Avaluació de models

Luis Carlos Molina Félix
Ramon Sangüesa i Solé

P03/05054/01040

www.uoc.edu

Índex

Introducció

El procés de mineria de dades, tal com s'ha descrit, no és en absolut lineal ni molt menys té un sol pas. Com s'ha comentat, correspon més a un procés iteratiu en què s'obté un primer model, s'avalua i, segons que aquesta avaluació sigui satisfactòria o no, es dóna per bo i s'acaba el procés, o bé es retorna a un pas anterior, que pot ser la selecció o la preparació de dades.

En aquest mòdul didàctic ens centrem en la problemàtica de l'avaluació del model, precisant-la i introduint els mètodes i les mesures que s'utilitzen més habitualment per a trobar la qualitat dels models obtinguts. 

Tot i que la qualitat d'un model sempre està en relació amb la tasca per a la qual s'ha construït (predicció, classificació, etc.), veurem que hi ha algunes mesures generals independents de la tasca que cal conèixer.

Objectius

Els materials didàctics associats a aquest mòdul permetran que l'estudiant assoleixi els objectius següents:

1. Ser capaç d'aclarir els procediments i les mesures per a establir la qualitat dels models obtinguts a partir d'un procés de mineria de dades.

2. Conèixer les mesures més adequades per a cada tipus de tasca.

3. Establir els procediments de comparació entre els resultats que s'obtenen amb diversos tipus de mètodes de mineria de dades.

1. Avaluació de models

Quan hem d'avaluar un model, partim d'una situació en què ja hem utilitzat un determinat conjunt de dades, hi hem aplicat un mètode de mineria de dades, hem obtingut un primer model i ens trobem en la necessitat de saber si aquest model és bo o no respecte a la tasca que ens hem proposat.

Aquest problema s'estén al fet que hem de comparar el funcionament de diversos mètodes de mineria de dades respecte al mateix conjunt de dades.

Exemple de problema de comparació de models

Hem construït diversos models per classificar clients d'un banc en 'Morosos' o 'No morosos': un model d'arbre de decisió, una xarxa neuronal o una llista de regles de classificació. Volem veure quin model augura un millor rendiment. Evidentment, ens interessa escollir el que asseguri un percentatge millor d'encerts en la classificació.

Pensem un moment què vol dir exactament que un model és bo. Per al cas d'una tasca de classificació, el model serà bo si assigna correctament les etiquetes de classe a cada objecte quan s'aplica a objectes nous que no s'havien utilitzat en la construcció del model classificatori. Per a una tasca d'agregació d'objectes, un model és bo quan és capaç de crear divisions del conjunt original que siguin prou diferents entre si i que, a més, els objectes que pertanyen a una mateixa partició mantinguin una alta similitud. A més, el model s'ha de poder enfrontar amb objectes nous del mateix domini i comportar-se correctament: predir-ne l'etiqueta de classe o integrar-los a la partició més adient.

Per tant, la **qualitat d'un model** és una estimació del comportament futur del model respecte a objectes del domini.

Qualitat d'un model

La qualitat d'un model és una estimació del seu comportament futur respecte a objectes del domini.

Una manera típica d'enfocar el procés de mesurar la qualitat d'un model consisteix a separar el conjunt de dades disponibles en dos grups:

- El **conjunt d'entrenament**: conjunt de les dades que realment s'empren per a construir el model.
- El **conjunt de prova**: conjunt de les dades que s'utilitzen per a veure si el model actua correctament.

Exemple de procés de mesura de la qualitat d'un model: els clients d'Hyper-Gym

En el nostre exemple del gimnàs podem separar el conjunt de socis en dos subconjunts.

Suposem que emprem el primer conjunt per a extreure un model classificatori, per exemple un arbre de decisió, que ens separi els socis en dues classes: els que pot ser que demanin servei d'entrenador personal i els que no en demanaran.

El segon conjunt de socis, el conjunt de proves, l'utilitzarem per a saber quin és el comportament de l'arbre de decisió construït. Observeu que, d'aquests clients del conjunt de proves, ja en coneixem la classe a què pertanyen (ja sabem si demanen entrenador o no). L'objectiu del nostre model és poder predir si un nou client el demanarà. Esperarem que, si el model classificatori construït és prou bo, tots o un gran percentatge dels clients del conjunt de proves siguin classificats correctament mitjançant l'arbre de decisió, en la classe que els correspon. Els clients del conjunt de proves no s'han utilitzat per a construir l'arbre de decisió i, per tant, són nous per a l'arbre.

En el procediment d'avaluació es plantegen una sèrie de qüestions, com les que esmentem a continuació:

1) Quina és la mesura de qualitat adient?

Exemple de mesura de qualitat

En el cas dels clients d'Hyper-Gym, tot i que no ho esmentem explícitament, fem servir la idea d'**error en la predicció**. Si el nombre de prediccions errònies és molt alt, considerem que l'arbre de decisió no és prou bo i no l'utilitzem per a classificar clients nous.

Vegeu l'"Exemple de procés de mesura de la qualitat d'un model" més amunt en aquest mateix subapartat.

Ara bé, per a altres tasques, per exemple per a l'agregació, quines mesures podríem fer servir?

2) Com fixem el nivell de qualitat?

Exemple de nivell de qualitat

En l'exemple dels clients d'Hyper-Gym ens plantegem la qüestió següent: estaríem satisfets amb un arbre de decisió que classifiqués correctament el 89% dels clients del conjunt de prova o potser voldríem que en classifiqués el 95%?

La resposta a aquesta qüestió no és tan senzilla com sembla. Per exemple, s'ha de tenir en compte el cost de classificar un objecte nou incorrectament. Què és més costós: assignar un client de la classe 1 a la classe 2, o viceversa? Aquest és el problema dels falsos positius o falsos negatius tan conegut en estadística, i especialment punyent en el disseny de mètodes de diagnòstic nous en medicina o en eines automàtiques de diagnosi de sistemes mecànics complexos.

3) Com escollim els objectes que han d'entrar als conjunts d'entrenament i de prova?

Tria dels conjunts d'entrenament i de prova

En el cas de l'exemple dels clients d'Hyper-Gym, els objectes que escollim són socis o clients.

És evident que, si escollim un conjunt d'entrenament amb característiques comunes semblants però molt diferents de les del conjunt de prova, el conjunt serà esbiaixat i poc representatiu. Llavors el model resultant segur que dóna valors de qualitat baixos.

L'estadística permetrà de respondre aquestes preguntes amb més precisió. De moment, avancem que hi ha dues situacions diferents amb vista a fer front a l'avaluació de models:

- Quan disposem de **grans quantitats de dades**, tant d'entrenament com de prova: en aquesta situació, en principi, el mètode que hem esbossat podria ser suficient si prenem algunes precaucions en la selecció dels objectes que s'incloguin en cada conjunt. Com més gran sigui la quantitat de dades, més suport tindrem per a poder valorar com a "bo" o "dolent" el model resultant.

- Quan disposem de **poca quantitat de dades**: en aquest cas allò que podem inferir amb el model té menys suport, i es fa més difícil de justificar que les qualitats del model es mantindran enfront d'objectes desconeguts, però és

igualment necessari tenir una idea de la manera en què es comportarà el model.

La literatura de mineria de dades es decanta normalment per ignorar la segona problemàtica, tenint en compte que les tècniques de mineria de dades s'apliquen justament a grans conjunts de dades. Ara bé, per a determinades aplicacions, el nombre de dades no és pas tan alt. Cal ser-ne conscient i conèixer les tècniques d'avaluació de models corresponents. 

1.1. Avaluació de models classificatoris

En aquest subapartat ens plantegem el problema de la manera en què cal mesurar la qualitat d'un model classificatori (arbre de decisió, xarxa neuronal de classificació, llista de regles de classificació, etc.).

L'**avaluació d'un model classificatori** és la capacitat per a predir correctament la classe a què pertanyen objectes no utilitzats en la seva construcció.

Primer veurem les mesures de qualitat i després descriurem el procediment d'avaluació.

Mesures de qualitat dels models classificatoris

Com hem dit, la base de la mesura de la qualitat en classificació és saber si el model etiquetarà correctament objectes nous.

Simplificarem el problema indicant que n'hi ha dues classes, *A* i *B*. En aquest cas, classificar es redueix a dir si un objecte pertany a la classe *A* o a la *B*.

La **mesura de la qualitat** l'establim comptant una operació de classificació com a 'correcta' quan un objecte que pertany a la classe *A* rep l'etiqueta *A* per part del classificador (i de la mateixa manera, quan un objecte de la classe *B* rep l'etiqueta *B*), i comptant una operació de classificació com a 'incorrecta' quan un objecte de la classe *A* rep l'etiqueta *B* i viceversa.

Una manera de mesurar la qualitat del model consisteix a comptar o predir el percentatge de classificacions incorrectes que efectuarà el model (o, en positiu, el percentatge d'operacions de classificacions correctes) a partir d'un conjunt d'observacions o casos (clients o socis en l'exemple d'Hyper-Gym). Cadascuna de les operacions de classificació incorrectes és un **error de classificació**.

Per tant, el paràmetre de qualitat que hem de considerar o estimar és la **taxa d'error**:

$$\text{Taxa d´error} = \frac{\text{Errors}}{\text{Casos}}.$$

Alternativament, podem considerar com a paràmetre de qualitat el percentatge d'èxits (operacions de classificació correctes), que també s'anomena **precisió del model**:

$$\text{Precisió} = \frac{\text{Èxits}}{\text{Casos}}.$$

Exemple de mesura de la qualitat d'un model

Aquí presentem un exemple senzill per al cas de la classificació de clients del gimnàs entre els que poden voler entrenador personal i els que no:

Cas	Classe real	Classe predita
34	Entrenador	Entrenador
24	Entrenador	Entrenador
56	No entrenador	Entrenador
110	No entrenador	Entrenador
21	No entrenador	Entrenador
1	No entrenador	Entrenador
23	Entrenador	Entrenador
77	Entrenador	Entrenador
29	No entrenador	No entrenador
101	No entrenador	No entrenador

La taxa d'error és del 40% o, si ho preferiu, la precisió és del 60%.

1.1.1. Grans quantitats de dades

Amb la mesura de qualitat per a un model de classificació que hem presentat en el subapartat anterior, ja podem endegar el procés d'avaluació del model. Vegem com cal fer-ho.

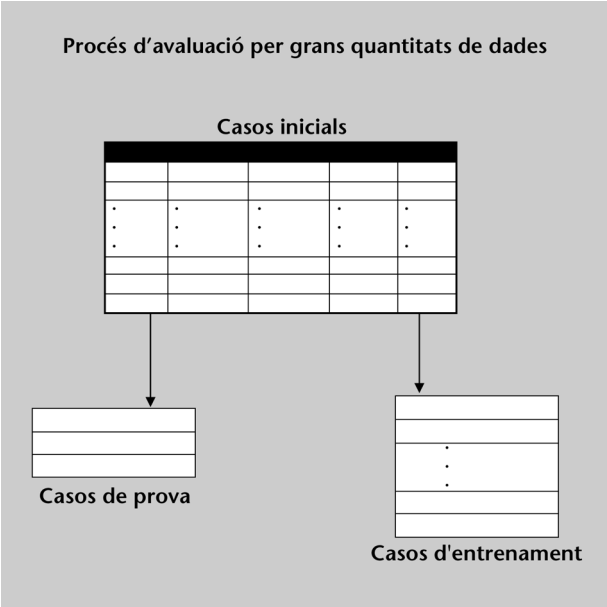
Considerarem una situació en què disposem de grans quantitats de dades. Aleshores, el procediment que se segueix consisteix a disposar de dos conjunts de dades:

Exemple de conjunts d'entrenament i de prova

Es poden extreure dades de dues oficines bancàries d'un mateix barri, i utilitzar les dades dels clients de la primera oficina per a construir el model i les dades dels clients de la segona per a avaluar-lo.

a) El **conjunt d'entrenament**: amb les dades que conté es construirà el model classificatori (arbre de decisió, xarxa neuronal de classificació o llista de regles de decisió).

b) El **conjunt de** prova: es farà servir per a comprovar la taxa d'error (o precisió del model obtingut).



D'aquesta manera, s'intenta reduir el problema de la sobreespecialització*, que obté models massa ben ajustats a dades molt particulars. Si ens descuidem, podem induir models massa específics que no donaran bons resultats en aplicar-los a dades reals noves.

* En anglès, *overfitting*.

Cal suposar que tots dos conjunts són una bona mostra de la població de la qual s'obtenen aquestes dades. Resulta convenient utilitzar més dades per al conjunt d'entrenament que per al de prova. També convé efectuar una extracció aleatòria de casos per a constituir cada conjunt.

Aquest esquema funciona bé amb grans nombres de dades bàsicament perquè, com més gran sigui la mostra de dades, més a prop dels paràmetres de la població haurien d'estar els valors dels paràmetres de la mostra:

a) Com més gran és el conjunt d'entrenament, millor és el classificador (encara que això té matisos: a partir d'un cert nombre de casos d'entrenament, el rendiment baixa).

b) D'altra banda, com més gran és el conjunt de prova, el percentatge d'error que obtenim és més a prop de la realitat.

Importància del volum de dades del conjunt de prova en la precisió del model

Si treballem amb un conjunt de prova de cinc-cents casos i obtenim un percentatge d'error del 2%, podem ser molt optimistes i creure que tenim un classificador fantàstic amb una precisió del 98%. Ara bé, podem considerar que cinc-cents casos és una mostra petita. Si disposem d'un milió de casos del mateix domini i resulta que el percentatge d'error és del 40% (i la precisió, del 60%), podem estar més segurs que la taxa d'error real que mostrarà el classificador estarà més a prop del 40% que no pas del 2%. Més endavant en veurem les raons.

Les suposicions que s'efectuen per a dur a terme aquest tipus de mètode de validació són les que esmentem a continuació:

- Els conjunts de dades que s'han separat són independents entre si.
- Els casos són independents entre si.

S'ha de procurar que, efectivament, aquestes dues condicions es compleixin. En cas contrari, els resultats no seran pas indicatius.

Activitat

1.1. Per què la grandària dels diversos conjunts aporta més o menys seguretat a les taxes d'error i precisió obtingudes?

Confiança en la mesura de l'error

L'**error d'una mostra** s'espera que se situï dins de dues desviacions típiques de la mitjana, cosa que representa un encert en el 95% de les vegades. Això es deu al fet que les mitjanes per a mostres de gran dimensió segueixen una distribució normal.

A partir d'obtenir una taxa d'error determinada mitjançant un test sobre un conjunt de prova, quina confiança podem tenir respecte al veritable valor de l'error del classificador? Suposem que hem obtingut un error del 2,5%. Com podem saber i amb quin grau de confiança si aquest és el veritable error que mostrarà el classificador? Com podem saber entre quins valors variarà l'error amb una confiança suficient? Un altre cop, l'estadística ens ajudarà a respondre aquestes qüestions.

1.1.2. Conjunts limitats de dades: validació creuada

Quan tenim poques dades, les hem d'aprofitar i extreure tant el conjunt de proves com el de dades de la mateixa base de dades original. No hi ha cap proporció fixada respecte al nombre relatiu de components de cada subconjunt, però la més emprada acostuma a ser 2/3 per al conjunt d'entrenament i 1/3 per al conjunt de prova.

Cal fer una extracció adequada d'observacions per a assegurar que no obtenim classificadors esbiaixats.

Exemple de classe sobrerepresentada

Si volem construir un classificador per destriar els clients que demanaran entrenador dels que no ho faran, hem d'evitar, per exemple, que en el conjunt d'entrenament hi hagi el 80% de clients que han demanat entrenador.

No hi ha recomanacions generals, però el sentit comú ens obliga que, després d'una extracció aleatòria de casos, efectuem una anàlisi de dades mínima per garantir que, per exemple, no hi hagi una classe sobrerepresentada en el conjunt d'entrenament. Un mínim ús de l'estadística descriptiva ens permet d'esbrinar

aquesta mena de qüestions encara que no ens en protegeix totalment: hi pot haver grups d'atributs i valors sobrerepresentats, més enllà de l'atribut de classe.

La validació creuada és el procediment més emprat per a assegurar que l'error calculat és proper al real malgrat que s'utilitzi una quantitat de dades relativament petita. El procediment de la validació creuada es pot dir que d'alguna manera "augmenta" la grandària de la base de dades i assegura la independència entre els conjunts de proves utilitzats.

Validació creuada

La validació creuada* és un mètode emprat per a calcular la precisió d'un classificador respecte a un conjunt de dades.

La validació creuada divideix el conjunt de dades en *k* subconjunts excloents mútuament i de grandària aproximadament igual.

*** En anglès, *cross validation*.**

Empíricament, s'ha trobat que la grandària més adequada, o el valor adequat per a *k*, és 10.

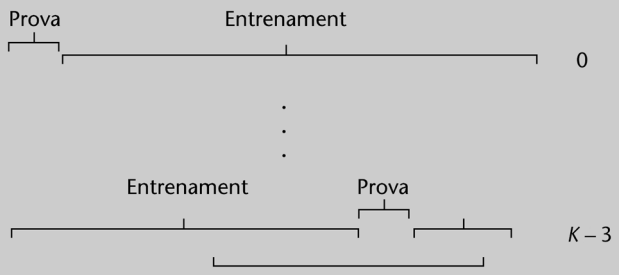
Un cop formats els subconjunts, el classificador agafarà $k - 1$ d'aquests subconjunts (nou subconjunts en el cas de $k = 10$) com a conjunt d'entrenament i el subconjunt que resta com a conjunt de prova, i en calcularà l'error de classificació.

Després, s'agafa el subconjunt següent com a conjunt de prova, es prenen els altres $k - 1$ com a conjunts d'entrenament, i es torna a calcular la taxa d'error.

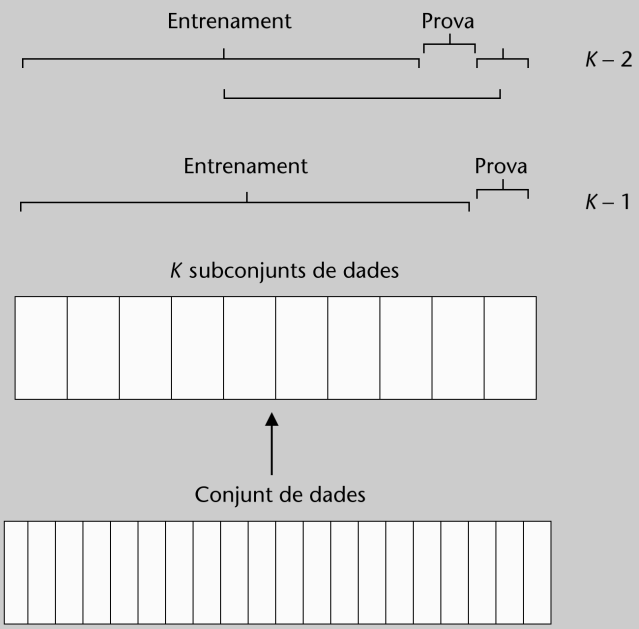
El procediment es repeteix fins que s'esgoten totes les combinacions de subconjunts possibles.

Mostrem gràficament el procediment esmentat en la figura següent:

Mètode de càlcul de l'error amb validació creuada



Mètode de càlcul de l'error amb validació creuada (cont.)



En cada pas es tenen $k - 1$ conjunts d'entrenament i un conjunt de prova. Un cop obtingudes totes les taxes d'error de classificació dels k experiments que s'han fet en total, se'n calcula la mitjana. El resultat és l'error de classificació obtingut mitjançant la validació creuada.

L'avantatge d'aquest mètode és que fa servir totes les observacions del conjunt de dades per a entrenament i per a prova, cosa que no passa amb altres mètodes de càlcul de l'error de classificació.

Si, a més, efectuem la selecció de les particions de manera que les classes existents estiguin representades igual que en el total del conjunt de dades, tant en el conjunt de prova com en el d'entrenament, evitem els problemes de biaix que abans havíem esmentat. Aquest procés s'anomena **estratificació**, i la seva combinació amb la validació creuada, **validació creuada estratificada**.

Suposem que tenim mil dades amb observacions que poden pertànyer a tres classes i la distribució d'observacions per classe és del 30% per a la primera classe, el 40% per a la segona i el 30% per a la tercera, i adoptem una validació creuada de cinc particions. Això vol dir que tindrem particions de dues-centes observacions cadascuna, però on cada partició ha de tenir aproximadament seixanta observacions de la primera classe, vuitanta de la segona i seixanta de la tercera.

El procediment estàndard consisteix a efectuar una validació creuada en deu particions i executar-lo deu vegades per a acomodar les variacions aleatòries. Al final del procés cal obtenir la mitjana dels errors obtinguts per a tenir una estimació de l'error que pugui presentar el model de classificació.

Com es pot veure, l'estimació de l'error de classificació d'un mètode determinat implica un esforç computacional gens menyspreable. No és estrany que, a vegades, quan la precisió no és una necessitat extrema i el temps d'obtenció del model és una consideració important, es prefereixin esquemes d'avaluació més simples.

Complexitat de la validació creuada

Una validació en deu particions implica que el mètode de construcció del model classificatori s'ha d'executar 10×10 vegades.

1.1.3. Comparació de rendiments

Hem comentat que una de les utilitats de l'avaluació de models és poder comparar quin mètode podria donar els millor resultats.

Utilitat de l'avaluació de models

L'avaluació de models ens permet de comparar si una xarxa neuronal ens podria donar un error de classificació més petit que un conjunt de regles de classificació o que un arbre de decisió extrets de les mateixes dades.

Semblaria que n'hi hauria prou d'establir un procediment com el que hem comentat abans, és a dir: obtenir la predicció d'error per a cada mètode i seleccionar el mètode que mostri la taxa d'error mínima. Però, un altre cop i a un altre nivell, ens tornem a posar en una situació típica de la inferència estadística: la diferència entre els resultats dels mètodes es deu a una variabilitat aleatòria o és realment significativa? Per a valorar els resultats i trobar la manera de plantejar la comparació, hem d'introduir algun altre canvi en el procés.

Suposem, per simplificar, que hem de comparar dos mètodes que anomenarem *mètode A* i *mètode B* (una xarxa neuronal de classificació i un arbre de decisió, per exemple).

Ja hem efectuat la validació creuada del model A i del model B obtinguts pels mètodes respectius, i ara tenim una estimació de l'error de A, e_A, i una altra de B, e_B. Ara hem de fer una prova d'hipòtesi: hem de decidir, per exemple, si e_A és significativament superior a e_B. Ho farem amb un test típic: la t de Student.

Suposem un conjunt de valors d'error de classificació obtinguts pel primer mètode, $e_{a1}, \ldots, e_{ak}$, on $k = 10$ si hem fet una validació creuada en deu particions. Per al cas del mètode B tenim $e_{b1}, \ldots, e_{bk}$, també amb $k = 10$. Indiquem la mitjana de la primera mostra d'errors amb $\overline{e_a}$ i la mitjana de la mostra dels errors obtinguts amb el mètode B amb $\overline{e_b}$. Expressem la diferència entre mitjanes amb $dif = e_a - e_b$.

La prova d'hipòtesi que ens plantegem consta de dues hipòtesis alternatives: $dif = 0$ i $dif \neq 0$.

Hem d'establir un nivell de significació i veure si la diferència apreciada supera el nivell que ens hem proposat: hem de derivar l'estadístic t de Student:

$$t = \frac{\overline{dif}}{\sqrt{\sigma^2_{dif} / k}}.$$

Cada factor d'aquesta expressió s'explica a continuació:

- $\overline{dif}$ és la mitjana de les diferències entre els errors observats per al mètode *A* i per al mètode *B*: $e_{a1} - e_{b1}$, (dif_1), $e_{a2} - e_{b2}$, (dif_2), $e_{ak} - e_{bk}$, (dif_k).

- *k*, en el nostre cas, és 10 (el nombre de particions i iteracions efectuades en la validació creuada i, per tant, el nombre d'errors que hem anotat).

- σ^2_{dif} és la variància calculada a partir dels valors de les diferències observades (de la mostra) entre els diversos errors: $e_{ai} - e_{bi}$, amb *i* que varia d'1 a 10.

La distribució que segueix aquest estadístic *t* és la distribució de Student amb $k - 1$ graus de llibertat. Per tant, s'han d'utilitzar les taules d'intervals de confiança corresponents a aquesta distribució.

En el cas que considerem aquí, partim de la suposició que cada error $e_{ai} - e_{bi}$ es refereix a la mateixa partició de les dades. Si no fos així, hauríem de seguir un mètode diferent.

1.2. Altres maneres d'estimar la qualitat de models predictius

Hem discutit i explicat les necessitats d'un tipus de models predictius determinat: els models classificatoris. Hi ha altres mètodes orientats a predir no un etiqueta de classe, sinó un valor numèric concret, com els mètodes de regressió.

En aquest cas els tipus d'error que emprem no són només 'èxit' o 'fracàs', sinó que ens podem equivocar també en grau.

La metodologia general d'avaluació no varia: validació creuada, càlcul de diferències entre els dos mètodes, establiment de la prova d'hipòtesi i ús de la distribució *t* de Student per a avaluar el nivell de confiança. Ara bé, cal donar una altra formalització de l'error.

En aquests casos suposem que partim d'un conjunt d'atributs coneguts $x_1, \ldots, x_n$ i hem de predir un valor numèric de sortida, *y*. De fet, el model que construïm és una funció que ens estableix la relació entre *y* i la resta de variables d'interès:

$$y = a_0 + w_1 x_1 + \ldots + w_n x_n.$$

Indiquem amb *y* el valor real i amb *y'* el valor predit. Aleshores, hi ha altres formalitzacions possibles de l'error, que presentem a continuació:

a) Error quadràtic:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 .$$

b) Error estàndard:

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - y'_i)^2}\,.$$

c) Error absolut mitjà:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i|^2\,.$$

d) Error quadràtic relatiu:

$$\frac{\sum_{i=1}^{n}|y_i - y'_i|^2}{\sum_{i=1}^{n}|y_i - \overline{y'_i}|^2},$$

on $\overline{y'_i}$ és la mitjana dels valors y_i.

La majoria d'aquestes mesures donen una aproximació adequada de l'error de predicció en què s'incorre. De totes maneres, les dues primeres donen un pes massa fort als valors estranys*.

*** En anglès, *outliers*.**

1.3. Cost

Fins ara hem estudiat la qualitat dels mètodes de classificació (o de predicció numèrica) tenint en compte únicament l'error de predicció que poden mostrar. Naturalment, tendim a preferir mètodes que tinguin un percentatge baix d'error en predicció. Ara bé, en comparar dos mètodes diferents, ens interessa tenir en compte també el cost de l'error. És a dir, volem poder decidir-nos per un mètode o un altre segons les conseqüències del seu comportament erroni.

El problema del cost es pot expressar de la manera més senzilla quan tenim únicament dues classes: + i – (o dos valors numèrics, per exemple 0 i 1). La taula següent ho esquematitza (matriu de confusió):

		Classe predita	
		+	–
Classe real	+	Positiu cert	Fals negatiu
	–	Fals positiu	Negatiu cert

El cost de predir un fals positiu o un fals negatiu depèn de l'aplicació a la qual es vulgui destinar el model. Per exemple, si volem crear un classificador que predigui correctament si un pacient pot patir una malaltia determina, un fals po-

sitiu posarà en tractament una persona sana, i un fals negatiu, la deixarà sense. Cadascun d'aquests dos casos té un clar cost econòmic (i humà).

En utilitzar les mesures anteriors, barrejàvem el nombre de positius correctes i negatius correctes i el dividíem pel total de casos de prova. Per a tenir en compte el cost hem de tenir una idea (no una mesura) de la utilitat del model de classificació que hem construït.

Lectura recomanada

Per a una discussió més profunda de les mesures d'utilitat, consulteu l'obra següent:

D. Heckerman (1996). "Bayesian Networks for Knowledge Discovery". A: U. Fayyad; G. Piatetsky-Shapiro; P. Smyth; U. Uthurusamy (ed.). *Advances in Knowledge Discovery and Data Mining* (pàg. 273-306). Menlo Park (Califòrnia): AAAI Press.

1.3.1. Aproximació de costos

Quan es desconeix el cost de cadascuna de les alternatives, es pot intentar treballar amb escenaris que permetin de relacionar alternatives i resultats possibles. Una manera de fer-ho és tenir en compte el concepte de *lift* d'un model predictiu.

El ***lift*** consisteix a avaluar el canvi que hi ha en la probabilitat d'una classe quan s'aplica el model predictiu a una població que mostra unes característiques especials. Es defineix amb l'expressió següent:

$$Lift = \frac{P(Classe|mostra)}{P(Classe|Població)}.$$

Aquest concepte procedeix del màrqueting. En entorns de màrqueting és molt típic fer una estimació del retorn que pot generar una campanya de *mailing*.

Suposem, doncs, que tenim un model que ens permet de predir si un client contestarà el *mailing* rebut o no ho farà. Suposem que escollim un conjunt de clients per avaluar el model respecte a dades que no s'han utilitzat per a construir-lo (el conjunt d'avaluació).

El model assigna l'etiqueta + als clients que prediu que contestaran el *mailing* i – als que no ho faran. Si el model és bo, caldrà esperar que dins els clients marcats com a + hi hagi una proporció més gran de clients que respondran efectivament que no pas la que hi ha a la resta de la població. És a dir, la proporció, que tot seguit presentem, hauria de ser més gran que 1.

$$Lift_{mailing} = \frac{P(Classe_{+}|\ mostra)}{P(Classe_{+}|\ població)}$$

Per exemple, si el conjunt d'avaluació conté el 3% de clients que responen *mailings* i el classificador assigna correctament l'etiqueta + al 60%, llavors el factor de *lift* és el següent:

$$Lift_{mailing} = \frac{0{,}60}{0{,}03} = 20.$$

Si ara tenim un altre mètode de predicció que té un factor de *lift* de 25, no podem afirmar que sigui un model millor, perquè aquest factor està en relació amb la grandària de la mostra i de la població, i també amb els altres costos associats.

En aquest exemple sempre cal preferir un model que ens retorni una quantitat significativa de clients que contestin el *mailing*. En l'exemple, si en la mostra tenim una proporció del 3% de clients que contesten, el màxim factor de *lift* que es pot aconseguir és el següent:

$$Lift_{mailing} = \frac{1{,}00}{0{,}03} = 33{,}\overline{3}.$$

Ara bé, no és el mateix un mètode que té un *lift* de 33,3 respecte a una població de cent mil clients que un de 25 respecte a una població d'un milió de clients. En principi ens interessa més un mètode que l'encerta el 25% de les vegades per a una població d'un milió de persones (dues-centes cinquanta mil respostes positives a la campanya de màrqueting) que un que l'encerta la tercera part de les vegades per a una població de 100.000 persones (33.333 clients que respondran positivament). Això encara s'ha de matisar més. Efectivament, hi ha altres factors de cost que cal tenir en compte:

- Quin dels dos mètodes és més car d'implementar?

- Quin és el cost associat de fer una tramesa de *mailing* a dos-cents cinquanta mil clients enfront dels costos d'una tramesa a 33.333 clients?

Si no disposéssim de cap model predictiu i enviéssim un *mailing* a tota la població objectiu (posem un milió de persones) incloent-hi tots els clients que poden respondre, ens trobaríem amb el següent: si només enviem publicitat al 10% de la població, només podrem arribar al 10% dels clients que responen; si n'enviem al 25%, al 25% dels qui responen, etc.

En un gràfic en què posem en l'eix *X* el percentatge població a la qual s'envia *mailing* i en l'eix *Y* la població de clients que responen, aquesta estratègia "desinformada" donarà una línia recta inclinada un angle de 45 graus.

En canvi, si utilitzem un bon mètode predictiu, cal esperar que la línia estigui per damunt de la de 45 graus. De fet, només ens interessen les accions que portin a quedar situats en la zona superior.

Resum

L'avaluació de models és important per dos motius:

a) D'una banda, l'avaluació de models ens permet de tenir una idea de la qualitat del model obtingut.

b) De l'altra, ens permet de comparar el rendiment de dos o més mètodes dissenyats per a resoldre la mateixa tasca. Per exemple, la classificació obtinguda per un mètode basat en xarxes neuronals enfront de la que s'obté amb un mètode que utilitza arbres de decisió o classificació bayesiana "ingènua".

En principi, el mètode més comú per a tenir una aproximació a la qualitat és, en el cas dels classificadors, el càlcul de la taxa d'error en classificació. És a dir, la proporció d'observacions mal classificades respecte al total d'observacions.

Distingim les dues situacions següents:

1) Grans volums de dades: quan hi ha moltes dades, s'acostuma a separar el conjunt original de dades en dos, un conjunt d'entrenament i un altre de prova.

a) El **conjunt d'entrenament** recull les dades que permeten de construir el classificador.

b) El **conjunt de prova** serveix per a calcular la taxa d'error.

A vegades s'aconsella utilitzar un tercer conjunt, el **conjunt de validació**, procedent d'un conjunt diferent de dades, tot i que es refereix al mateix domini. Un cop validat el model, es pot tornar a provar sobre el segment de dades separat inicialment com a conjunt de prova.

2) Poc volum de dades: quan hi ha poques dades, és recomanable utilitzar el mètode de validació creuada i estratificada.

Per a comparar el rendiment de diversos mètodes, el més corrent és veure les diferències en les respectives taxes d'error. El més assenyat, però, és establir una prova de comparació que es refereixi a les taules de la *t* de Student.

Finalment, pot ser interessant tenir en compte altres factors en la comparació de mètode, com el cost de cada mètode o el retorn esperat. El *lift* és un paràmetre que intenta reflectir aquestes qüestions.

L'avaluació dels mètodes d'agregació s'efectua sobre bases semblants, però aquí el paràmetre per a comparar no és la taxa d'error, com en classificació. Normalment, en el procés de validació creuada es comparen les mateixes mesures de qualitat utilitzades per a construir el model.

Exercicis d'autoavaluació

1. Suposant que, després d'un procés de validació creuada aplicada sobre el mateix conjunt de dades a dos mètodes diferents, heu obtingut els resultats que presentem en la taula següent per a l'error de classificació (expressat en fraccions de la unitat), digueu quin dels dos mètodes és millor.

Mètode 1	0,25	0,05	0,34	0,08	0,02	0,10	0,04	0,08	0,09	0,23
Mètode 2	0,38	0,15	0,04	0,16	0,05	0,08	0,01	0,09	0,12	0,23

2. Suposant que teniu dos mètodes que us han tornat les respectives matrius de confusió que us presentem a continuació, digueu quin dels dos considereu que és millor.

		Classe predita	
		Entrenador personal	**No entrenador personal**
Classe real	**E**	0,45	0,60
	$\neg E$	0,55	0,40

		Classe predita	
		Entrenador personal	**No entrenador personal**
Classe real	**E**	0,65	0,40
	$\neg E$	0,35	0,60