Avaluació de models

PID_00284569

Raúl Montoliu Colás

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Raúl Montoliu Colás

Enginyer en Informàtica per la Universitat Jaume I (UJI) de Castelló. Doctor en mètodes avançats informàtics per la mateixa universitat. Actualment treballa com a docent en el departament d'Enginyeria i Ciència dels Computadors de l'UJI i com a investigador en el grup de recerca Machine Learning for Smart Environments de l'Institut de Noves Tecnologies de la Imatge (INIT). Des del 2017 col·labora com a docent en la Universitat Oberta de Catalunya (UOC).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Julià Minguillón Alfonso

Primera edició: setembre 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoria: Raúl Montoliu Colás
Producció: FUOC

Índex

Introducció

L'avaluació de models és un procés fonamental en la mineria de dades. Es fa servir per comprovar en quina mesura el model que hem creat és capaç de resoldre el problema plantejat. Si no és el cas, haurem de prendre decisions amb la finalitat d'obtenir un nou model capaç de resoldre de la millor manera el problema en qüestió.

Aquest mòdul s'ha dividit en tres parts, corresponents als tres tipus principals de problemes de mineria de dades: problemes supervisats de classificació, problemes supervisats de regressió i problemes no supervisats. Per a cada tipus, s'explicaran les tècniques més comunes emprades per avaluar de manera correcta els models creats.

1. Avaluació de models supervisats de classificació

Imaginem que ens han contractat per dissenyar un algorisme capaç de predir si un client comprarà un determinat producte o no. Per a això, necessitem un conjunt ampli de mostres amb les dades dels clients i, a més, per a cadascuna, és necessari que un expert les etiqueti amb una de les dues possibilitats: comprarà o no comprarà el producte. En aquest cas, l'expert etiqueta amb +1 les mostres dels clients que finalment van comprar el producte, i amb –1 les mostres dels que no. Fent servir aquest conjunt de dades etiquetades, es crearà un model de classificació supervisada. Una vegada tinguem el model, i ateses les dades d'un nou client, podrem fer servir el model per predir l'etiqueta per a aquest client. És a dir, podrem predir si el client comprarà el producte o no.

Per saber si el model que hem creat funcionarà bé o no, només podem fer servir les dades que tenim etiquetades, ja que així podem comparar les etiquetes predites amb les reals.

1.1. Avaluació de problemes binaris

Un problema binari és el que només té dues classes, com per exemple el problema plantejat abans, en què es vol predir si un client comprarà un producte o no.

Per comprovar la qualitat del model desenvolupat, hem de dividir el conjunt original de mostres (les mostres que tenim etiquetades) en dos conjunts: entrenament i test. El conjunt d'entrenament ens servirà per obtenir un model preliminar que tindrà un comportament similar al que podríem obtenir fent servir tot el conjunt de dades original, però no igual, ja que té menys mostres. El conjunt de test s'utilitzarà per validar el model preliminar entrenat. El valor que obtinguem serà una estimació optimista del resultat que s'obtindrà quan es prediguin les mostres reals futures.

Seguint amb l'exemple anterior, suposem que tenim una base de dades de 1.000 clients, dels quals 500 van comprar el producte i 500 no el van comprar. Dividim el conjunt en 800 mostres d'entrenament i 200 de test. És molt important que el nombre d'elements de cada classe estigui equilibrat en tots dos conjunts. Per exemple, un error important seria que el conjunt d'entrenament fos de 500 clients que van comprar el producte i 300 clients que no el van comprar; per tant, el conjunt de test tindria únicament mostres de clients que no van comprar el producte i cap de clients que sí que el van comprar. El

correcte seria, en aquest cas, que el conjunt d'entrenament en tingués 400 de cada tipus i, per tant, el de test 100 de cada tipus.

La mesura més emprada per avaluar la qualitat d'un model és l'*exactitud*, que es defineix com el nombre de mostres per les quals el model ha predit bé la seva classe enfront del nombre total de mostres. Formalment, es defineix com es mostra a continuació:

$$exactitud = \frac{1}{N}\sum_{i=1}^{N}\lambda(f(x_i),l(x_i)) \quad (1)$$

en què x_i és una mostra del conjunt de mostres per avaluar (conjunt de test), $l(x_i)$ és l'etiqueta veritable d'aquesta mostra, f és la funció de predicció del model que retorna l'etiqueta predita de la mostra x_i i λ és una funció que retorna 1 si l'etiqueta predita $f(x_i)$ és igual a la veritable $l(x_i)$ i 0 en un altre cas. L'exactitud obté un valor entre 0.0 i 1.0. Com més proper sigui el valor obtingut a 1.0, més bon comportament presenta el model.

Suposem que hem creat un model amb les 800 mostres d'entrenament. Per obtenir una mesura de la qualitat del model, prediem la classe de les 200 mostres de test: obtenim, per exemple, 180 mostres en què s'ha predit la classe correcta i 20 en què no. En aquest cas, l'exactitud serà 180/200 = 0.9. És a dir, el model encerta en 90 % dels casos.

Una qüestió important és com fem la divisió entre els dos conjunts. Una possibilitat és fer la divisió a l'atzar. No obstant això, podria passar que just un tipus de clients molt particular no estigués present en un dels conjunts. Per exemple, imaginem que succeeix que els clients d'edat avançada que sí que van comprar el producte cauen tots en el conjunt de test, i cap en el d'entrenament. El sistema podria confondre's quan intentés predir un client amb aquestes característiques, ja que no ha estat entrenat per a aquesta mena de mostres.

Per evitar aquest problema, s'empra una tècnica coneguda com a *validació creuada*.

1.1.1. Validació creuada

La validació creuada (o *cross validation*) és una tècnica per estimar l'error que produeix un model. La tècnica consisteix a dividir el conjunt de mostres en diverses carpetes (o *folds*), cadascuna amb un nombre similar de mostres de cada classe.

En el problema plantejat, podríem fer servir 10 carpetes. Per tant, a cada carpeta hi hauria 50 mostres de clients que sí que van comprar el producte i 50

mostres de clients que no el van comprar. Per a cada carpeta, s'entrena amb les mostres pertanyents a totes les carpetes menys l'actual (és a dir, amb 900 mostres) i es valida amb l'actual (amb 100 mostres), i s'obté un valor d'exactitud. D'aquesta manera, per validar la primera carpeta F_1, el conjunt d'entrenament estaria compost per les mostres pertanyents a la resta de carpetes, és a dir, amb $\{F_2,F_3,\ldots,F_{10}\}$. De manera similar, per validar la cinquena carpeta F_5, el conjunt d'entrenament estaria compost per les mostres pertanyents a les carpetes $\{F_1,\ldots,F_4,F_6,\ldots,F_{10}\}$. Finalment, per validar l'última carpeta F_{10}, el conjunt d'entrenament estaria compost per les mostres pertanyents a les carpetes $\{F_1,\ldots,F_9\}$.

Per a cada carpeta s'obté un valor d'exactitud. L'exactitud total és la mitjana de les exactituds obtingudes en totes les carpetes.

Un cas particular és el mètode *leaving one out*, en què les carpetes tenen només una mostra. Per a cada mostra, s'entrena amb la resta i es valida si l'encerta o no. L'exactitud total serà el nombre d'encerts dividit pel total de mostres.

1.2. Avaluació de problemes binaris no equilibrats

Un problema binari no equilibrat és aquell en què el nombre de mostres d'una classe és molt superior al nombre de mostres de l'altra. Suposem, per exemple, que hi ha un sistema d'aprenentatge automàtic capaç de detectar de manera primerenca una malaltia mortal. El sistema pren com a entrada un conjunt de dades del pacient i retorna 1 si el pacient té la malaltia, i –1 en el cas contrari. En el cas de detectar la malaltia, podran prendre's les mesures oportunes per augmentar les probabilitats de superar-la. No obstant això, la medicació té uns efectes secundaris molt desagradables, per això no es recomana que un pacient se la prengui.

En aquest exemple, la classe objectiu és la classe positiva, és a dir, detectar que el pacient té la malaltia. L'altra classe és la classe negativa.

Amb la finalitat de comprovar si el sistema desenvolupat funciona correctament, s'han fet 10,000 prediccions, de les quals 9,500 eren de pacients sans i 500 de pacients amb la malaltia. El sistema ha predit com a sans 9,450 dels pacients sans, i com a malalts 300 pacients amb la malaltia. Per tant, el sistema ha encertat en 9,450 + 300 = 9,750 casos dels 10,000. Una possible mesura de la qualitat del procés és, tal com s'ha explicat a l'apartat anterior, calcular l'exactitud del sistema. En aquest cas és 9,750/10,000 = 0.975. És a dir, el sistema encerta en 97.5 % dels casos.

El valor obtingut per a l'exactitud podria portar-nos a confusió: tot i que certament és un valor molt alt, a la classe objectiu només ha encertat el del 60 % de casos (300 de 500). En realitat, el resultat és molt dolent, ja que hi ha 200 pacients (el 40 %) que tenen la malaltia i el sistema no ho ha detectat. En

aquest tipus de problemes, és crucial encertar la gran majoria dels casos de la classe objectiu, encara que això impliqui augmentar lleugerament les fallades a la classe negativa. En aquesta situació, és desagradable medicar un pacient sa (errar a la classe negativa), però pot ser mortal no medicar-ne un de malalt (errar a la classe positiva). En el nostre exemple, hi ha 50 persones que sofririen els efectes secundaris per error, però 200 persones que podrien morir si no es medicaven.

En aquest apartat, es presentarà una mesura per avaluar aquest tipus de problemes anomenada *F-measure*, *F-score* o *F1-score*, que té en compte com de bo o dolent és el model a l'hora de predir correctament la classe objectiu.

Prèviament, cal definir quatre conceptes importants:

- *Veritable positiu* (TP): és una mostra positiva que el sistema ha predit com a positiva.
- *Veritable negatiu* (TN): és una mostra negativa que el sistema ha predit com a negativa.
- *Fals positiu* (FP): és una mostra negativa que el sistema ha predit com a positiva.
- *Fals negatiu* (FN): és una mostra positiva que el sistema ha predit com a negativa.

En els casos de TP i TN el model encerta en la predicció, mentre que en els altres dos casos el model s'equivoca.

A partir de les definicions anteriors, s'obtenen els valors següents:

- *Precisió*: també anomenada *valor de la predicció positiva*, és la fracció de mostres positives predites com a positives (TP) entre el total de mostres predites com a positives (TP + FP) i es defineix com:

$$Precisió = \frac{TP}{TP + FP} \tag{2}$$

- *Sensibilitat*: també anomenada *recall*, és la fracció de mostres positives predites com a positives (TP) entre el total de mostres realment positives (TP + FN) i es defineix com:

$$Sensibilitat = \frac{TP}{TP + FN} \tag{3}$$

Finalment, la mida *F-measure* s'obté mitjançant l'equació següent:

$$2 \times \frac{Precisió \times Sensibilitat}{Precisió + Sensibilitat} \tag{4}$$

Aquesta mesura obté un valor entre 0.0 (mal resultat) i 1.0 (bon resultat).

Tornant a l'exemple plantejat a l'inici d'aquest subapartat, els valors de TP, TN, FP, FN, Precisió, Sensibilitat i *F-measure* es mostren a la segona columna de la taula 1. Com pot comprovar-se, el valor de *F-measure* no és gaire elevat, a causa de la capacitat deficient del model a l'hora de predir els veritables positius.

Taula 1. Dos possibles resultats d'un problema binari no equilibrat

Mida	Model original	Model millorat
TP	300	450
TN	9,450	9,250
FP	50	250
FN	200	50
Exactitud	$\frac{9,450+300}{10,000} = 0.975$	$\frac{9,250+450}{10,000} = 0.97$
Precisió	$\frac{300}{300+50} = 0.86$	$\frac{450}{450+250} = 0.64$
Sensibilitat	$\frac{300}{300+200} = 0.6$	$\frac{450}{450+50} = 0.9$
F-measure	$2 \times \left(\frac{0.86 \times 0.6}{0.86+0.6}\right) = 0.71$	$2 \times \left(\frac{0.9 \times 0.64}{0.9+0.64}\right) = 0.75$

Després de veure els resultats, els experts decideixen modificar el model perquè eviti punts falsos negatius. El nou model prediu ara com a sans 9,250 dels pacients sans, i com a malalts 450 pacients amb la malaltia. Per tant, amb el nou model, el sistema l'ha encertat en 9,250 + 450 = 9,700 casos dels 10,000. En aquest cas, l'exactitud és 9,700/10,000 = 0.97. És a dir, el sistema l'encerta en 97.0 % dels casos, que, com pot comprovar-se, és una mica menys que en el model original. No obstant això, tal com mostra la tercera columna de la taula 1, la *F-measure* és millor en aquest cas i, per tant, el nou model és preferible a l'original.

1.3. Avaluació de problemes multiclasse

Sovint ens trobem davant problemes de classificació supervisada que tenen més d'una classe. A aquest tipus de problemes els anomenem *problemes multiclasse*. Per exemple, un possible problema multiclasse és aquell en què, donada una imatge en què apareix un cotxe, se'n prediu la marca. És un problema multiclasse, ja que hi ha més de dues possibles marques.

Hi ha estratègies principals per validar problemes multiclasse: *one versus one* i *one versus all*. En tots dos casos, es transforma el problema multiclasse en múltiples problemes binaris.

A la primera estratègia, *one versus one*, es calcula una mesura de bondat del model (com l'exactitud o la *F-measure* explicades abans) per a cada parell de classes. Per exemple, si el problema té quatre classes $\{C_1,C_2,C_3,C_4\}$, es calcularà la mesura de bondat del model per tots els possibles problemes binaris que poden plantejar-se amb les quatre classes, és a dir: $\{C_1\}^+$ vs $\{C_2\}^-$, $\{C_1\}^+$ vs

$\{C_3\}^-$, ..., $\{C_3\}^+$ vs $\{C_4\}^-$, en què $\{\cdot\}^+$ fa referència a les etiquetes que formen part de la classe positiva i $\{\cdot\}^-$ a les etiquetes que formen part de la classe negativa. La mesura de bondat final del problema multiclasse serà la mitjana de les mesures de bondat obtingudes per a tots els problemes binaris plantejats.

La segona estratègia, *one versus all*, consisteix a formar tants problemes binaris com classes hi hagi, de manera que la classe positiva estarà composta per les mostres d'una de les classes existents i la negativa amb les mostres pertanyents a la resta de classes. Seguint amb un exemple amb quatre classes, es calcularà la mesura de bondat del model per als següents quatre problemes binaris: $\{C_1\}^+$ vs $\{C_2,C_3,C_4\}^-$, $\{C_2\}^+$ vs $\{C_1,C_3,C_4\}^-$, $\{C_3\}^+$ vs $\{C_1,C_2,C_4\}^-$ i $\{C_4\}^+$ vs $\{C_1,C_2,C_3\}^-$. Com en l'estratègia anterior, la mesura de bondat final del problema multiclasse serà la mitjana de les mesures de bondat obtingudes per a tots els problemes binaris plantejats.

En realitat, la majoria de les implementacions dels algorismes de classificació supervisada són capaces de tractar amb problemes multiclasse de manera transparent per a l'usuari. Segons el mètode, fan servir internament una de les dues estratègies comentades.

L'usuari haurà de seguir les mateixes recomanacions que s'han comentat per als problemes binaris. D'una banda, a l'hora de seleccionar la mesura de bondat més adequada, s'ha de tenir en compte si el nombre de mostres d'alguna de les classes és molt diferent del de la resta. D'altra banda, quan s'aplica la validació creuada, caldrà tenir especial cura que el nombre de mostres de cada classe incloses en cada carpeta sigui similar.

1.4. Eines d'interès

A més de les tècniques comentades als subapartats anteriors, hi ha un conjunt d'eines que poden ajudar-nos a l'hora d'interpretar els resultats obtinguts en la validació d'un model. Les més comunes són les matrius de confusió i les corbes ROC (*receiver operating characteristic*).

1.4.1. Matriu de confusió

La matriu de confusió és una forma gràfica de comprovar com de bé o malament ha funcionat un model. Per als problemes binaris, és una matriu de dues per dues, en què a la primera fila s'hi posaran els veritables negatius i els falsos positius, i a la segona fila s'hi posaran els falsos negatius i els veritables positius. Per tant, la matriu de confusió pot escriure's tal com apareix a la taula 2. La taula 3 mostra la matriu de confusió de l'exemple corresponent a la segona columna de la taula 1.

Taula 2. Matriu de confusió per a problemes binaris

		Predicció	
		Negatiu	Positiu
Valor real	Negatiu	TN	FP
	Positiu	FN	TP

Un bon model serà el que té valors grans a la diagonal principal i propers a zero a la resta de posicions de la matriu.

Taula 3. Matriu de confusió de l'exemple corresponent a la segona columna de la taula 1

		Predicció	
		Negatiu	Positiu
Valor real	Negatiu	9,450	50
	Positiu	200	300

En problemes multiclasse, la matriu de confusió tindrà tantes files i columnes com nombre de classes hi hagi. De manera semblant al cas dels problemes binaris, obtindrem més bons models quan la diagonal principal tingui valors alts i la resta de posicions de la matriu siguin properes a zero.

Taula 4. Matriu de confusió d'un hipotètic problema multiclasse amb 4 classes

		Predicció			
		A	B	C	D
Valor real	A	50	5	5	40
	B	0	70	2	3
	C	5	0	50	0
	D	10	0	0	60

La taula 4 mostra un hipotètic resultat per a un problema de classificació multiclasse amb quatre possibles classes: A, B, C i D. En aquest exemple, el model ha estat capaç de predir correctament (la diagonal principal) $50+70+50+60 = 230$ de les 300 existents. És a dir, el model obté una precisió del 76.6 % (230/300 = 0.766). Una anàlisi més detallada dels resultats obtinguts ens permet comprovar que de les 100 mostres existents de la classe A, només 50 han estat predites correctament. De les 50 predites incorrectament, la gran majoria (40) han estat predites com a pertanyents a la classe D. Gràcies a la matriu de confusió obtinguda, podem deduir que hi ha un problema de confusió entre les mostres de les classes A i D. El model té un comportament més correcte a les mostres de la resta de classes. Per tant, en aquest cas, i després d'analitzar la matriu de confusió resultant, hauríem de centrar els nostres esforços a esbrinar la raó per la qual tantes mostres de la classe A es prediuen com a pertanyents a la classe D.

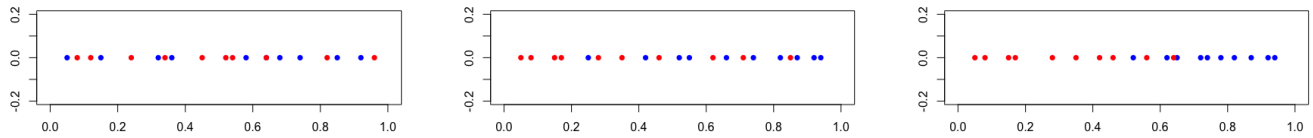
1.4.2. Corbes ROC

Les corbes ROC (*receiver operating characteristic*) són un mètode molt efectiu per validar el funcionament d'un model de classificació supervisada en problemes binaris. Imaginem que estem fent servir un classificador supervisat que, en

comptes d'obtenir un +1 quan prediu que la mostra és positiva i –1 quan és negativa, ens proporciona un nombre entre 0.0 i 1.0 que indica la probabilitat que la mostra pertanyi a la classe positiva. Per exemple, si el resultat és 0.2, voldrà dir que hi ha una probabilitat petita que la mostra sigui positiva. No obstant això, si el resultat és 0.8, voldrà dir que hi ha una probabilitat alta que la mostra sigui positiva. En realitat, la gran majoria d'algorismes de classificació supervisada funcionen així.

Suposem que, per a un problema de classificació supervisada binari, hem entrenat tres algorismes diferents, i que per a un conjunt de 20 mostres de test (10 de cada classe), hem obtingut la probabilitat que la mostra sigui de la classe positiva. La figura 1 mostra el resultat obtingut per als tres models entrenats. Els punts vermells són les mostres negatives i els punts blaus són les mostres positives. La figura 1a mostra un mal resultat, ja que és molt difícil trobar una frontera entre les mostres d'ambdues classes. La figura 1b presenta un resultat intermedi i la figura 1c mostra el millor resultat dels tres, ja que, en aquest cas, és més fàcil establir una frontera entre ambdues classes.

Figura 1. Resultats de tres models diferents de classificació supervisada. Els punts vermells són les mostres negatives i els punts blaus són les mostres positives. Es presenta per a cada mostra la probabilitat que pertanyi a la classe positiva.



a. A l'esquerra, mal resultat. **b.** Al centre, resultat intermedi. **c.** A l'esquerra, bon resultat

Les taules 5, 6 i 7 mostren, respectivament, els TP, TN, FP i FN per als tres models entrenats, per a un conjunt de llindars de 0.0 a 1.0. Aquestes taules també mostren la sensibilitat i 1 menys l'especificitat de cada model.

La sensibilitat (o *recall*) quantifica la proporció de mostres positives (TP + FN) que són classificades com a positives (TP) i s'ha definit anteriorment com:

$$S = \frac{TP}{TP + FN} \tag{5}$$

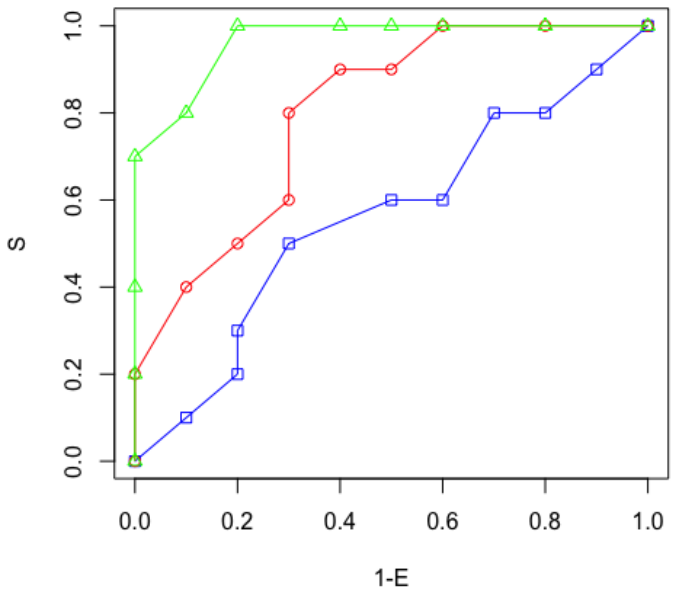
D'altra banda, l'especificitat quantifica la proporció de mostres negatives (TN + FP) que són classificades com a negatives (TN) i es defineix com:

$$I = \frac{TN}{TN + FP} \tag{6}$$

La figura 2 mostra les corbes ROC dels tres models. La corba ROC representa per a cada llindar un punt en què la coordenada x és la sensibilitat i la coordenada y és 1 menys l'especificitat. Una mesura de la qualitat del model és l'àrea

que queda sota la corba. Com pot comprovar-se, aquesta àrea serà més gran en el model representat per la corba verda (mostres de la figura 1c i la taula 7) que en el model intermedi representat per la corba vermella (mostres de la figura 1b i la taula 6), i el model pitjor serà el representat per la corba blava (mostres de la figura 1a i la taula 5).

Figura 2. Corbes ROC dels tres models entrenats (vegeu figura 1). Les corbes verda, vermella i blava es corresponen amb el model millor, mitjà i pitjor, respectivament.



En termes generals, la corba ROC com més s'assembli a una diagonal, més mal model s'haurà obtingut. No obstant això, com més a l'esquerra i a dalt estigui la corba, més bon model s'haurà obtingut.

Taula 5. TP, TN, FP, FN, sensibilitat i 1 menys l'especificitat del pitjor model (representat a la figura 1a), per a cadascun dels valors del llindar.

Llindar	TP	TN	FP	FN	S	1-I
0.0	10	0	10	0	1.0	1.0
0.1	9	1	9	1	0.9	0.9
0.2	8	2	8	2	0.8	0.8
0.3	8	3	7	2	0.8	0.7
0.4	6	4	6	4	0.6	0.6
0.5	6	5	5	4	0.6	0.5
0.6	5	7	3	5	0.5	0.3
0.7	3	8	2	7	0.3	0.2
0.8	2	8	2	8	0.2	0.2
0.9	1	9	1	9	0.1	0.1
1.0	0	10	0	10	0.0	0.0

Taula 6. TP, TN, FP, FN, sensibilitat i 1 menys l'especificitat del model intermedi (representat a la figura 1b), per a cadascun dels valors del llindar.

Llindar	TP	TN	FP	FN	S	1-I
0.0	10	0	10	0	1.0	1.0
0.1	10	2	8	0	1.0	0.8
0.2	10	4	6	0	1.0	0.6
0.3	9	5	5	1	0.9	0.5
0.4	9	6	4	1	0.9	0.4
0.5	8	7	3	2	0.8	0.3
0.6	6	7	3	4	0.6	0.3
0.7	5	8	2	5	0.5	0.2
0.8	4	9	1	6	0.4	0.1
0.9	2	10	0	8	0.2	0.0
1.0	0	10	0	10	0.0	0.0

Taula 7. TP, TN, FP, FN, sensibilitat i 1 menys l'especificitat del millor model (representat a la figura 1c), per a cadascun dels valors del llindar.

Llindar	TP	TN	FP	FN	S	1-I
0.0	10	0	10	0	1.0	1.0
0.1	10	2	8	0	1.0	0.8
0.2	10	4	6	0	1.0	0.6
0.3	10	5	5	0	1.0	0.5
0.4	10	6	4	0	1.0	0.4
0.5	10	8	2	0	1.0	0.2
0.6	8	9	1	2	0.8	0.1
0.7	7	10	0	3	0.7	0.0
0.8	4	10	0	6	0.4	0.0
0.9	2	10	0	8	0.2	0.0
1.0	0	10	0	10	0.0	0.0

2. Avaluació de models supervisats de regressió

Un problema supervisat de regressió es diferencia d'un problema supervisat de classificació en el fet que el que es prediu és un valor continu, i no una etiqueta entre un conjunt finit. Per exemple, un problema de regressió és predir el valor de venda d'una casa. El valor de la casa és un nombre real positiu que és una quantitat contínua. Si es discretitza aquest valor en diversos segments, llavors transformaríem el problema en un de classificació supervisada.

Un concepte important, a l'hora de validar la qualitat d'un model de regressió, és el residu o error d'una mostra, que es defineix com la diferència entre el valor real de la mostra y_i i el valor predit pel model $\hat{y}_i$. A partir del residu, poden definir-se les dues mesures més comunes que se solen fer servir: l'error quadràtic mitjà (ECM o MSE, en anglès) i l'arrel de l'error quadràtic mitjà (RECM o RMSE, en anglès); es defineixen com es mostra a continuació:

$$ECM = \sum_{i=1}^{N} \frac{1}{N}(\hat{y}_i - y_i)^2 \quad (7)$$

$$RECM = \sqrt{ECM} \quad (8)$$

en què N és el nombre de mostres del conjunt de dades.

Per fer correctament la validació del model de regressió, es pot fer servir la validació creuada. En aquest cas, també és molt important distribuir correctament les mostres entre les diferents carpetes. Suposem que la nostra base de dades té 800 mostres de cases amb preus entre 100,000 i 500,000 euros. Per distribuir correctament les mostres, podem crear arbitràriament tres grups diferents segons el preu de les cases. Per exemple, el primer grup estaria compost per les cases entre 100,000 i 200,000 euros; el segon grup, entre 200,000 i 300,000 euros, i l'últim grup, amb les de més de 300,000. El pas següent seria explicar el nombre de mostres de cada grup. Per exemple, suposem que tenim 400, 300 i 100 mostres en cada grup. El pas següent és distribuir les mostres de cada grup a les diferents carpetes intentant que el nombre de mostres de cada grup a cada carpeta sigui similar. Seguint amb l'exemple i assumint que fem servir 10 carpetes, posaríem 40, 30 i 10 mostres de cada grup a cada carpeta.

3. Avaluació de models no supervisats

Els models no supervisats, com els mètodes d'agrupament, agregació o *clustering*, també han d'avaluar-se per validar la qualitat de l'agrupament obtingut. No obstant això, al contrari dels mètodes supervisats, en els mètodes no supervisats és complicat definir quin és el resultat correcte, ja que les úniques dades de què partim són les mateixes mostres, sense l'existència d'informació addicional que pugui confirmar-nos si cada mostra ha estat assignada al grup o clúster correcte. Hi ha un conjunt de tècniques que poden utilitzar-se amb l'objectiu de validar la qualitat de l'agrupament o el mètode.

Si volem avaluar la qualitat de l'agrupament obtingut, hem d'emprar tècniques de validació interna. En la validació interna es fa servir un conjunt d'índexs de qualitat que només depenen de les dades de què es parteix. El que es vol és obtenir un índex de com de bé o malament s'han agrupat les mostres en els diferents grups. Aquestes tècniques també es poden emprar per obtenir el nombre òptim de grups en algorismes com el *k-means*.

Si el que volem és fer una comparativa entre diversos algorismes d'agrupament, podem fer servir tècniques de validació externa. Per a això, serà necessari disposar d'informació addicional, com per exemple l'etiqueta de classe de cada mostra. A la pràctica, no es disposa d'aquesta informació, ja que estem tractant amb problemes no supervisats. No obstant això, i amb l'única fi de comprovar si l'algorisme d'agrupament que hem desenvolupat obté bons resultats, podem fer servir una base de dades que inclogui informació de la classe a què pertany cada mostra (definida per a problemes supervisats), i utilitzar aquesta informació per comprovar la qualitat de l'agrupament obtinguda amb el nostre algorisme.

3.1. Tècniques de validació interna

Les tècniques de validació interna es fonamenten en el càlcul d'un conjunt d'índexs basats únicament en informació obtinguda de les mateixes dades. En termes generals, aquestes tècniques es basen en els dos criteris següents:

- *Cohesió*: una mostra pertanyent a un grup ha d'estar prop de la resta de mostres del mateix grup.

- *Separació*: una mostra pertanyent a un grup ha d'estar lluny de les mostres pertanyents als altres grups.

Per tant, basant-se en aquests dos criteris, un bon agrupament és el que maximitza tant la cohesió com la separació. És a dir, quan les mostres d'un grup estan prop de les mostres del seu propi grup i lluny de les mostres de la resta de grups.

A continuació, es presenta un conjunt d'índexs que es poden fer servir per mesurar la qualitat d'un agrupament. En un cas real, és recomanable provar diversos índexs diferents per poder obtenir més informació sobre la qualitat de l'agrupament aconseguit.

Figura 3. Dos exemples de possibles resultats d'un model d'agrupament. El resultat de l'esquerra (**a**) és clarament millor que el de la dreta (**b**).

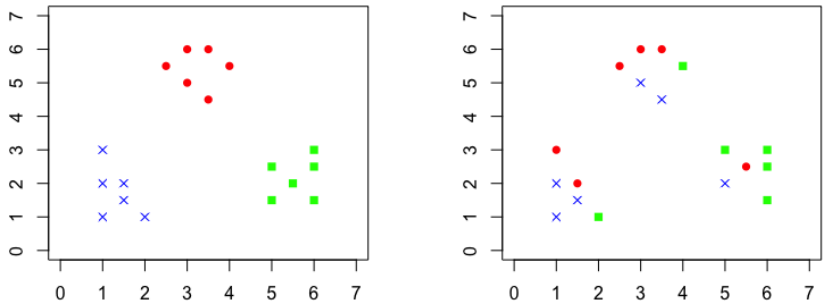
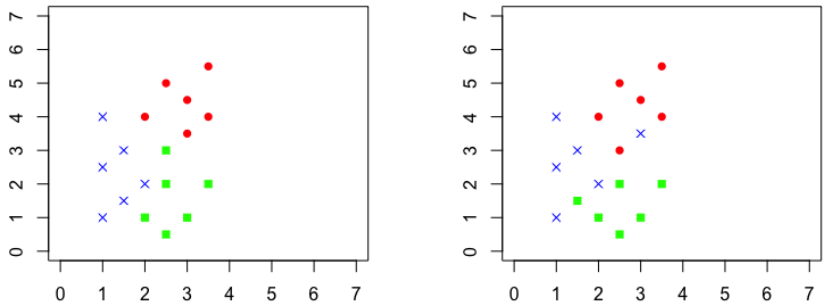


Figura 4. Dos exemples de possibles resultats d'un model d'agrupament. No és tan clar, com en el cas mostrat a la figura 3, quin resultat és el millor.



Per facilitar la comprensió de cada tècnica, es presenten dos resultats diferents d'un model d'agrupament per a dos conjunts de dades diferents. Les figures 3a 3i b mostren el primer conjunt de dades. La figura de l'esquerra (figura 3a) presenta un resultat de més bona qualitat que la figura de la dreta (figura 3b), per la qual cosa és d'esperar que els índexs que s'obtinguin siguin millors en el cas de l'esquerra que en el de la dreta. Les figures 4a i 4b mostren el segon conjunt de dades. En aquest cas, no és tan clar quin model d'agrupament ha aconseguit més bons resultats. Seran els valors que obtinguem dels índexs els que ens resoldran el dubte.

3.1.1. *Sum of squared within* (SSW)

Aquest índex s'utilitza per mesurar la cohesió dels grups obtinguts. S'obté mitjançant l'ús de l'equació següent:

$$SSW = \sum_{i=1}^{k} \sum_{x_j \in G_i} (x_j - \mu_i)^2 \tag{9}$$

en què k és el nombre de clústers, x_j una mostra del grup G_i i μ_i és el centroide del i-èsim grup G_i.

Com més petit sigui el nombre obtingut, més cohesionats estaran els grups, ja que les distàncies entre les mostres i la seva centroide seran més petites. Cal tenir en compte que el valor aconseguit depèn del nombre de mostres de cada grup. A l'exemple mostrat a les figures 3 i 4, tots els grups tenen el mateix nombre de mostres, fet que facilita la comparació. Però en un cas real no té per què passar el mateix. De fet, el més normal serà que no passi.

La primera fila de la taula 8 mostra els valors obtinguts de l'índex *SSW* per als exemples mostrats a les figures 3 i 4. Tal com s'esperava, el resultat aconseguit confirma que el millor agrupament dels quatre exemples és el presentat a la figura 3a, que és molt superior al valor obtingut per a l'altra agrupació amb el mateix conjunt de dades, mostrada a la figura 3b. En el cas de les dues agrupacions presentades a les figures 4a i 4b, el resultat està més igualat, i és lleugerament preferible l'exemple de la figura 4a.

3.1.2. *Sum of squared between* (SSB)

Aquest índex es fa servir per mesurar la separació entre els grups obtinguts. S'obté mitjançant l'ús de l'equació següent:

$$SSB = \sum_{i=1}^{k} |G_i|(\mu - \mu_i)^2 \tag{10}$$

en què k és el nombre de clústers, $|G_i|$ és el nombre de mostres del grup G_i, μ_i és el centroide del i-èsim grup G_i i μ és la mitjana de tot el conjunt de dades.

Com més gran sigui el nombre, més separació hi haurà entre els grups. Tal com passava en el cas de l'índex *SSW*, el valor obtingut depèn del nombre de mostres.

La segona fila de la taula 8 mostra els valors obtinguts de l'índex *SSB* per als exemples mostrats a les figures 3 i 4. Igual que passava amb l'índex SSW, el millor agrupament dels quatre exemples és el presentat a la figura 3a. De

manera semblant, en el cas de les dues agrupacions mostrades a les figures 4a i 4b, el resultat està de nou més igualat, i també és lleugerament preferible l'exemple de la figura 4a.

3.1.3. Índexs basats en SSW i SSB

A partir dels dos índexs anteriors SSW i SSB, pot obtenir-se un altre conjunt d'índexs per valorar la qualitat de l'agrupament aconseguit per un model. De manera general, com més petit sigui el valor de SSW i més gran el de SSB, això voldrà dir un agrupament millor. L'article de Zhao i Fränti (2014) presenta un estudi molt complet d'índexs que es poden emprar per valorar la qualitat de l'agrupament. A continuació, es presenta un subconjunt dels més utilitzats:

- *Ball and Hall* (Ball i Hall, 1965): $\frac{SSW}{k}$
- *Caliński and Harabasz* (Caliński i Harabasz, 1974): $\frac{SSB}{k-1}$
- *Hartigan* (Hartigan, 1975): $\log\left(\frac{SSB}{SSW}\right)$
- *XU-index* (Xu, 1997): $d \times \log\left(\sqrt{\frac{SSW}{dN^2}}\right) + \log(k)$
- *WB-index* (Zhao i uns altres, 2009): $k \times \frac{SSW}{SSB}$

en què k és el nombre de grups, N és el nombre total de mostres i d és la dimensió del problema. En els exemples mostrats a les figures 3 i 4, $k = 3$, $N = 18$ i $d = 2$.

Per als índexs *Ball and Hall*, *XU-index* i *WB-index*, són preferibles els valors baixos. No obstant això, en el cas dels índexs *Caliński and Harabasz* i *Hartigan*, els valors alts indiquen una qualitat millor de l'agrupament.

La taula 8 mostra els valors que s'obtenen dels índexs anteriors per als dos exemples mostrats a les figures 3 i 4. Com pot comprovar-se, i tal com era esperat, el primer exemple aconsegueix sempre els millors resultats. En el cas de les dues agrupacions mostrades a les figures 4a i 4b, el resultat està sempre molt igualat, i és lleugerament preferible l'exemple de la figura 4a.

Taula 8. Resultats obtinguts fent servir els índexs més habituals i les dades mostrades a les figures 3 i 4. El millor resultat per a cada índex es mostra en negreta.

Índex	Figura 3a	Figura 3b	Figura 4a	Figura 4b
SSW	9.38	82.20	16.63	19.04
SSB	102.72	27.64	34.56	32.13
BH	3.78	27.40	5.54	6.34
CH	78.34	2.52	15.59	12.65
H	2.34	–1.09	0.73	0.52
XU	8.47	10.60	9.00	9.13
WB	0.29	8.92	1.44	1.78
DB	1.35	18.60	9.51	9.76
SC	0.72	–0.01	0.30	0.22

3.1.4. Davies Bouldin

L'índex de Davies Bouldin (DB) no està directament relacionat amb els índexs SSW i SSB, malgrat que els principis en què es basa són similars. Per calcular aquest índex, es fa servir l'equació següent:

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{11}$$

$$R_i = \max_{j=1,\ldots,k} R_{ij} \tag{12}$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \tag{13}$$

en què d_{ij} és la distància entre els centroides dels grups G_i i G_j, S_i és la distància mitjana entre cada punt del grup G_i i el seu centroide μ_i i S_j és la distància mitjana entre cada punt del grup G_j i el seu centroide μ_j.

Com més baix sigui el valor d'aquest índex, això indicarà grups més compactes els centroides dels quals estan ben separats els uns dels altres.

La penúltima fila de la taula 8 mostra el valor obtingut per als exemples mostrats en les figures 3 i 4. El resultat és el mateix que quan s'han usat els altres índexs.

3.1.5. Coeficient de Silhouette

El coeficient de Silhouette tampoc està directament relacionat amb els índexs SSW i SSB i es calcula per a cada mostra del conjunt de dades. Proporciona un valor entre –1.0 i 1.0 que indica com de bé o malament està agrupat aquest punt en el seu grup. Els valors propers a 1.0 indiquen que la mostra és al grup correcte. Els valors propers a –1.0 indiquen que la mostra és al grup incorrecte.

Aquesta mesura es calcula per a cada punt x_i del conjunt de dades, com es mostra a continuació:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(a(x_i), b(x_i))} \tag{14}$$

en què $a(x_i)$ és la distància mitjana d'una mostra a totes les del seu mateix grup i $b(x_i)$ és la distància mínima de la mostra en qualsevol altra mostra de la resta de grups. $a(x_i)$ és una mesura de la cohesió, mentre que $b(x_i)$ és una mesura de la separació.

Una mesura de la qualitat de l'agrupament es pot obtenir calculant la mitjana dels coeficients de Silhouette de totes les mostres del conjunt de dades (SC), com es mostra a l'equació següent:

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x_i) \tag{15}$$

L'última fila de la taula 8 mostra el valor obtingut per als exemples mostrats a les figures 3 i 4. De nou, el resultat coincideix amb l'aconseguit quan s'han fet servir els altres índexs.

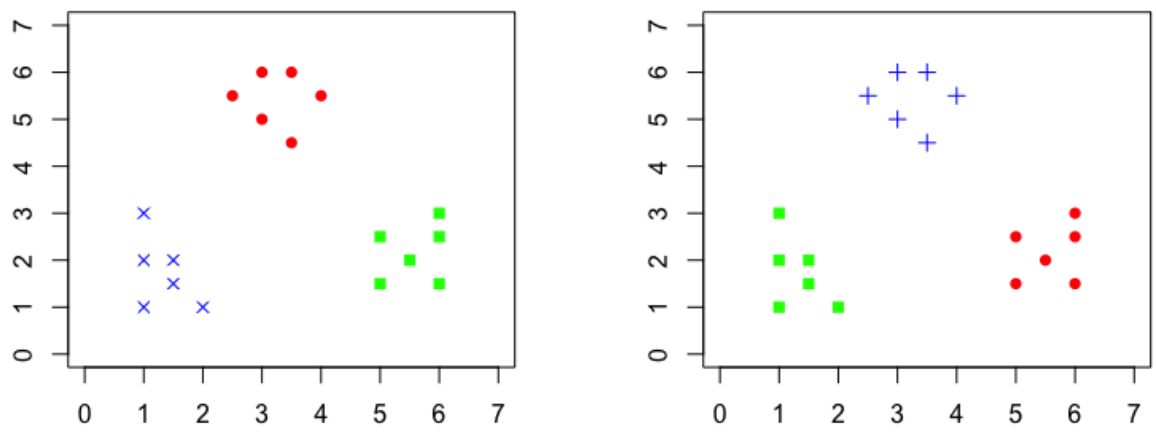
3.2. Tècniques de validació externa

Quan l'objectiu és validar la qualitat d'un algorisme d'agrupament o comparar el funcionament de diversos algorismes, podem fer servir, a més de les tècniques de validació interna descrites, un conjunt de dades etiquetades prèviament per un expert. En primer lloc, es procedirà a executar l'algorisme d'agrupament fent servir les dades d'entrada però excloent la variable que etiqueta les mostres. Una vegada que l'algorisme ha obtingut una partició de les mostres en els diferents grups, es procedirà a comparar si el grup en què s'ha classificat cada mostra coincideix amb l'etiqueta original.

Una possible mesura de la qualitat de l'agrupament és l'exactitud, definida com el total de mostres ben classificades dividit pel nombre total de mostres. També podrien emprar-se altres mesures, com les comentades a l'apartat 1.

És important tenir en compte que el mètode d'agrupament proporciona una agrupació de les mostres en els diferents grups, però no identifica la classe a què pertanyen. La figura 5 presenta aquest problema. Imaginem un problema de classificació d'imatges d'animals. Les mostres originals estan etiquetades com a *Gats*, *Gossos* o *Vaques*. A la figura 5a hi apareixen les mostres de la base de dades fent servir aspes blaves per representar els gats, cercles vermells per representar els gossos i quadrats verds per representar les vaques.

Figura 5. A l'esquerra, conjunt de dades original (**a**). A la dreta, un possible resultat d'un algorisme d'agrupament (**b**).



La figura 5b mostra un possible resultat de l'aplicació d'un algorisme d'agrupament. L'algorisme d'agrupament agrupa les mostres en tres grups, i amb la finalitat d'informar l'usuari del grup a què pertany cada classe, assigna a les mostres un nombre entre 1 i el nombre de grups (en aquest cas, 3). Aquest identificador del grup no proporciona cap informació de caràcter semàntic. És simplement un identificador per poder diferenciar a quin grup pertany cada mostra. Per representar el resultat, s'han fet servir els quadrats verds per mostrar les mostres del primer grup, els cercles vermells per al segon i les aspes blaves per al tercer.

Abans de poder comprovar si l'agrupament obtingut ha classificat les mostres correctament, hem d'interpretar els grups obtinguts. És a dir, hem d'arribar a la conclusió que el grup 1 (quadrats verds a la figura 5b) correspon als gossos (aspes blaves a la figura 5a), que el grup 2 (aspes blaves a la figura 5b) correspon als gats (cercles vermells a la figura 5a) i que, finalment, el grup 3 (cercles vermells a la figura 5b) correspon a les vaques (quadrats verds a la figura 5a). En aquest exemple, és relativament senzill, ja que només tenim dues dimensions, però en altres problemes amb més dimensions aquest procés pot ser molt complex.

Bibliografia

Ball, G.; Hall, D. (1965). *ISODATA. A novel method of data analysis and pattern classification*. Stanford Research Institute.

Caliński, T.; Harabasz, J. (1974). «A dendrite method for cluster analysis». *Communications in* Statistics (vol. 3, núm. 1, pàg. 1-27).
<https://www.tandfonline.com/doi/abs/10.1080/03610927408827101>

Hartigan, J. A. (1975). *Clustering algorithms*. Nova York: Wiley.

Xu, L. (1997). «Bayesian Ying–Yang machine, clustering and number of clusters». *Pattern Recognition Letters* (vol. 18, núm. 11, pàg. 1167-1178).
<http://www.sciencedirect.com/science/article/pii/s0167865597001219>

Zhao, Q.; Fränti, P. (2014). «WB-index: A sum-of-squares based index for cluster validity». *Data and Knowledge Engineering* (vol. 92, pàg. 77-89).
<http://www.sciencedirect.com/science/article/pii/s0169023x14000676>.

Zhao, Q.; Xu, M.; Fränti, P. (2009). «Sum-of-Squares Based Cluster Validity Index and Significance Analysis». A: M. Kolehmainen; P. Toivanen; B. Beliczynski. *Adaptive and Natural Computing Algorithms* (pàg. 313-322). Berlín / Heidelberg: Springer.