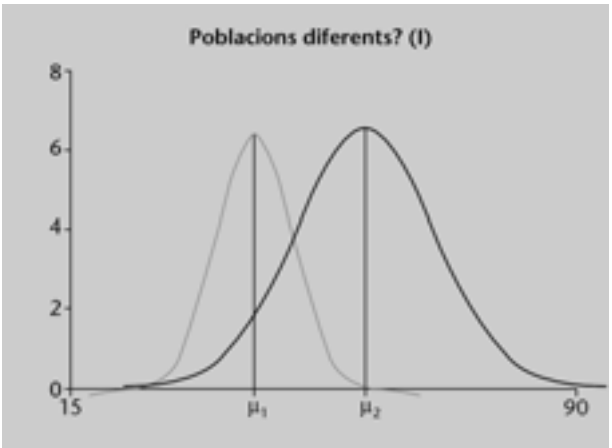
D'entrada, només a partir de l'aspecte gràfic de la distribució i si coneixem els valors de la mitjana i la desviació, podem tenir una primera aproximació a la semblança o la diferència entre dues poblacions.

Exemple de deducció a partir de dues distribucions normals

Aquí tenim dues distribucions amb la mateixa mitjana però amb desviacions diferents.



A continuació veiem dues distribucions de les quals tindríem la temptació de dir que pertanyen a poblacions diferents:



Aquesta darrera gràfica correspon a la distribució d'edats entre els homes i les dones que són socis del nostre club imaginari.

1.2.1. Propietats de la distribució normal

La primera cosa que cal tenir en compte és la relació entre la mitjana i la variància poblacional. Si tenim una població en què l'edat té una mitjana de trenta anys i una desviació estàndard de sis anys, una persona de vint-i-quatre anys està a una desviació estàndard de la mitjana poblacional i una persona de quaranta-dos anys està a dues desviacions estàndard.

Per a facilitar les comparacions, s'acostuma a transformar les distribucions de manera que la mitjana sigui 0, i la desviació típica, 1. Això es fa mitjançant el canvi de variable següent:

$$z = \frac{(x - \mu)}{\sigma}$$

on x és un valor concret de la variable X.

Ara hem de tenir en compte la simetria de la distribució normal. El punt màxim correspon just a la mitjana poblacional (μ). La funció decau tant cap a l'esquerra com cap a la dreta. El primer punt d'inflexió és a una distància d'una desviació estàndard de la mitjana.

Els valors *z* tenen la peculiaritat que permeten de comparar mesures semblants que poden procedir de poblacions diferents.

Comparació de mesures de poblacions diferents

Ens diuen que una persona procedent d'un país on es menja molt de greix té un grau de colesterol al qual correspon un valor *z* d'1, i que una altra persona d'un país on la població és majoritàriament vegetariana té un grau de colesterol al qual correspon un valor *z* de 3. Quina de les dues està en millors condicions?

La importància d'aquesta manera de representar la distribució normal es veu en calcular la superfície situada sota la funció entre els punts que són a una desviació estàndard cap a l'esquerra i a una desviació estàndard cap a la dreta: en aquesta zona se situa el 68% de les observacions; a dues desviacions estàndard s'aplega el 90% de les observacions. L'interval normal deixa a la seva dreta el 5% dels casos. Ja veurem quina utilitat té això.

1.2.2. Estimació de la probabilitat d'un valor mostral

Les característiques de la distribució normal ens ajudaran a poder contestar preguntes com la de si un determinat valor procedeix de la població o no.

Per exemple,...

... si fem una enquesta entre la gent que accedeix al web d'Hyper-Gym i algú ens contesta que té cent anys i és soci actiu del gimnàs, fins a quin punt ens l'hem de creure? Més que una persona que afirmi que n'és sòcia activa i que en té vint-i-vuit? Menys que una que afirmi que en té tretze?

Interval de probabilitat 1 – α

L'interval de probabilitat $1 - \alpha$ és el que aplega el $100 - \alpha$ per cent d'observacions de la població. Aquest interval permet d'avaluar fins a quin punt les diferències de valor respecte a la mitjana poblacional corresponen a variacions purament aleatòries. El valor α associat a l'interval de probabilitat indica la probabilitat que un valor no pertanyi a l'interval.

Interval de confiança 1 – α

Si tenim una mostra que procedeix d'una població desconeguda, es pot definir un interval simètric respecte a una proporció observada d'un valor determinat que tingui la probabilitat de l'$1 - \alpha$ de contenir el valor desconegut d'aquesta proporció a la població. També s'acostuma a definir amb $\alpha = 0{,}05$.

L'**interval de probabilitat** pressuposa que es coneix la distribució de la població. Per tant, indica dins de quina part de la distribució estarien situades certes proporcions de valors. L'**interval de confiança**, però, parteix del coneixement d'una proporció d'aparicions d'un valor determinat dins una mostra de grandària *n*, i ens dóna un interval que té una probabilitat elevada de contenir la veritable proporció. El primer permet de resoldre problemes de predicció, i el segon, d'estimació.

Veurem que algunes d'aquestes idees s'aplicaran a l'hora d'estimar la qualitat dels models obtinguts.

Vegeu l'estimació de la qualitat dels models obtinguts en el mòdul "Avaluació de models" d'aquesta assignatura.

A partir de l'error estàndard, és possible calcular l'interval de confiança. En efecte, si en una mostra de n observacions s'ha observat una proporció p per a un valor determinat, llavors l'error estàndard es defineix de la manera següent:

$$es = \sqrt{\frac{p(1-p)}{n}}$$

L'interval de confiança del 95% per a un valor p_0 es calcula multiplicant el seu error estàndard per 2 (això és una aproximació, ja que el valor z que correspon a $\alpha = 0{,}05$ és 1,96).

Proves d'hipòtesi

A partir del concepte d'*interval de probabilitat*, *ens* podem començar a contestar preguntes interessants. Podem fer hipòtesis i veure si queden validades. En altres paraules, tenim un primer mecanisme per a justificar informació i generar coneixement. Ja veurem que això també tindrà importància en l'avaluació i validació de models.

Vegeu el mòdul "Avaluació de models" d'aquesta assignatura.

Proves de conformitat

Amb relació a les proves de conformitat, es tracta de saber si els resultats que s'han observat en una mostra s'adiuen amb la distribució de població que suposem.

Exemple de proves de conformitat

Hem observat que entre deu socis d'un dels centres d'Hyper-Gym, la proporció persones de trenta anys és del 33%, i volem veure si això es correspon amb les característiques globals de tots els centres. Què diríem si la proporció general fos del 60%?

Ho podem resoldre comparant la proporció observada amb la teòrica. Les dades de les edats dels socis són les següents:

Edats	30	34	45	30	45	30	60	32	45	43

Calculem el valor z corresponent:

$$z = \frac{|p - p_0|}{es}$$

En aquest cas tenim el següent:

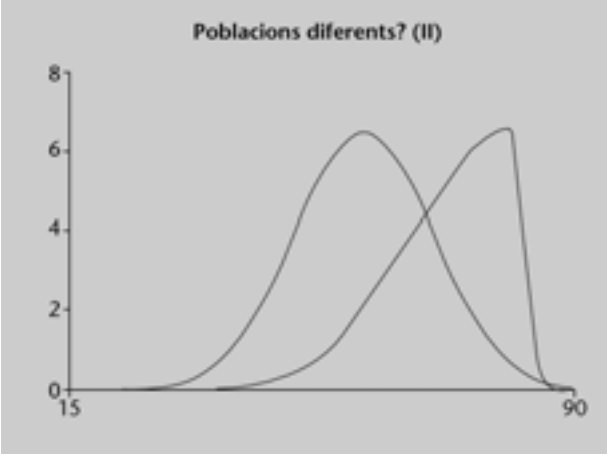
$$es = \sqrt{\frac{0{,}33 \cdot 0{,}67}{10}} = \sqrt{0{,}022} = 0{,}148 \text{ , i}$$

$$z = \frac{|0{,}60 - 033|}{0{,}148} = 1{,}824 .$$

Podeu consultar els valors z en unes taules d'Estadística per saber què indica aquest resultat. Fixeu-vos que el valor de z per a un nivell de confiança del 5% és 0,46.

1.2.3. Comparació de distribucions (valors *t*)

Paga la pena tenir en compte els paràmetres d'una distribució per a poder fer comparacions. Hi ha distribucions amb una dispersió molt alta i d'altres amb una distribució molt baixa. Les primeres són més "aplanades" que les segones. Donem, doncs, una primera aproximació per saber si dues distribucions representen la mateixa població o poblacions diferents.



Introduirem uns quants conceptes i el procediment per a poder respondre aquesta mena de preguntes. Per exemple, si dos valors procedeixen de distribucions diferents, o si el rendiment d'un mètode de mineria de dades és significativament millor que el d'un altre. En un altre mòdul es torna a fer referència a aquest darrer punt, molt important.

Vegeu el mòdul "Avaluació de models" d'aquesta mateixa assignatura.

Significació o significança

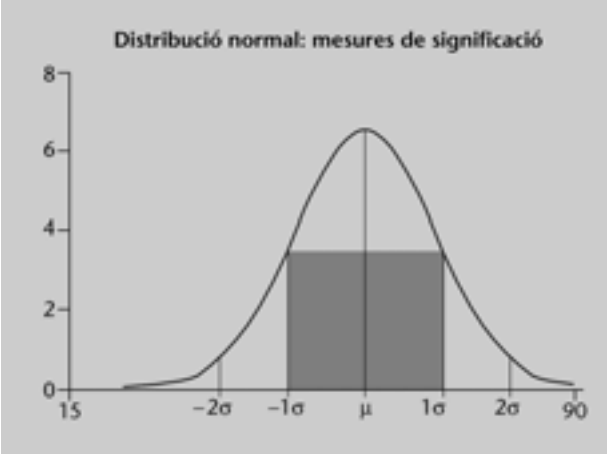
Suposem que hem extret dues mostres de la mateixa població i apreciem diferències en la freqüència observada entre el mateix valor de cada mostra. Per exemple, en la mostra de clients d'Hyper-Gym, extraiem dues mostres de cent clients cadascuna. En la primera mostra trobem que la mitjana d'edat és de trenta-quatre anys, i en la segona, de trenta-vuit. Això és una casualitat o és una diferència significativa? És a dir, les edats d'una mostra i l'altra són prou diferents entre si?

> En general, si la diferència observada és superior a dues desviacions típiques i mitja, considerarem que, efectivament, la diferència és significativa perquè a aquesta distància de la mitjana, sota la corba de la distribució normal, no queda més que un 5% de la població.

És a dir, només tenim el 5% de possibilitats de trobar una diferència així. Estem segurs que les mostres procedeixen de dues poblacions diferents amb un nivell de confiança del 95%. Veurem altres mesures de significació quan calgui.

Precisarem una mica més l'aspecte de la significació. Suposem que tenim dues mostres que no sabem si procedeixen de la mateixa població o de poblacions diferents. Anomenem $x_1, \ldots, x_n$ els valors de la primera, i $y_1, \ldots, y_n$ els de la segona. Les seves mitjanes respectives són $\overline{X}$ i $\overline{Y}$. Volem saber si $\overline{X}$ és significativament diferent de $\overline{Y}$.

Si el nombre d'observacions és prou alt, les mitjanes d'un conjunt d'observacions que suposem independents (és a dir, que no tenen influència entre si) segueixen una distribució normal.



Suposem que la mitjana de població és μ. Si coneguéssim la variància d'aquesta distribució normal, podríem obtenir el límits de confiança de μ. Com que no la tenim, la podem estimar a partir de la variància mostral. Podem calcular la variància de la mitjana $\overline{X}$ dividint la variància mostral (s^2, en la nostra notació) pel nombre d'observacions, n. Finalment, podem intentar aproximar el valor z corresponent:

$$\frac{\overline{X} - \mu}{\sqrt{s^2/n}}.$$

Aquest valor, però, no segueix una distribució normal, sinó una distribució anomenada t de Student amb $n - 1$ graus de llibertat.

El problema que ens plantegem ara és el que s'anomena *provar si les dues mitjanes són iguals*. Això és el mateix que veure si la diferència entre mitjanes, $\overline{X} - \overline{Y}$, és zero.

En aquest cas es defineix una variable nova, m, que recull les diferències entre els valors de les observacions corresponents, x_i i y_i. Ara cal trobar l'estadístic de Student que li correspon:

$$t = \frac{\overline{m} - 0}{\sqrt{s^2/n}}.$$

Consultem les taules de *t* amb els graus de llibertat corresponents a un valor de confiança del α%. Si el valor de *t* és més gran que el que apareix en la taula, llavors rebutgem la hipòtesi nul·la (que les dues mitjanes són iguals).

Exemple de significació a partir de l'estadístic de Student

Podem comprovar amb aquests dinou valors per a l'edat dels homes i dinou valors per a l'edat de les dones, seleccionats entre els socis del mateix centre d'Hyper-Gym, que la diferència no és significativa.

Edat (D)	Edat (H)
30	24
32	35
32	36
36	41
30	21
55	56
32	23
40	33
60	32
41	20
36	67
59	24
33	22
34	61
34	30
35	30
32	28
30	23
18	29

Trobareu que el valor $P(T \Leftarrow t)$ és 0,8 i el valor corresponent a la confiança del 5% és de 2,09.

Com que aquest valor està per sota, hem d'acceptar que les mitjanes de les dues edats no són significativament diferents.

Independència entre dues variables

Una propietat interessant pel que fa a la relació que hi ha entre dues variables és la de si els valors que pren una tenen influència sobre els que pren l'altra. Si, efectivament, en tenen, les variables són dependents.

Una manera de mesurar la dependència o independència entre les variables és efectuant la **prova de khi quadrat** (χ^2). Més endavant, en parlar de discretització, fem una descripció més detallada d'aquesta prova.

Vegeu la prova de khi quadrat en el subapartat 3.2 d'aquest mòdul didàctic.

Una altra manera de mesurar-ne la dependència consisteix a calcular-ne el **grau de correlació**. Les variables que tenen una correlació alta es poden interpretar com a més dependents.

1.3. Altres distribucions

No és qüestió aquí d'entrar a discutir i explicar altres distribucions ni les seves propietats. En tot cas, a mesura que ens faci falta al llarg del text, ja en farem esment en el moment en què calgui. Només anotarem que les propietats que hem esmentat fins ara fan referència a distribucions definides sobre una única variable (distribucions univariants o monovariants), mentre que en mineria de dades els problemes sempre tenen més d'una variable.

Per exemple,...

... en el cas del nostre gimnàs fictici:

P (*Sexe* = 'Dona', *Edat* = 43, *Horari* = 'Matí', *Anys al club* = 4,*Act1* = 'loga',*Act2* = '*Steps*')

i no tenint en compte tots els atributs.

En efecte, un conjunt de tuples extretes d'una base de dades que té com a atributs X_1, ..., X_n es pot considerar una mostra extreta d'una població definida sobre el mateix conjunt de variables X_1, ..., X_n. La probabilitat d'observar una combinació de valors concrets que corresponen a una de les observacions de la base de dades $P(X_1 = X_{11}, X_2 = X_{23}, X_3 = X_{33}, ..., X_n = X_{n3})$ segueix una distribució multivariant on cadascuna de les variables pot seguir un mateix tipus de distribució (per exemple, normal), però no necessàriament amb els mateixos paràmetres.

Per exemple, els paràmetres que caracteritzen una distribució normal, mitjana i variància de població no han de ser necessàriament les mateixes per a X_1 que per a X_3. Igualment, ni tan sols el tipus de distribució que segueix X_1 ha de ser el mateix que el que segueix X_3. Per exemple, X_1 pot seguir una distribució uniforme, i X_3, una distribució normal. Ja veurem quan serà el moment fins a quin punt ens pot afectar això en la formalització i caracterització dels mètodes de mineria de dades. De moment n'hi ha prou de tenir presents els conceptes rudimentaris sobre estadística que s'han presentat fins ara.

2. Terminologia de preparació de dades: tipus d'atributs

Recordem breument aquí els diversos tipus d'atributs amb què es pot descriure un domini:

a) Numèrics: prenen valors en els reals o enters o naturals (en general, en un conjunt numèric que pot prendre valors infinits).

b) Lògics: prenen els valors 'Cert' o 'Fals', normalment representats per 0 ('Fals') i 1 ('Cert').

c) Categòrics o **discrets**: els que prenen valors en un conjunt finit, per exemple {1, 2, 3} o {'Baix', 'Mitjà', 'Alt'}.

d) Ordenats: tenen una relació d'ordre entre si. Per exemple, en principi els colors no tenen cap ordre establert; per tant, si un atribut pren els valors {'Vermell', 'Verd', 'Blau'}, no necessàriament hem de considerar que és un atribut amb valors ordenats.

Podem utilitzar també altres criteris: les dimensions del rang de valors i l'escala de mesura emprada.

Lectura complementària

Trobareu altres criteris per a descriure dominis en l'article següent:
M. Anderberg (1973). "Cluster Analysis for Applications". *Academic Press*.

Atenint-nos al rang de valors de la variable, podem distingir els tipus de variables següents:

- Les **variables contínues**: tenen un conjunt de valors infinit no numerable. Cas típic: els nombres reals.

- Les **variables discretes**: prenen valors en un conjunt finit o, com a molt, infinit numerable. Exemple: els enters.

- Les **variables binàries**: variables discretes que només poden prendre dos valors. Cas típic: la representació dels valors lògics 'Cert' i 'Fals' mitjançant els nombres 1 i 0.

Atenint-nos a l'escala en què es mesuren les variables, en podem distingir els tipus que esmentem a continuació:

- Les **escales nominals** només distingeixen dues classes; és a dir, per a dos valors X i Y, només podem dir si són iguals ($X = Y$) o diferents ($X \neq Y$). Per exemple, si tenim un atribut que recull el valor del color d'una peça del magatzem, on l'atribut *Color* s'ha definit sobre el conjunt {'Vermell', 'Verd', 'Blau'}.

- Les **escales ordinals** corresponen als valors sobre els quals es pot definir una relació d'ordre. Per a dos valors X i Y, a més de dir si són iguals ($X = Y$) o diferents ($X \neq Y$), podem distingir $X > Y$ de $X < Y$. Ens permeten d'ordenar valors quan calgui.

- Una **escala per interval** assigna significats a la diferència de valors. No solament podem dir que $X > Y$, sinó que X és $X - Z$ unitats diferent de Y. Per exemple, les lletres de l'alfabet en la codificació ASCII.

- Una **escala de proporció** (o **escala de mesura**) és un interval on s'ha fixat un valor com a origen (o zero). A més de poder dir que $X > Y$, $X < Y$ o $X = Y$, podem dir que X és X/Y vegades més gran que Y. Exemple: la temperatura en graus Celsius.

Combinant totes aquestes escales, pràcticament podem definir les característiques de qualsevol tipus d'atribut.

La raó per la qual els valors continus es tracten separadament és que podem considerar que poden arribar a tenir tants valors diferents que cadascun d'aquests valors concrets apareix amb una freqüència molt baixa. Això pot donar problemes de precisió en alguns mètodes que han de fer comparacions entre valors, ja que dos valors d'aquesta mena difícilment seran del tot iguals entre si. Per això s'apliquen tècniques de suavització de valors o discretització enfocades a reduir el nombre de valors que cal comparar.

Vegeu les tècniques de discretització al subapartat 3.2 d'aquest mòdul didàctic.

Observacions

Els atributs nominals (o atributs simbòlics) són atributs discrets, els valors dels quals no segueixen necessàriament un ordre lineal. Com ja hem assenyalat, un cas típic és el dels colors. Una variable que representi el color podria prendre els valors 'Vermell', 'Verd', 'Blau', 'Groc', 'Negre' i 'Blanc'. A efectes de representació, potser els podríem codificar amb els nombres enters de l'1 al 6 {('Vermell',1), ('Verd',2), ('Blau',3), ('Groc',4), ('Negre',5), ('Blanc',6)}, però, atenció: tot i que hi ha un ordre entre els enters que assegura que $1 < 6$, o que 6 és a una distància de 5 d'1, no té cap sentit utilitzar aquesta mena de conceptes en aquest cas, perquè no té sentit dir que el negre és més gran que el blau (si no és que, implícitament, parlem de les seves propietats en termes de longitud d'ona). S'ha d'anar molt amb compte amb aquesta mena de codificacions. A l'hora de fer comparacions o a l'hora de reduir valors, podem introduir una interpretació absolutament absurda i distorsionar moltíssim tant el procés de preparació de dades com els resultats posteriors que se'n puguin extreure.

3. Operacions de preparació de dades

En aquest apartat descriurem les principals operacions utilitzades en la preparació de dades, i les dividirem en dos grans grups: la transformació de valors i la reducció del nombre d'atributs que cal considerar.

3.1. Transformació de valors

Per **transformació de valors** entenem modificacions dins el tipus de valors que poden prendre tots o alguns dels atributs.

Les operacions més habituals són la normalització i la discretització de dades. Comentarem breument la problemàtica que sorgeix quan ens manquen els valors d'un o més atributs d'algunes observacions.

3.1.1. Normalització de dades

La **normalització** consisteix a posar les dades sobre una escala de valors equivalent que permeti la comparació d'atributs que prenen valors en dominis o rangs diferents.

Exemple de normalització de dades

Es pot normalitzar de manera que els valors numèrics quedin dins l'interval real (0, 1) o (–1, 1).

Per exemple, si tenim que l'atribut *Edat* del nostre exemple del gimnàs oscil·la entre divuit i vuitanta-set anys, podem fer que vagi a parar als valors (0, 1). Si l'atribut *Anys al club* (que oscil·la entre 0 i 14) també es transforma de manera que doni valors entre 0 i 1, llavors podem saber que un valor 0,8 en *Edat* i un valor 0,8 en *Anys al club*, tot i que no corresponen al mateix valor, sí que representen valors en l'escala alta dels atributs respectius.

La normalització és útil, o necessària, per a diversos mètodes de construcció de models com, per exemple, les xarxes neuronals o alguns mètodes basats en distàncies, com el dels veïns més propers. En efecte, si no hi ha normalització prèvia, aquests mètodes tendeixen a quedar esbiaixats per la influència dels atributs amb valors més alts, fet que distorsiona el resultat.

Comentem a continuació algunes de les tècniques de normalització habituals.

Normalització pel màxim

La normalització pel màxim consisteix a trobar el valor màxim, x_{max}, de l'atribut que cal normalitzar X i dividir tota la resta de valors per x_{max}.

Amb aquest tipus de normalització ens assegurem que el màxim rep el valor 1.

Exemple de normalització pel màxim

Si tenim aquest conjunt de valors per a l'atribut *Renda*:

53.000.000	15.400.890	7.978.999	3.500.000	12.780.000

llavors x_{max} és 53.000.000. Un cop transformats els valors, queden així:

1	0,29	0,15	0,07	0,24

Normalització per la diferència

La normalització per la diferència tracta de compensar l'efecte de la distància del valor que tractem respecte al màxim dels valors observats.

La normalització per la diferència consisteix a fer la transformació següent:

$$z_i = \frac{|x_i - x_{min}|}{|x_{max} - x_{min}|}.$$

Exemple de normalització per la diferència

Si tenim el conjunt de valors següents per a l'atribut *Renda*:

53.000.000	15.400.890	7.978.999	3.500.000	12.780.000

llavors x_{max} és 53.000.000 i x_{min} és 3.500.000. Un cop transformats els valors, queden així:

1	0,24	0,15	0	0,18

D'aquesta manera, sempre es generen els valors 0 i 1.

Escalat decimal

L'escalat decimal permet de reduir en un cert nombre de potències de 10 el valor d'un atribut.

Aquesta transformació resulta especialment útil en tractar amb valors grans (per exemple, rendes o volums de negoci).

Suposem un atribut X que mostra un valor X_i. Aquí tenim l'expressió per a dur a terme el seu escalat decimal:

$$z_i = \frac{x_i}{10^j}.$$

on j ha de ser tal que mantingui el valor màxim que pot prendre x_i per sota d'1.

Exemple d'escalat decimal

Continuant amb el mateix conjunt de valors per a la variable *Renda*:

53.000.000	15.400.890	7.978.999	3.500.000	12.780.000

volem que els valors quedin normalitzats entre 0 i 1. Llavors tenim que el valor màxim és 53.000.000. Hem de trobar el valor de j que faci que, un cop transformat, 53.000.000 sigui inferior i tan a prop com sigui possible d'1. Aquest valor de j és 8. El resultat de fer la transformació

$$z_i = \frac{x_i}{10^8}.$$

és el que es veu a continuació:

0,53	0,15	0,08	0,04	0,13

Normalització basada en la desviació estàndard: estandarditzación de valors

Els mètodes anteriors no tenen en compte la distribució dels valors existents.

El **mètode d'estandardització de valors** assegura que s'obtenen valors dins el rang escollit que tenen la propietat que la seva mitjana és zero i la seva desviació estàndard val 1.

L'estandardització consisteix a fer la transformació següent sobre els valors dels atributs:

$$z_i = \frac{x_i - \mu}{\sigma}.$$

Aquesta normalització resulta adequada per a treballar després amb mètodes que utilitzen distàncies.

Exemple d'estandardització de valors

Tenim aquest conjunt de valors per a l'atribut *Edat* de la columna de l'esquerra en què, com es pot comprovar, la mitjana és 37,3 i la desviació estàndard és 19,11. Llavors els resultats que obtenim de l'estandardització són els que veiem en la columna de la dreta:

Edat	Estandardització
22	–0,8006279
43	0,2982732
31	–0,3296703
22	–0,8006279
34	–0,1726845
22	–0,8006279
45	0,4029304
43	0,2982732
23	–0,7482993
88	2,6530615

3.2. Discretització

Recordem que una bona part dels mètodes de mineria de dades treballen amb la suposició que els atributs amb què es descriuen les instàncies o observacions existents en la base de dades de què es parteix són categòrics i no numèrics. Per aquest motiu aquesta és una de les tècniques més emprades en la preparació de dades.

La **discretització** consisteix bàsicament a establir un criteri per mitjà del qual es puguin dividir els valors d'un atribut en dos o més conjunts disjunts.

Però no solament és aquest el motiu per a discretitzar dades. Citem-ne aquests altres:

a) Cost computacional: tenint en compte que el conjunt de valors sobre el qual es treballarà després de discretitzar representa una reducció de valors que cal tractar, el nombre de comparacions i càlculs que haurà de fer el mètode corresponent de mineria de dades és més petit.

Lectura complementària

Amb relació a la velocitat en el procés d'aprenentatge consulteu l'obra següent:

J. Catlett. "On Changing Continuous Atributes into Ordered Discrete Atributes". A: Y. Kodratoff (ed.). *Proceedings of the European Working Session on Learning: Machine Learning* (pàg 164-178). Springer-Verlag.

b) Velocitat en el procés d'aprenentatge: s'ha demostrat empíricament que el temps necessari per a dur a terme un procés d'entrenament d'un mètode de mineria de dades és més curt utilitzant dades discretitzades.

c) Emmagatzemament: en general els valors discrets necessiten menys memòria per a ser emmagatzemats.

d) Grandària del model resultant: quan es treballa amb dades contínues, els models classificatoris que s'obtenen són comparativament més grans. Per exemple, els arbres de decisió que s'obtenen acostumen a tenir un factor de

ramificació més alt quan es treballa amb dades contínues que quan es treballa amb dades discretes.

e) Comprensió: la comprensió d'alguns models millora molt en descriure'n els elements utilitzant menys termes.

Evidentment, tot mètode de discretització està obligat a mantenir o millorar les característiques del model a la construcció del qual serveix. Per exemple, si introduïm un procés de classificació i obtenim taxes d'error més altes o de predicció més baixes que sense discretitzar, poca cosa hem guanyat. Per tant, el que cerquen la majoria de mètodes de discretització és mantenir la informació associada a l'atribut que es discretitza. Un inconvenient que normalment se cita en parlar de mètodes de discretització és precisament la pèrdua d'informació sobre els valors continus. Aquesta mena d'efectes pot tenir una influència notable en la precisió dels mètodes d'aprenentatge.

Un exemple típic de pèrdua d'informació...

... en el procés de discretització és que els dos nombres situats en els extrems de l'interval de discretització es consideren el mateix. Aquest efecte influeix en la precisió dels models d'aprenentatge.

Els mètodes de discretització es poden dividir en diverses categories:

- **Supervisats** o **no supervisats**: mètodes que tenen en compte els valors de l'atribut classe o no (els primers adreçats a classificació; els segons, a altres tasques).

- **Locals** o **globals**: mètodes limitats a un atribut cada vegada o a un subconjunt de les dades originals o conjunt de dades, o que actuen sobre tots els atributs i totes les dades.

- **Parametritzats** i **no parametritzats**: els primers comencen coneixent el nombre màxim d'intervals que cal generar per a un atribut específic, mentre que els altres han de trobar aquest nombre automàticament.

Lectura complementària

Trobareu més informació amb relació a les diverses categories dels mètodes de discretització en l'obra següent:

J. Dougherty; R. Kohavi; M. Sahami (1995). "Supervised and Unsupervised Discretizations of Continuous Features". *Proceedings of the 12th International Conference on Machine Learning* (pàg. 194-202). Morgan Kaufmann Publishers.

El nombre de conjunts al qual es vol arribar depèn de l'aplicació i el domini concrets en què es treballi. Per exemple, segons el tipus de resultat que volem obtenir, aquestes dues discretitzacions poden ser igualment útils o no. En tots dos casos utilitzem l'atribut *Anys al club* que pren valors entre 0 i 16:

Anys al club	*Conjunt*
0-2	Inicial
2-4	Estable
4-6	Fidel
6-11	Sènior
11-16	Veterà

o bé aquest altre:

Anys al club	*Conjunt*
0, 4	Estable
< 5	Fidel

És clar que els mètodes per a discretitzar depenen molt de la tasca que es vol dur a terme. Per tant, la qualitat del mètode varia segons la millora de la qualitat del model que resulti d'aplicar una tècnica de mineria de dades sobre el conjunt de dades discretitzades.

En classificació es pot utilitzar la informació respecte a quina és la variable de classe per a poder millorar el resultat de la discretització; en agregació el criteri és diferent. Els mètodes per a discretitzar variables en un cas o un altre haurien de ser diferents, ja que les tasques són diferents. Tot i així, s'han desenvolupat diverses maneres de discretitzar que no tenen en compte el tipus d'aplicació final que tindran i que, per tant, han de recórrer a altres criteris per a tenir una base on puguin assegurar que el resultat no serà pitjor que sense discretitzar.

L'objectiu de tots aquests mètodes, repetim, és obtenir una divisió en intervals a partir d'un atribut numèric continu, de manera que a cada interval es pugui associar una etiqueta (categoria).

Farem una descripció dels mètodes progressant des dels que semblen més naturals i senzills fins a d'altres de més refinats.

3.2.1. Mètodes no supervisats

En els mètodes no supervisats no tenim cap informació sobre la tasca que ha de complir el model que es construirà a partir de les dades discretitzades. Per tant, només utilitzem la informació procedent de les observacions.

Lectures recomanades

Consulteu els mètodes no supervisats en les obres següents:

J.R. Martín (1999). *Discretització de variables contínues i la seva aplicació*. Projecte de final de carrera, Enginyeria en Informàtica. Departament de Llenguatges i Sistemes Informàtics. Barcelona: Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya.

J.J. Valdés (1997). "Fuzzy Clustering and Multidimensional Signal Processing in Enviromental Studies". *Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing*. Aache (Alemanya).

Una bona referència per a aquest subapartat és la lectura recomanada al marge (Martín, 1999), que resumeix i implementa una bona part dels mètodes següents amb una descripció dels resultats experimentals obtinguts amb diversos conjunts de dades de referència. Hem adaptat alguns dels algorismes que s'hi proposen, però encara en recull més, prou interessants (Valdés, 1997).

Mètodes de partició en intervals d'igual amplitud

El mètode de partició en intervals d'igual amplitud és el mètode més senzill. Consisteix en l'algorisme següent.

Donat un atribut X numèric amb valor mínim x_{min} i màxim x_{max}, se segueixen els passos següents:

1) Fixar un nombre k d'intervals que cal assolir.

2) Dividir el rang de valors (x_{min}, x_{max}) en k intervals $\{(x_{1min}, x_{1max}), ..., (x_{kmin}, x_{kmax})\}$ tals que $x_{1min} = x_{min}$ i $x_{kmax} = x_{max}$.

Es tracta de fer que la distància entre el màxim i el mínim de cada interval sigui la mateixa: $(x_{min} - x_{max})/k$.

Exemple de partició en intervals d'igual amplitud

Aquí teniu els intervals que s'obtenen amb $k = 4$ i l'atribut *Anys al club* de la base de dades d'Hyper-Gym:

Tenim $Anys\ al\ club_{min} = 0$ i $Anys\ al\ club_{max} = 16$. Aleshores:

Intervals
0-4
5-8
9-12
13-16

Com tot mètode massa fàcil el mètode de partició en intervals d'igual amplitud té els seus inconvenients:

a) Dóna la mateixa importància a tots els valors, independentment de la freqüència d'aparició que tinguin.

Per exemple, si resulta que només tenim un soci que fa setze anys que és al club, hauríem de pensar que 16 no es pot considerar un valor representatiu i que, si el mantenim, farem que tant la capacitat predictiva com la descriptiva d'aquest atribut quedin alterades per la importància exagerada que donem a un valor poc normal.

b) En cas que vulguem utilitzar el mètode com a punt de partida de classificació, ens podem trobar que es barregin dins un mateix interval valors que corresponen a classes diferents, intervals poc homogenis.

Els mètodes que, com els arbres de decisió, es basen en la propietat que acabem d'esmentar per decidir quan un atribut és útil per a separar el conjunt de dades en classes poden quedar afectats per l'ús d'aquest mètode de discretització.

Vegeu els arbres de decisió en el mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

c) Amb aquest mètode no hi ha manera de trobar un valor de k que sigui prou bo.

Obtenció d'intervals de discretització d'igual freqüència

Un dels problemes que hem citat en comentar el mètode anterior és que pot generar intervals en què les diverses classes o valors es distribueixin amb fre-

qüències diferents. Per aquest motiu es pot introduir informació sobre la freqüència requerida de la manera següent:

- indicar el nombre d'intervals que cal obtenir;
- indicar la freqüència que es vol obtenir per als intervals:

$$\text{Freqüència} = \frac{\text{Nombre de valors}}{\text{Nombre dintervals}}.$$

El procediment consisteix en l'algorisme següent. Donat un vector **A** que guarda els n valors de l'atribut que es vol considerar, X, cal seguir els passos següents:

1) Ordenar els valors de l'atribut X que es vol considerar de més petit a més gran: $x_1, ..., x_n$.

2) Fixar un nombre d'intervals k.

3) Calcular el valor de la freqüència per interval $freq = n/k$.

4) Executar el bucle següent:

```
Total := 0
Per a Interval_i = 1 fins a k fer
    Per a x_i = 1 fins a freq fer
        Per a punt = 1 fins a freq fer
        Assignar A (Total + punt) a Interval_i
        fper
fper
    Total := Total + freq
fper
```

El mètode és senzill d'aplicar, però depèn críticament de la suposició que hi ha una distribució uniforme dels valors de l'atribut que cal discretitzar. Com que el criteri és mantenir la mateixa freqüència dins el mateix interval, valors molts dispars poden anar a parar al mateix interval.

Exemple d'obtenció d'intervals de discretització d'igual freqüència

Aquí tenim un conjunt de valors per a l'atribut *Edat*:

55	22	27	40	28	22	28	31	27	31	31	55

Els ordenem de més petit a més gran:

22	22	27	27	28	28	31	31	31	40	55	55

i guardem la seva informació de freqüència:

Valor	22	27	28	31	40	55
Freq.	2	2	2	3	1	2

Farem una discretització en tres intervals ($k = 3$); per tant, la freqüència per interval hauria de ser n/k (on n indica els valors existents, en aquest cas dotze). Per tant, la freqüència desitjable per interval és quatre.

22 22 27 27	28 28 31 31 31	40 55 55

Els punts de tall són a < 28 i < 31. Com podeu veure, els intervals no són gaire uniformes.

Suposició de normalitat (utilització d'estimador)

Quan s'està en condicions de suposar que la distribució que segueixen els valors de l'atribut és normal, es pot aprofitar aquest punt de partida per a obtenir discretitzacions que millorin alguns dels problemes que presenten els mètodes anteriors.

Lectura recomanada

Trobareu la suposició de normalitat explicada amb més detall en l'article següent:

M. Anderberg (1973). "Cluster Analysis for Applications". *Academic Press*.

Recordem que la funció de densitat de probabilitat pròpia de la distribució normal és la següent:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}.$$

Si l'atribut que volem discretitzar prové d'una distribució normal, llavors per a cada interval que podem introduir s'ha de complir que en l'interval i-èsim (x_{imin}, x_{imax}) la proporció d'observacions de la classe i-èsima és igual a l'àrea que hi ha sota la funció de densitat entre els punts límit de cada interval.

La proporció de la població que és entre tots dos extrems és $P(x_{imax}<x\leq x_{imin}) = F(x_{imax}) - F(x_{imin})$, on $F(x)$ dóna el valor de densitat de probabilitat corresponent al punt X un cop hem efectuat la transformació $((x - \mu)/\sigma)$, que ens permet d'establir la comparació amb la distribució normal de mitjana 0 i la desviació estàndard 1.

Vegeu l'estandardització de valors en el subapartat 3.1.1 d'aquest mòdul

Mirarem si la proporció de valors que apareix en cada interval s'adiu amb la hipòtesi de normalitat. El procediment és el que expliquem a continuació. Donat un vector **A** amb els valors, l'algorisme efectua els passos següents:

1) Ordenar els valors de l'atribut x presents en les dades $x_1, ..., x_n$.

2) Fixar el nombre d'intervals k.

3) Fixar els valors z mínims i màxims de cada interval d'acceptació: z_{imin} i z_{imax}.

4) Calcular el rang: $|z_{imin} - z_{imax}|$.

5) Calcular el rang de cada interval: $|z_{imin} - z_{imax}|/k$.

6) Per a cada valor extrem calculem la quantitat P_i (proporció de dades) que li correspon.

7) Mentre la proporció de valors que s'ha llegit fins ara sigui més petita que P_i posem el valor llegit $\boldsymbol{A}(i)$ dins l'$Interval_i$.

El primer interval tindrà $n \cdot F(x_{1x})$ observacions si el nombre total d'observacions en la base de dades és n. El segon interval en tindrà $n \cdot F(x_{2x})$, i així successivament.

En principi, el mètode dóna resultats més que acceptables quan la distribució que segueixen les dades es pot considerar normal. A vegades és útil emprar el valor màxim observat en la mostra com a possible punt de tall del darrer interval. Però també pot passar que sigui un valor extrem o bé que sigui un valor inferior al que li correspondria en la mostra distribucional.

El mètode es pot fer més senzill si s'evita demanar a l'usuari els valors z i, en canvi, es calculen a partir de les dades.

Mètode *k-means*

Tornarem a trobar la idea del mètode *k-means* en un altre context, en parlar de mètodes d'agregació. De fet, es pot considerar que una manera d'afrontar el problema de la discretització o reducció de valors per a una variable consisteix a repartir els valors entre els diversos intervals de manera que, donats k intervals, es minimitzi la distància entre cada valor i el valor mitjà del seu interval corresponent. En el fons, el que es fa és generar diversos grups i anar fusionant els grups (conglomerats o *clusters*) més propers.

Lectura recomanada

Trobareu el mètode *k-means* explicat en l'article següent:

J.I. Hartigan; M.R. Wong (1979). "A *k*-means Algorithm AS136: Clustering Algorithm". *Applied Statistics* (vol. 28, núm. 1, pàg. 100-108).

Tractarem de les funcions de distància en parlar dels mètodes d'agregació. N'hi ha diverses, però aquí només avançarem una de les més utilitzades en aquest context, que és la distància euclidiana.

La **distància euclidiana** entre dos valors x_1 i x_2 es calcula mitjançant la fórmula següent:

$$Dist(x_1,x_2) = \sqrt{\sum (x_1 - x_2)^2}.$$

El **mètode de *k-means*** consisteix a comparar cada valor x_j amb el valor mitjà dels dos intervals adjacents: $Interval_i$, que és el que ocupa en un moment donat, i el següent, $Interval_{i+1}$:

- Si la distància al valor mitjà del seu interval $\overline{X}_i$ és més petita que respecte al $\overline{X}_{i+1}$ següent, el valor roman en el seu interval.
- En cas contrari, es passa a l'interval següent.

En el cas de la distància euclidiana, la fórmula de càlcul de la distància és la següent:

$$Dist(x_j,\bar{x}_i) = \sqrt{\sum (x_j - \bar{x}_i)^2}$$

A partir d'aquesta distància i de la partició en intervals que hi ha en un moment donat, es pot veure com són d'homogènies les particions totals existents. Una manera de calcular això és veure quin és l'error mitjà que es comet amb l'assignació actual de valors a intervals. Es calcula la distància mitjana dels valors de cada interval als seus "centres" (o medianes) corresponents i se'n treu la mitjana aritmètica per a l'actual configuració d'intervals. Per a un conjunt de valors X i un conjunt de particions k amb mitjanes d'interval $\bar{x}_1,...,\bar{x}_k$, es té l'expressió de l'error següent:

$$Error_{actual}(configuració) = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\sum_{j=1}^{i_{max}} (x_j - \bar{x}_i)^2},$$

on i_{max} és el nombre de valors que té l'interval i en la configuració actual.

Si en la iteració següent l'error de la configuració augmenta, és una indicació que introduïm intervals encara més desordenats que els que teníem en el pas anterior. En aquest cas, l'algorisme s'atura.

L'algorisme de discretització per *k-means* funciona de la manera que expliquem a continuació. Donat un vector **A** amb els valors $x_1, ..., x_n$, s'efectuen els passos següents:

1) Ordenar els valors de l'atribut X presents en les dades $x_1, ..., x_n$.

2) Fixar el nombre d'intervals k i el nombre de valors diferents n.

3) Executar el bucle següent:

Per a cada $Interval_i$ **fer**
  Assignar a $Interval_i$ els n/k valors següents
Fper {això genera la $Configuració_1$}
Calcular $Error_{actual}$ ($Configuració_1$)
Repetir

4) Guardar la configuració actual a $Configuracions_i$:

$$Error_{anterior}(Configuracions_i) = Error_{actual}(Configuracions_i)$$

5) Assignar els valors a cada interval:

Per a cada valor x_j de $Interval_i$ **fer**
  Si $Dist(x_j, \overline{x}_i) > Dist(x_j, \overline{x}_{i-1})$ **llavors assignar** valor x_j a $Interval_{i-1}$ **fsi**
  Si $Dist(x_j, \overline{x}_i) > Dist(x_j, \overline{x}_{i+1})$ **llavors assignar** valor x_j a $Interval_{i+1}$ **fsi**
Fper {Aquí s'ha generat una nova configuració, $Configuracions_{i+1}$}
Calcular $Error_{actual}$ $(Configuracions_{i+1})$
Fins que $Error_{actual}$ $(Configuracions_i) \geq Error_{actual}$ $(Configuracions_{i+1})$

Exemple de mètode de discretització per k-means

Podem veure un exemple amb un conjunt d'edats:

22	22	22	27	27	28	31	31	31	40	55	55

El nombre de valors diferents és $n = 6$. Suposem que volem obtenir una partició amb $k = 3$ intervals. Llavors els intervals inicials estan formats per n/k elements (6/3 = 2) i són els següents:

Interval 1	Interval 2	Interval 3	Interval 4	Interval 5	Interval 6
22 22	22 27	27 28	31 31	31 40	55 55

Les seves mitjanes i errors són els següents:

Interval	Mitjana	Distància mitjana (error de l'interval)
1	22	0
2	24,5	1,666
3	27,5	0,333
4	31	0
5	35,5	3
6	55	0

L'error mitjà d'aquesta configuració és 1,666. La nova configuració és la següent:

Interval 1	Interval 2	Interval 3	Interval 4
22 22 22	27 27 28	31 31 31 40	55 55

i les mitjanes i distàncies corresponents són les següents:

Interval	Mitjana	Distància mitjana (error de l'interval)
1	22	0
2	27,333	0,444
3	33,25	4,5
4	55	0

Es pot veure fàcilment que la configuració final obtinguda és: ⇒∧

Interval 1	Interval 2	Interval 3
22 22 22 27 27 28	31 31 31 40	55 55

Altres variants del mètode encara uniformitzarien més el resultat assignant a cada interval el valor majoritari.

3.2.2. Mètodes supervisats

Els mètodes que hem discutit fins ara no estaven dirigits a cap tasca concreta: és a dir, es podrien utilitzar tant per a classificació com per a agregació. Ara bé, en el cas de la classificació ens interessa assegurar que dins de cada interval generat els valors de les diverses classes que hi pot haver es distribueixen tan uniformement com és possible. Aquesta és una problemàtica que es torna a trobar a un altre nivell en parlar d'arbres de decisió. Algunes de les solucions que s'han aportat són comunes tant al problema de la discretització com a alguns passos de la construcció d'arbres de decisió.

Vegeu els arbres de decisió en el mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

Mètode *chi merge*

Una bona discretització de dades hauria de mostrar la propietat que dins de cada interval estiguessin representades totes les classes de manera semblant, i molt diferent respecte als altres intervals. En termes de freqüències, això voldria dir que dins un interval caldria esperar que la freqüència d'aparició de valors de cada classe fos semblant, i molt diferent de la d'altres intervals. Si la primera condició no es compleix, vol dir que tenim un interval massa heterogeni i que caldria subdividir-lo. Si no es compleix la segona, vol dir que l'interval és molt semblant a algun altre interval adjacent. El que caldria, llavors, seria unir-los. Una manera de mesurar aquestes propietats és recorrent a l'estadístic χ^2 per a esbrinar si les freqüències corresponents a dos intervals són significativament diferents (i, per tant, ja està bé que els intervals estiguin separats), o bé que no ho són i cal unir-los.

Lectura recomanada

Consulteu el mètode *chi merge* en l'article següent:

R. Kerber (1992). "Chimerge: Discretization of Numeric Attributes". *Proceedings of the 10th National Conference on Artificial Intelligence* (pàg. 123-127).

Recordem que l'estadístic es calcula mitjançant aquesta expressió:

$$x^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

on:

- n és el nombre d'intervals que cal comparar entre si; per a simplificar, farem $n = 2$.
- k és el nombre de classes que hi ha.
- N_{ij} és el nombre de valors de la classe j presents en l'interval i.

En aquest mètode resulta crític situar el nivell de significació de χ^2 prou adequat per a evitar situacions a les quals no es vol arribar. Per exemple, si el valor

és massa alt, obligarem a efectuar moltes unions d'intervals consecutius; per contra, un valor baix crearà un nombre elevat d'intervals de poca amplitud.

La idea és fusionar intervals fins que cada interval mostri un nivell de χ^2 adequat. Alternativament, es pot prefixar un nombre mínim o màxim d'intervals que cal construir.

L'algorisme *chi merge* consta dels passos següents:

1) Ordenar els valors de l'atribut $x_1, \ldots, x_m$.

2) Fixar el valor de significació de χ^2, s.

3) Fixar els n graus de llibertat per a la distribució de χ^2; $n = k - 1$, on k és el nombre de classes.

4) Obtenir una primera discretització en què cada valor és el seu propi interval: $Interval_1, \ldots, Interval_m$.

5) Repetir els passos següents fins que per a tot parell $Interval_i$, $Interval_j$ adjacents $significació(Interval_i) > s \wedge significació(j) > s$:

a) Calcular χ^2 per a cada parell d'intervals adjacents $Interval_i$, $Interval_j$.
b) Trobar el parell d'intervals amb χ^2 més petit $Interval_{imin}$, $Interval_{jmin}$.
c) Fusionar els intervals.

Aquest mètode té l'avantatge de fer servir un estadístic que no depèn del nombre de classes o exemples. En general, es comporta correctament sense introduir punts de tall que no separin intervals bons; en aquest cas sense proposar punts de talls intermedis.

El problema és que exigeix conèixer quin és l'atribut de classe. També és un algorisme massa local: només considera els intervals adjacents dos a dos.

Mètodes basats en mesures d'entropia

Aquest mètode parteix de la mateixa idea que hi ha darrere els algorismes de construcció d'arbres de decisió basats en mesures d'informació. En efecte, en el cas dels arbres, es tractava de trobar particions del conjunt original de dades que fossin internament tan homogènies com es pogués, i respecte a la resta de particions, tan diferents com fos possible. Aquestes són les mateixes característiques dels intervals d'una bona discretització. Com ho podem aprofitar, doncs?

Lectura recomanada

Consulteu els mètodes basats en mesures d'entropia en l'obra següent:

U.M. Fayyad; K.B. Irani (1993). "Multi-interval Discretization of Continous Valued Atrtibutes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pàg. 1022-1027).

En el cas dels arbres de decisió que treballen sobre variables contínues per crear un node de decisió sobre un atribut determinat, es vol trobar el valor de l'atribut que permet de separar millor les dades. D'aquesta manera s'estableix un

límit o llindar. Les observacions que per a l'atribut considerat presenten un valor més petit en el llindar s'assignen al subarbre esquerre, i les altres, al dret.

La idea general del mètode és la següent:

1) Ordenar els valors de l'atribut continu X_i que es consideri: $x_1, ..., x_m$.

2) Crear el conjunt de punts de tall candidats:

a) Considerar com a llindar possible el punt mitjà, x_{mig}, entre dos valors consecutius de X_i: $x_{mig} = (x_i + x_{i+1})/2$:

$$candidats = \{x_{mig}, \ldots, x_{(1+m)/2}\}.$$

3) Per a cada $x_i \in candidats$ executar els passos següents:

a) Establir C_{i1} = {observacions C_j de la base de dades tals que el valor de $X_i < x_i$}.

b) Establir C_{i2} = {observacions C_j de la base de dades tals que el valor de $X_i \geq x_i$}.

c) Calcular $E_1 = Entropia(C_{i1})$.

d) Calcular $E_2 = Entropia(C_{i2})$.

4) Escollir x_i, tal que $Entropia(C_{i1} \cup C_{i2})$ sigui mínima.

Podem utilitzar aquí els conceptes d'entropia i entropia associada en una partició que també esmentem en parlar d'arbres de decisió, semblantment a la manera en què hem definit l'entropia de classe en discutir els arbres de decisió.

Vegeu els conceptes d'entropia i entropia associada en el subapartat 2.1 del mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

En efecte, suposem que hem escollit un valor de tall x_t. Si tenim un conjunt inicial de valors de l'atribut X, el punt de tall el divideix en dos: els que són inferiors a x_t (que denotem per X_1) i els que són superiors a x_t (que denotem per X_2). Suposant que hi ha k classes, $C_1, ..., C_k$, i que la proporció de valors de X que corresponen a la classe C_i sigui $P(X, C_i)$, llavors l'entropia de la classe C_i per a un conjunt de valors X_i és la següent:

$$I(X_i) = -\sum_{i=1}^{k} P(C_i,X_i)\log_2 P(C_i,X_i)\,.$$

Si definim X com el conjunt de valors de l'atribut corresponent, X_1 com el conjunt de valors inferiors al punt de tall i X_2 com el conjunt de valors superiors, l'entropia de classe deguda a l'elecció d'un punt de tall x_t és la següent:

$$I(X_i,x_t) = \frac{|X_1|}{|X|} I(X_1) + \frac{|X_2|}{|X|} I(X_2)\,.$$

Hi ha dues possibilitats:

- Dividir els valors de l'atribut X en un sol punt de tall que faci que l'entropia de les particions sigui mínima.

- Dividir el conjunt de valors pel punt de tall que tingui l'entropia mínima (a menys que tots els valors del conjunt siguin iguals o corresponguin a la mateixa classe).

Cal aclarir el criteri que s'utilitza per a definir el final d'aquest procés de partició repetida sobre el punt d'entropia mínima. Hem de parar en el moment en què tinguem intervals prou homogenis. Com caracteritzem aquest moment?

Lectures complementàries

Trobareu més informació sobre el criteri de la longitud de la descripció en les obres següents:

J.R. Quinlan (1989). "Inferring Decision Trees Using the Minimum Description Length Principle". *Information and Computation* (núm. 80, vol. 3. pàg. 227-248).

M. Rissanen (1985). "The Minimum Description Length Principle". A: S. Kotz; N.L. Johnson (ed.). *Encyclopedia of Statistical Sciences* (vol. 5.). Nova York: John Wiley & Sons.

Un criteri molt útil i que dóna bons resultats per a saber en quin moment cal parar el procés de partició és el de minimització de la longitud de la descripció. Aquest concepte molt general es pot resumir dient que el millor model (M) és, donat un conjunt de dades (D), el que minimitza la grandària del model (és a dir, el que ens dóna el model més compacte) i, simultàniament, també requereix el mínim d'informació (dades) per a construir-lo. Aquesta intuïció es formalitza tenint en compte que ambdues coses, el model i les dades, es poden codificar mitjançant bits (unitats d'informació). Per tant, donat un model i unes dades, es tracta de minimitzar la longitud de la codificació.

En altres paraules, donades unes dades D, un model M i un esquema de codificació C, llavors la **longitud de codificació** és la següent:

$$L_c(D) + L_c(M|D);$$

és a dir, la longitud de codificació és la longitud de la codificació de les dades observades segons l'esquema de codificació escollit, més la longitud de la codificació del model M, tenint en compte que procedeix del conjunt de dades D.

Quin és el model M i quines són les dades en el nostre problema de discretització?

La codificació d'algun tipus d'informació es fa mitjançant bits. Es pot demostrar que el nombre de bits necessaris per a codificar un valor x_i és el logarisme en base 2 de la seva probabilitat (això és, de fet, una mesura de la informació d'aquest valor).

En el nostre cas, el model és el punt de tall, les dades, els valors i les classes associades. Volem comparar la grandària de codificació del model i la quantitat de dades que necessitem per a especificar-lo. Si no decidim introduir una partició en aquest punt, si mantenim l'interval tal com està, llavors hem de

descriure les dades resultants utilitzant les etiquetes de classe de cada valor. Si introduïm un punt de tall, llavors hem de codificar el següent:

- El valor del punt de tall: necessitem $\log_2(N-1)$ per a codificar-lo (on N és el nombre de valors que hi ha).
- Les classes dels valors per sota del punt de tall.
- Les classes dels valors per damunt del punt de tall.

Depenent de l'homogeneïtat dels intervals resultants, pagarà la pena o no introduir un punt de tall. Si la codificació obtinguda introduint el punt de tall és més curta que sense fer-ho, la introduirem; si no ho és, no la introduirem.

Si hi ha exactament el mateix nombre de valors que pertanyen a la classe 1 que els que pertanyen a la classe 2, llavors codificar cada valor costa aproximadament 1 bit. Ara bé, si introduïm el punt de tall de manera que per sota tots els punts pertanyen a una classe, i per sobre, a l'altra, llavors codificar cada valor té un cost molt proper a zero.

La partició que genera un punt de tall determinat és vàlida si aporta un guany d'informació que supera un llindar determinat. No entrarem a precisar com es deriva aquesta fórmula, però tenim que, si el punt introdueix una partició en dos intervals i definim els elements següents:

- N és el nombre de valors total.
- L'atribut del qual considerem valors és X.
- k és el nombre de classes que cal considerar.
- El nombre de classes diferents associades als valors del primer interval X_1 (el que quedaria per sota del punt de tall) és k_1.
- El nombre de classes diferents associades als valors del segon interval X_2 (el que quedaria per damunt del punt de tall) és k_2.
- L'entropia del conjunt de valors inicial, sense partir, és I.
- L'entropia del conjunt que queda per sota del punt de tall és I_1.
- L'entropia del conjunt que queda per damunt del punt de tall és I_2.

Llavors l'expressió del **guany d'informació** sobre un conjunt de N valors de l'atribut X introduïda per un punt de tall X_t que parteix el conjunt original en dos subconjunts X_1 i X_2 és la següent:

$$Guany\ (X,\ x_t,\ X_1,\ X_2) = \frac{\log_2(N-1)}{N} + \frac{\log_2(3^k-2) - kI + k_1I_1 + k_2I_2}{N}.$$

Observeu que la primera part d'aquesta fórmula indica la quantitat d'informació necessària per a codificar el valor del punt de tall; la segona part indica la quantitat d'informació que es necessita per a codificar a quines classes pertanyen els valors de l'interval inferior i les del superior.

Lectura complementària

Trobareu la derivació de l'expressió del guany d'informació en l'obra següent:

U.M. Fayyad; K.B. Irani (1993). "Multi-interval Discretization of Continous Valued Atrtibutes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pàg. 1022-1027).

Es pot minimitzar la longitud de descripció total si a cada pas de discretització s'escull el punt que maximitza el guany d'informació. Es pot demostrar que tot punt que minitmizi l'entropia de la partició ha de ser un punt de tall i, no solament això, sinó que s'assegura que els punts de tall escollits així sempre estan entre valors que pertanyen a classes diferents.

En conjunt, aquest és un comportament molt adequat i la combinació de la discretització basada en l'entropia, conjuntament amb el criteri de no introduir particions o aturar el procés de partició quan no augmenta el guany d'informació, dóna discretitzacions molt bones.

Exemple de mètode basat en mesures d'entropia

Vegem un exemple senzill d'aplicació d'un mètode basat en mesures d'entropia.

Aquí tenim informació sobre el valor de l'atribut *Edat* per a un conjunt de socis d'HyperGym que pertanyen o bé a la classe 1 o bé a la classe 2 (per a l'exemple és irrellevant, però remarquem que la classe 1 correspon als socis que demanen entrenador personal, i la classe 2, als que no en demanen):

Valors	22	24	28	35	38	42	43	47	49	55
Classe	1	2	2	12 21	2 2	1 1	2	1	1	2

Com es pot veure, hi ha valors que són menys homogenis que d'altres. Per exemple, el 35 acumula dues observacions de la classe 2 i dues observacions de la classe 1.

Quants llocs possibles pot ocupar el punt de tall? Recordem que inicialment s'escull com a punt de talls el punt mitjà entre dos valors consecutius $(x_i+x_{i+1})/2$, que són els següents:

23	26	31,5	36,5	40	42,5	45	48	52

Hem de veure la qualitat de cada un i hem de calcular el contingut d'informació dels intervals que generaria cada punt de tall. Per exemple, el 23 deixa a la seva esquerra un únic valor de classe 1 i a la seva dreta vuit valors de la classe 2 i sis valors de classe 1. Quina quantitat d'informació té aquesta partició?

$$I(Edat,23) = \frac{1}{10}I(Edat,Edat_1) + \frac{9}{10}I(Edat,Edat_2)\ ,$$

on $Edat_1$ correspon al conjunt de punts inferiors a 23, i $Edat_2$, al conjunt de punts amb valors superiors a 23.

Calculem els valors corresponents:

$$I(Edat,Edat_1) = -P(C_1,Edat_1)\log_2 P(C_1,Edat_1) - P(C_2,Edat_1)\log_2(C_2,Edat_1) =$$
$$= -(1/1)\cdot\log_2 1 - (0/1)\cdot\log_2 0 = 0.$$

$$I(Edat,Edat_2) = -P(C_1,Edat_2)\log_2 P(C_1,Edat_2) - P(C_2,Edat_2)\log_2(C_2,Edat_2) =$$
$$= -(6/14)\cdot \log_2 (6/14) - (8/14)\cdot \log_2 (8/14) =$$
$$= 0{,}42 \times 1{,}22 + 0{,}57 \times 0{,}80 = 0{,}98.$$

Per tant,

$$I(Edat,\ 23) = \frac{1}{10} I(Edat,\ Edat_1) + \frac{9}{10} I(Edat,\ Edat_2) = 0{,}88.$$

Si repetíssim el procés, trobaríem el conjunt de punts de tall final.

4. Reducció de dimensionalitat

Un cop tenim les dades en el format adequat per al tipus de model que es vol obtenir i el mètode per a construir-lo, encara és possible aplicar una sèrie d'operacions nova per a assegurar dos objectius: la reducció del nombre d'atributs que cal considerar, i la reducció del nombre de casos que cal tractar, assegurant que es mantindrà la qualitat del model resultant.

El motiu per a efectuar la reducció acostuma ser doble:

1) El programa de construcció del model escollit no pot tractar tantes dades com la quantitat d'aquestes de què disposem.

2) El programa les pot tractar, però el temps requerit per a construir el model és inacceptablement llarg (dies d'UCP, per exemple).

Hi ha algun altre motiu una mica més subtil que ens diria que no sempre més dades vol dir que tindrem un model millor. En efecte, quan es consideren moltes dades, cal ajustar el model a més i més casos, amb la qual cosa ens perdem detalls que poden ser importants.

Amb les **operacions de reducció de dimensionalitat** que comentarem, volem mantenir les característiques de les dades inicials, perquè eliminem dades que no són necessàries.

Altres mètodes alteren les dades, bé per fusió d'atributs, bé per canvi de valors, de manera que costa recuperar les dades originals per a interpretacions futures.

4.1. Reducció del nombre d'atributs

La reducció del nombre d'atributs consisteix a trobar un subconjunt dels atributs originals que permeti d'obtenir models de la mateixa qualitat que els que s'obtindrien utilitzant tots els atributs. Aquest problema s'anomena **problema de la selecció òptima d'atributs**.

La selecció d'atributs...

... és un procés que intenta seleccionar el subconjunt dels atributs originals que mantindrà la qualitat dels models que es puguin extreure a partir de les dades.

El sentit comú ens imposa un procediment que sembla el més evident, però que té desavantatges clars. Es tracta d'anar generant tots els subconjunts possibles de $n - 1$ atributs, escollir-ne el millor, generar els de $n - 2$, etc., i

així fins que no hi hagi cap millora o es comenci a degradar la qualitat dels models.

Només cal considerar quants subconjunts es generen per mitjà d'aquest procés per a adonar-se que no és pas pràctic. Per tant, s'han desenvolupat mètodes indirectes que, a partir de mesures sobre les propietats dels atributs o de les relacions que hi ha entre aquests, ens permetin de detectar quins són irrellevants, o redundants, o equivalents.

4.1.1. Mètodes de selecció d'atributs

Els mètodes de selecció d'atributs consisteixen a saber quin subconjunt d'atributs pot generar un model amb la mateixa qualitat que el que es podria extreure amb tot el conjunt d'atributs inicials. Es tracta de veure quines mesures sobre els atributs es poden utilitzar per a assegurar que es pot introduir una reducció sense pèrdua d'informació.

En part, tot es redueix a saber fins a quin punt determinats valors pertanyen a la distribució que es correspon amb les dades. Si, per exemple, un atribut concret no té correspondència amb la distribució, el podríem descartar com a erroni o irrellevant. El problema és que no coneixem la distribució de la població i no podem fer aquest tipus de tractament. Només podem aproximar-nos-hi a força d'estimar-ne els paràmetres corresponents.

Ens centrarem en els mètodes de selecció d'atributs desenvolupats per a classificació. 

> Els **mètodes de selecció d'atributs desenvolupats per a classificació** pretenen obtenir un model que, amb menys atributs, tingui el mateix poder de classificació que el model que s'obtindria a partir del conjunt original d'atributs.

Conegut el valor C_i que pot prendre l'atribut classe dins un conjunt finit de valors de classe $\{C_1, \ldots, C_n\}$, es tracta de veure, per exemple, si dos atributs diferents, X_1 i X_2, mostren valors de la mitjana que són molt iguals o no per a un atribut classe determinat. Si, efectivament, són molt iguals, llavors tots dos atributs es poden considerar igualment predictius per a aquella classe i l'un pot ser substituït per l'altre. En canvi, en cas que els valors siguin molt diferents, som davant un atribut interessant, ja que permet de discriminar millor si una observació ha d'anar a una classe o a una altra. Evidentment, aquest tipus d'anàlisi s'ha d'estendre a tots els atributs que cal considerar, i ponderar adequadament la importància de cada un.

Aquí mostrem una aproximació senzilla a aquest problema basada en l'anàlisi dels paràmetres estadístics més corrents.

Prova de significació

La prova de significació compara si un atribut X_i és realment rellevant o no a partir dels seus valors.

Una manera de fer-ho, prou clàssica i procedent de l'estadística, consisteix a fer un test de significació per als seus valors mitjans en relació amb les diverses classes que hi ha. Si l'atribut té el mateix valor per a cada classe, llavors el podríem considerar poc rellevant per a la classificació i viceversa. Simplificarem la presentació suposant un problema senzill de classificació en què només hi pot haver dues classes, la 1 i la 2. L'anàlisi es pot estendre fàcilment a problemes de multiclassificació.

La prova de significació consisteix a fer una prova d'hipòtesi per a esbrinar si la mitjana de l'atribut X_i en els casos observats que tenen com a etiqueta la classe 1 és igual a la mitjana dels casos corresponents a la classe 2 o és diferent. Per tant, es tracta de fer una prova d'hipòtesi sobre la igualtat o no de la mitjana de l'atribut per a cada classe:

$$est = \sqrt{\frac{\text{var}(X_{i1})}{n_1} + \frac{\text{var}(X_{i2})}{n_2}}$$

on X_{i1} és el valor de l'atribut X_i en les observacions que corresponen a la classe 1; X_{i2} és l'equivalent per a les observacions que corresponen a la classe 2; n_1 és el nombre de casos corresponents a la classe 1, i n_2, el dels corresponents a la classe 2; *var* és la variància de X_i.

Es comparen els dos atributs i es veu si el seu nivell de significació realment indica que les diferències corresponen a una cosa més que una senzilla variació aleatòria. Si no hi ha diferència, es pren la decisió d'eliminar un dels dos atributs.

Per al cas d'haver de comparar més de dos atributs, es pot reiterar aquesta comparació d'atributs dos a dos.

Cal tenir en compte que la suposició inicial d'aquesta manera de procedir és que els atributs són independents entre si.

Exemple de prova de significació

Aquí podem veure un exemple senzill d'anàlisi de prova de significació. Suposem l'atribut *Edat* que presenta els valors següents per a les classes 1 i 2:

Classe 1	Classe 2
22	23
34	34

Classe 1	Classe 2
23	23
21	56
34	67
23	87
21	56
34	23
65	59
12	77

La mitjana de l'edat a la classe 1 és de 28,9 i a la classe 2 és de 50,5. Sembla que tenim un possible atribut interessant per a discriminar entre els elements de les classes 1 i 2.

Obtenim la diferència entre mitjanes, 21,6. El valor de *est* és 8,3, la qual cosa ens dóna un valor de 2,6. Consultant les taules de significació, es pot comprovar que la diferència és significativa i, en principi, sembla que l'atribut *Edat* pot ser rellevant per a discriminar entre una classe i l'altra.

Anàlisi de grups d'atributs

El mètode de selecció d'atributs anterior, com hem remarcat, només seria realment aplicable sobre atributs independents entre si. En la majoria dels casos ens dóna una idea aproximada de la rellevància final de cada atribut i dóna peu a decisions prou encertades. De totes maneres, en la realitat la suposició d'independència pot ser una mica excessiva. Per aquest motiu s'ha de recórrer a altres mètodes que no en fan ús.

Pot passar que determinats atributs siguin útils quan es consideren aïlladament, però deixin de ser-ho en conjunció amb d'altres. Per tant, interessa detectar quines agrupacions d'atributs $X_1, \ldots, X_k$ són realment rellevants i si ho continuen essent quan es deixa de considerar un o més atributs X_j, $1 < j < k$.

Moltes vegades un grup d'atributs pot ser substituït per un altre atribut (pertanyent al conjunt o no) sense alterar en absolut la qualitat del model resultant. Per a fer-ho, cal estudiar conjuntament els grups d'atributs.

Suposem que treballem sobre un conjunt de m atributs, $X_1, \ldots, X_m$. Cada atribut es pot considerar una variable aleatòria. Presos conjuntament, els m atributs segueixen una distribució de probabilitat conjunta $P(X_1, \ldots, X_m)$. Es pot fer la suposició que aquesta distribució de probabilitat segueix una distribució normal multivariant.

Una distribució d'aquest tipus es pot caracteritzar mitjançant un vector que recull les mitjanes de totes les variables $\boldsymbol{X}$ i per la matriu de covariàncies de les mitjanes $\boldsymbol{C}$, que és una matriu $m \times m$.

Cadascun dels elements c_{ij} de la matriu $\boldsymbol{C}$ reflecteix la relació entre els atributs X_i i X_j segons la fórmula següent:

$$c_{ij} = \frac{1}{n}\sum_{k=1}^{n}[(x_{ki}\,\bar{x}_i)\cdot(x_{kj}\,\bar{x}_j)]\text{, on:}$$

- n és el nombre total d'observacions que hi ha en el conjunt de dades.
- x_{ki} és el valor de l'atribut x_i en l'observació k-èsima.
- x_{kj} és el valor de l'atribut x_j en l'observació k-èsima.
- $\bar{x}_i$ és la mitjana de l'atribut x_i sobre tot el conjunt de dades.
- $\bar{x}_j$ és la mitjana de l'atribut x_j sobre tot el conjunt de dades.

Els elements de la diagonal de la matriu C, c_{ii}, corresponen a les variàncies de cada atribut i. Els termes de fora de la diagonal de la matriu, c_{ii} amb $i \neq j$, indiquen la correlació entre els atributs corresponents x_i i x_j. A partir de la informació de les correlacions i variàncies, es pot detectar l'existència d'atributs redundants.

En efecte, suposem que tornem a estar en una situació de classificació que admet dues classes, la 1 i la 2. Ens interessa tornar a trobar si hi ha diferències significatives entre les característiques dels atributs respecte a una classe i l'altra. La diferència ara és que, en comptes de fer aquesta prova atribut per atribut o de dos en dos, com que no suposem independència entre els atributs, establim una prova simultàniament sobre totes les mitjanes de tots els atributs.

Donada la classe 1, podem indicar per $\overline{X}_{c1}$ el vector que conté les mitjanes de tots els atributs calculades a partir de les observacions corresponents a la classe 1. Indiquem per $\overline{X}_{c2}$ el vector corresponent per a la classe 2.

La nostra intuïció ens diu que, si aquests dos vectors són molt semblants entre si, llavors el grup de n atributs que considerem no és gaire rellevant per a discriminar entre una classe i l'altra.

Per a establir la similitud entre els dos vectors, podem recórrer un altre cop al concepte de *distància*, encara que en aquesta ocasió la seva expressió ha de tenir en compte les característiques de l'espai on calculem aquestes distàncies. Així estenem la prova de significació entre les diverses mitjanes de tots els atributs simultàniament. Definim la distància següent:

$$Dist = (\overline{X}_{c1} - \overline{X}_{c2})(C_1 + C_2)^{-1}(\overline{X}_{c1} + \overline{X}_{c2})^T$$

on:

$$(\overline{X}_{c1} - \overline{X}_{c2})^T$$

el vector transposat del resultat de sumar els vectors de mitjanes per la classe 1 i 2, i:

$$(C_1 + C_2)^{-1}$$

és la inversa de la suma de la matriu de covariàncies.

Si els atributs són independents entre si, llavors tots els elements que no siguin a la diagonal de la matriu inversa de covariàncies són zero i els valors de la diagonal són:

$$\frac{1}{var_{xi}}$$

per a cada atribut X_i. Hem de trobar els atributs per als quals es maximitzi la distància; és a dir, aquells per als quals cadascun dels components del vector mostri els valors màxims. Cada component del vector de distàncies és el valor donat per l'expressió:

$$\frac{(\overline{X}_{iC1} - \overline{X}_{iC2})^2}{var_{XiCi} + var_{XiCi}}$$

on:

$\overline{X}_{iC2}$ és la mitjana de X_i per a la classe C_1 (anàlogament per a C_2).

var_{XiCi} és la variància de X_i per a la classe C_1 (anàlogament per a C_2).

Per tant, el problema consisteix a trobar un conjunt de *k* atributs amb *k* inferior al nombre d'atributs total present en el domini *m*, tals que el valor corresponent a aquesta fórmula quedi maximitzat. Per a trobar-lo, cal fer el següent:

1) Trobar els subconjunts de *k* atributs dins el conjunt de *m* atributs inicial.

2) Avaluar-los.

3) Quedar-se amb el subconjunt que mostri un valor de *Dist* més gran.

El nombre de combinacions és evidentment molt alt, però hi ha algorismes d'optimització que permeten d'afrontar-lo.

Una alternativa interessant a l'anterior consisteix a començar amb un únic atribut, i anar afegint-hi atributs, mesurant el valor de *Dist*. La idea es basa a incrementar el nombre d'atributs del conjunt inicial de manera que aquest valor millori.

A cada pas de la iteració també s'ha de veure si cal eliminar algun dels atributs considerats fins aquest moment perquè no aporta prou guany. La idea consis-

teix a introduir a cada pas els atributs que maximitzin la qualitat i eliminar-hi els que la minimitzin. La condició de final de la iteració és quan ja no s'hi afegeix ni s'hi elimina cap atribut, o bé quan s'ha superat la qualitat del conjunt total de n atributs en una proporció prou significativa (per exemple, entre el 75% i el 95%).

A continuació presentem l'esquema de l'algorisme associat a aquest mètode d'anàlisi de grups d'atributs. Donats $X_1, \ldots, X_n$ atributs que descriuen un domini i donat un llindar de qualitat λ, l'algorisme és el següent:

Seleccionar un atribut X_i.
Conjunt_Actual = $\{X_i\}$.
No_final = Cert
Mentre *No_final* **fer**
  Resta = $\{X_1, \ldots, X_n\}$ – *Conjunt_Actual*
  Màxim = 0; *Atribut_màxim* = {}
    Per a tot $X_j \in$ *Resta* **fer**
    Si Qualitat (*Conjunt_ Actual* $\cup$ $\{X_j\}$) > *Màxim*
    aleshores *Atribut_màxim* = $\{X_j\}$
    Màxim = *Qualitat*(*Conjunt_Actual* $\cup$ *Atribut_màxim*)
    fsi
  fper
  Conjunt_Actual = *Conjunt_Actual* $\cup$ {*Atribut_màxim*}
  Mínim = *Qualitat*(*Conjunt_Actual*);
  Per a tot $X_j \in$ *Conjunt_Actual* **fer**
    Si *Qualitat*(*Conjunt_Actual* – $\{X_j\}$) < *Mínim*
      aleshores *Atribut_mínim* = $\{X_j\}$
      Mínim = *Qualitat*(*Conjunt_Actual* – *Atribut_mínim*)
    fsi
  fper
  Conjunt_Actual = *Conjunt_Actual* – {*Atribut_mínim*}
fmentre

4.1.2. Fusió i creació de nous atributs: anàlisi de components principals

El mètode de selecció d'atributs que hem comentat fins aquest punt és l'equivalent a eliminar en la forma tabular de les dades tota una columna: la resta de columnes queda inalterada completament.

Exemple de selecció d'atributs

Si decidim que, en el cas d'Hyper-Gym, per a una determinada operació de construcció de models l'atribut *Renda* resulta irrellevant, l'eliminarem però no hi n'introduirem cap de nou ni modificarem els valors dels atributs que restin en la base de dades.

Una alternativa a aquest mètode consisteix a fusionar atributs, introduint si es vol atributs "híbrids", de manera que se'n creïn de nous que corresponguin a valors nous. Igualment podem influir en la resta d'atributs modificant-ne els valors.

Donats n atributs, X_i, ..., X_n, es pot convertir en un atribut nou X_{n+1} a força d'aplicar una combinació lineal de pesos als atributs X_1, ..., X_n.

$$X_{n+1} = \sum_{i=1}^{n} X_i \cdot w_i$$

on w_i, $1 < i < n$ és un pes, un factor numèric, que indica en quin grau l'atribut X_i contribueix a l'atribut nou.

Evidentment, la combinació lineal de tots els atributs ens portaria a una reducció extrema d'atributs en la qual n atributs quedarien reemplaçats per un de sol, el nou atribut X_{n+1}. Això, en general, no ens garantirà comportaments correctes.

El **mètode dels components principals** efectua fins a n transformacions. A cada etapa es fa una única transformació. En cadascuna d'aquestes etapes o transformacions s'aplica un vector **W** de pesos diferent del que actua com a component principal. Un cop obtingudes n transformacions, s'escullen les k que tenen una millor qualitat.

Suposant un conjunt de dades original definit com un vector amb tants elements com atributs té el domini **D**, podem establir la relació entre aquest conjunt original i el que resulta després d'aplicar i transformacions, $\mathbf{D_i}$, mitjançant una matriu **P** que recull els diversos components principals. En efecte, es pot veure que es compliria aquesta relació:

$$\mathbf{D = DP}$$

on cada columna de **P** és un component principal format per m pesos. El resultat de multiplicar, en la transformació i-èsima, l'observació k-èsima per la columna j-èsima de la matriu **P** és el valor del nou atribut j per a l'observació k-èsima.

Per a trobar els pesos dels components principals, primer es normalitzen les dades per estandardització. El primer component principal és la línia que s'ajusta millor a les dades. Això es pot mesurar utilitzant la distància euclidiana mínima:

$$D_{eu} = \sum_{i,j} (D_{ij} - w_j D_{ij}) .$$

Es pot demostrar que l'atribut que s'ha generat en aquesta transformació és el que té variància màxima. En principi, sembla que els atributs que tenen aquesta propietat haurien de ser bons candidats per a separar classes.

4.2. Mètodes de reducció de casos

Una altra manera de reduir la dimensionalitat és eliminar el nombre de casos que cal considerar. Aquest problema ja s'ha afrontat tradicionalment des de l'estadística i l'anàlisi de dades.

El **mètode de reducció de casos** consisteix a trobar una mostra, un subconjunt del conjunt original de casos, que doni un comportament semblant.

El **mostreig aleatori** intenta extreure el subconjunt de casos adequat. Considerarem dos mètodes de mostreig:

a) Obtenció incremental de la mostra: es tracta d'escollir un conjunt inicial de mostra i anar afegint-hi casos fins que s'obté un nivell de qualitat acceptable prefixat.

b) Consens de mostres: es tracta d'obtenir diversos models per a *N* conjunts de mostres de dimensions relativament petites i intentar fusionar els models obtinguts.

És important adonar-se que tots dos mètodes requereixen aplicar el programa de construcció de models de què es tracti.

Obtenció incremental de mostres

L'obtenció incremental de mostres consisteix a començar per un conjunt petit de casos escollits aleatòriament; per exemple, començar amb un subconjunt inicial corresponent al 10% dels casos. Llavors, cal avaluar el procés i anar-lo repetint amb percentatges de dades progressivament més grans.

Consens de mostres

El consens de mostres consisteix a dividir els *N* casos en un nombre determinat de mostres. A cadascuna s'hi aplica el mètode que cerquem i finalment s'hi aplica un algorisme de consens.

Curiosament, amb aquest mètode s'obtenen millors resultats que aplicant el mètode a tots els casos.

5. Tractament de la manca de dades

Un dels problemes més habituals en el tractament previ de les dades és l'absència de valors per a un atribut determinat.

Dues opcions típiques són les següents:

- Substituir el valor que manca per la mitjana dels valors que presenta l'atribut en les dades.
- Substituir-lo pel valor més freqüent.

Exemples típics de tractament de manca de dades

En el cas de les edats dels clients d'Hyper-Gym, podríem tenir aquesta situació:

22	43	31	22	?	22	45	43	23	88

Les dues solucions típiques serien les següents:

- Substituir el valor que manca per la mitjana dels valors que tenim. En aquesta solució seria 37,3.
- Substituir el valor que manca pel més freqüent. En aquest cas seria 22.

Cada alternativa pot arribar a donar resultats diferents, especialment quan el nombre de valors que manquen és alt. En general, aquests mètodes massa senzills porten a introduir biaix a les dades.

Acostuma a passar que els casos per als quals no s'ha recollit una observació reben algun valor especial, per exemple –1 o 999. És important no tractar-los com un valor numèric més, sinó com el que realment són: absència d'informació respecte a un atribut en un cas observat.

Altres solucions inclouen el següent:

- No tractar els atributs que presenten un nombre alt de valors no observats.
- No tractar els casos que presenten un atribut no observat.
- No tractar els casos que presenten més d'un atribut no observat.

També és important saber si es tracta de manca d'observació o bé que aquell atribut no es considera significant o rellevant. En el segon cas, no cal tractar l'atribut en absolut.

Resum

La **fase de preparació de dades** introdueix modificacions sobre els valors presents en el conjunt de dades per dos motius, principalment:

1) Exigències de format dels mètodes de mineria de dades, com el cas de les xarxes neuronals (valors d'entrada entre 0 i 1) o de certs mètodes d'aprenentatge que només admeten determinats tipus de variables.

2) Necessitat de simplificació dels càlculs que ha d'efectuar el mètode que s'apliqui, bé reduint el nombre de valors dels atributs, bé reduint el conjunt d'atributs inicials o reduint el conjunt de casos que cal tractar.

Hem vist les característiques dels atributs segons l'escala, l'ordre i la grandària del seu rang. Hi ha diversos **tipus de transformacions** segons la divisió en variables contínues, discretes, ordinals i nominals.

S'han descrit les **normalitzacions**, que porten els valors sobre un rang estandarditzat i comparable, i també les discretitzacions, que redueixen el nombre de valors dels atributs partint-los en diversos intervals que defineixen punts de tall sobre els quals es poden efectuar comparacions.

Els **mètodes de discretització** es poden aplicar de manera general, sense conèixer quina mena de mètodes de mineria de dades s'aplicaran sobre les dades discretitzades. Aquesta mena de mètodes s'anomenen **mètodes no supervisats**.

Per contra, els **mètodes supervisats** parteixen del coneixement de quina mena de tasca s'ha de desenvolupar amb el model que resulti de tractar les dades discretitzades. Hem revisat dos mètodes supervisats per a discretitzar dades per a classificació: *k-means* i mètodes basats en entropia de classe. Aquests darrers ens han permès d'introduir tres conceptes importants:

a) D'una banda, el concepte de **mesura de distància**, que també fem servir en parlar d'agregació.

Vegeu l'agregació en el mòdul "Agregació" d'aquesta assignatura.

b) D'una altra, l'**entropia** com a mesura de l'homogeneïtat d'una classe (que tornem a trobar en parlar d'arbres de decisió).

Vegeu els arbres de decisió en el mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

c) Finalment, també hem presentat el **principi de la mínima longitud de descripció** (que s'aplica de manera general per a la construcció i avaluació de molts tipus de models per a diverses tasques).

També hem presentat mètodes per a reduir el nombre d'atributs que cal utilitzar. Hem après a detectar atributs irrellevants dos a dos i grups d'atributs que poden ser substituïts per un de sol.

Finalment, hem comentat la possibilitat d'utilitzar un conjunt més petit de dades que el conjunt original per a evitar el tractament de volums massa grans d'observacions. Només hem esmentat la possibilitat d'aconseguir mostres incrementalment i de consensuar diverses mostres.

Tots aquests canvis obliguen a assegurar-se que el model final mantingui unes característiques de qualitat bones. 

Activitats

1. Compareu els diversos mètodes de discretització que ofereixen els sistemes:
- WEKA
- SIPINA
- CViz

Quines limitacions imposen?

2. Consulteu l'article original de Fayyad i Irani per tal de veure com es desenvolupa la prova que el criteri MDL és apropiat per a donar una condició de final a la discretització per entropia de classe.

3. Respecte a la primera proposta de projecte que havíeu efectuat en les activitats del mòdul anterior:
a) Estudieu quines necessitats de transformació de dades es podrien donar.
b) Estudieu quines necessitats de discretització es podrien donar.
c) Aporteu al fòrum algun conjunt de dades que necessiteu per a dur a terme el projecte, i comenteu quins mètodes de normalització i discretització hi aplicaríeu.
d) Heu trobat cap atribut irrellevant?

4. Compareu les vostres propostes i dificultats amb les dels vostres companys i companyes; discutiu els avantatges i inconvenients que hi heu trobat.

Lectura recomanada

Trobareu l'article original de Fayyad i Irani en l'obra següent:

U.M. Fayyad; K.B. Irani (1993). "Multi-interval Discretization of Continous Valued Atrtibutes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pàg. 1022-1027).

Vegeu l'activitat 1 del mòdul "El procés de descobriment de coneixement" d'aquesta assignatura.

Exercicis d'autoavaluació

1. Classifiqueu els tipus de variables següents com a contínues, ordinals, discretes o nominals:
a) Horari d'Hyper-Gym.
b) Renda dels socis d'Hyper-Gym.
c) Sexe dels socis d'Hyper-Gym.
d) Professió dels socis d'Hyper-Gym.
e) Dia de la setmana.
f) Anys de pertinença al club.

2. Normalitzeu, si podeu, pel màxim, per diferència i per escala decimal els valors següents:
a) Valors de *X*: –999, 985,34, 234,06 i 2.002,999, on *X* pot prendre valors reals entre 1.000 i 2.003 sota del 1.000.
b) Valors de *X*: 1, 224, 30.000, 154, on *X* pot prendre valors enters entre 0 i 150.000. Heu d'aconseguir que els valors resultants estiguin entre –1 i 1
c) Valors de la variable *X* que recull valors enters corresponents a les rendes dels membres del club Hyper-Gym. El valor mínim que pot prendre és 0, i el valor màxim, 10.000.000. El rang final s'ha de trobar entre 0 i 1 (valors reals).

3. Discretitzeu els conjunts de valors següents per a obtenir tres intervals:

Variable 1	Variable
6.000.000,0	40,0
0,0	30,0
0,0	32,0
0,0	32,0
4.000.000,0	30,0
1.200.000,0	55,0
0,0	32,0
3.500.000,0	40,0
2.500.000,0	60,0
4.000.000,0	41,0
3.500.000,0	36,0
1.800.000,0	59,0
3.200.000,0	33,0

Variable 1	Variable
5.000.000,0	34,0
4.000.000,0	34,0

a) Pel mètode d'igual amplitud d'interval.
b) Pel mètode d'igual freqüència.
c) Pel mètode de la distribució normal.
d) Pel mètode *k-means*.
e) Compareu el resultat:
- Sense transformar les dades.
- Normalitzant els valors de cada variable perquè quedin en el rang (0, 1).
- Amb els resultats que obtingueu amb una discretització amb cinc intervals.

3. Discretitzeu aquest conjunt de dades, que és una mostra de la base de dades IRIS del dipòsit de la UCI:

Classe	Longitud del sèpal	Amplada del sèpal	Longitud del pètal	Amplada del pètal
Iris setosa	5,1	3,5	1,4	0,2
Iris setosa	4,9	3	1,4	0,2
Iris setosa	4,7	3,2	1,3	0,2
Iris setosa	4,6	3,1	1,5	0,2
Iris setosa	5	3,6	1,4	0,2
Iris-setosa	5,4	3,9	1,7	0,4
Iris setosa	4,6	3,4	1,4	0,3
Iris-virginica	6,5	3	5,8	2,2
Iris virginica	7,6	3	6,6	2,1
Iris-virginica	4,9	2,5	4,5	1,7
Iris virginica	7,3	2,9	6,3	1,8
Iris virginica	6,7	2,5	5,8	1,8
Iris virginica	7,2	3,6	6,1	2,5
Iris virginica	6,5	3,2	5,1	2
Iris setosa	5,8	4	1,2	0,2
Iris setosa	5,7	4,4	1,5	0,4
Iris setosa	5,4	3,9	1,3	0,4
Iris setosa	5,1	3,5	1,4	0,3
Iris setosa	5,7	3,8	1,7	0,3
Iris setosa	5,1	3,8	1,5	0,3
Iris versicolor	6,3	2,3	4,4	1,3
Iris versicolor	5,6	3	4,1	1,3
Iris versicolor	5,5	2,5	4	1,3
Iris versicolor	5,5	2,6	4,4	1,2
Iris versicolor	6,1	3	4,6	1,4

- Per *chi merge*.
- Per minimització de l'entropia de classe.

5. Donades les dades següents, determineu els atributs rellevants. Aquestes dades procedeixen d'una simplificació de la qualificació per barris de la ciutat de Boston, als Estats Units,

que intenta relacionar diversos factors del barri on es resideix amb el nivell de renda assolit per una persona.

Nivell de renda (classe)	Índex de criminalitat	Ocupació industrial	Nivell d'òxid nítric a l'aire	Nombre d'habitants per habitatge	Nivell d'escolarització
0-10K	20,0849	18,1	0,7	4,368	20,200001
10-20K	0,17004	7,87	0,524	6,004	15,2
20-30K	0,00632	2,31	0,538	6,575	15,3
30-40K	0,02729	7,07	0,469	7,185	17,799999
40-50K	0,08187	2,89	0,445	7,82	18
30-40K	0,03237	2,18	0,458	6,998	18,700001
30-40K	0,06905	2,18	0,458	7,147	18,700001
20-30K	0,02985	2,18	0,458	6,43	18,700001
20-30K	0,08829	7,87	0,524	6,012	15,2
20-30K	0,14455	7,87	0,524	6,172	15,2
10-20K	0,21124	7,87	0,524	5,631	15,2

6. Amb les mateixes dades de l'exercici d'autoavaluació número 5, comproveu la rellevància dels atributs dos a dos.

Bibliografia

Anderberg, M. (1973). "Cluster Analysis for Applications". *Academic Press*.

Catlett, J. (1991). "On Changing Continuous Attributes into Ordered Discrete Attributes. A: Y. Kodratoff (ed.). *Proceedings of the European Working Session on Learning: Machine Learning* (pàg. 164-178). Springer-Verlag.

Dougherty, J.; Kohavi, R.; Sahami, M. (1995). "Supervised and Unsupervised Discretizations of Continuous Features". *Proceedings of the 12th International Conference on Machine Learning* (pàg. 194-202). Morgan Kaufmann Publishers.

Fayyad, U.M.; Irani, K.B. (1993). "Multi-interval Discretization of Continous Valued Atrtibutes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pàg. 1022-1027).

Hartigan, J.I.; Wong, M.R. (1979). "Algorithm AS136: A *k*-means Clustering Algorithm". *Applied Statistics* (vol. 28, núm. 1, pàg. 100-108).

Huan Liu, H.; Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Maning*. Boston: Kluwer Academic Publishers.

Kerber, R. (1992). "Chimerge: Discretization of Numeric Attributes". *Proceedings of the 10th National Conference on Artificial Intelligence* (pàg. 123-127).

Lebart, L.; Morineau, A.; Fenelon, J.P. (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.

Martín, J.R. (1999). *Discretització de variables contínues i la seva aplicació*. Projecte de final de carrera, Enginyeria en Informàtica. Departament de Llenguatges i Sistemes Informàtics. Barcelona: Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya.

Pyle, D. (1999). *Data Preparation For Data Mining*. Morgan Kaufmann Publishers.

Quinlan, J.R. (1989). "Inferring Decision Trees Using the Minimum Description Length Principle". *Information and Computation* (núm. 80, vol. 3, pàg. 227-248).

Rissanen, M. (1985). "The Minimum Description Length Principle". A: S. Kotz; N.L. Johnson (ed.). *Encyclopedia of Statistical Sciences* (vol. 5.). Nova York: John Wiley & Sons.

Simoudis, E.; Fayyad, U.M. (1997, març). "Data Mining Tutorial". *First International Conference on the Practical Applications fo Knowledge Discovery in Data Bases*. Londres.

Valdés, J.J. (1997). "Fuzzy Clustering and Multidimensional Signal Processing in Enviromental Studies". *Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing*. Aache (Alemanya).