Disseny experimental en analítica de dades

PID_00247922

Ester Bernadó

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Índex

Introducció

Fa un temps que Pedro González treballa com a consultor d'analítica de dades. Els seus clients li demanen sovint que analitzi dades dels seus negocis i que els proporcioni la millor solució als seus problemes d'analítica. Després d'uns anys d'experiència s'adona que tots els seus clients es troben, en el fons, amb problemes similars. Fins i tot fa uns dies el programador amb el qual treballa li va preguntar: «Quin és el millor mètode de classificació? M'han parlat dels arbres de decisió. Si implemento una versió molt optimitzada, ens estalviem futures implementacions d'altres algoritmes». Un altre dia, un client, la Sra. Martínez, directora de la secció d'analítica del departament de màrqueting d'una empresa, comentava: «Crec que les màquines de suport vectorial són uns mètodes molt bons de classificació. A partir d'ara, les aplicaré en tots els problemes d'analítica de màrqueting». Altres clients pregunten pel percentatge d'encert dels classificadors o l'error mitjà de les prediccions. Per exemple, el Sr. Dupont va preguntar si el sistema de predicció de vendes del seu comerç electrònic era fiable.

Amb els anys d'experiència, el Pedro s'ha adonat d'algunes veritats:

- *No hi ha un mètode infal·lible que tingui el millor rendiment per a tots els problemes d'analítica. Tot i això, els seus clients continuen demanant la «recepta màgica». El Pedro respon que no existeix el* one-size fits all.

- *Per a un problema determinat, cal realitzar una estimació de la precisió o error del model i comparar diferents models entre si per identificar la millor aproximació al problema.*

- *L'estimació de la qualitat del model s'ha de fer amb afecte, diu el Pedro. És a dir, cal evitar biaixos, tant si són pessimistes com optimistes. Diu que és difícil justificar a un client que la seva estimació de precisió era del 90% i, finalment, el model opera amb precisions del 50%.*

El Pedro ha signat un projecte fa poc per al desenvolupament d'un sistema de predicció de frau en transaccions amb targetes de crèdit. El client disposa d'una base de dades de transaccions a partir de la qual es pot construir un model. El Pedro, amb l'ajuda de l'informàtic, desenvoluparà el sistema de predicció de frau. A més, han de proporcionar una estimació del percentatge d'encert del sistema. És a dir, quan es predigui una transacció com a fraudulenta o no fraudulenta, quina probabilitat té d'encertar en la predicció. En la primera fase, s'ha de realitzar el disseny experimental del projecte.

Com es pot observar en la història que inicia aquest mòdul, el disseny experimental de l'anàlisi de dades és d'una importància vital. En un projecte d'analítica no només cal descobrir patrons interessants en les dades, sinó que

també és important poder estimar quina qualitat presenta el model extret o els patrons identificats. Aquesta qualitat pot ser mesurada de diferents maneres, i la precisió o l'error en són algunes de les més habituals. En aquest mòdul es revisen els elements essencials del disseny experimental d'anàlisi de dades, les mesures de qualitat dels models i la validació estadística dels resultats.

1. Disseny experimental d'anàlisi de dades

Hi ha diverses metodologies que especifiquen els passos necessaris per a realitzar una anàlisi de dades. Una de les més conegudes és CRISP-DM (*Cross Industry Standard Process for Data Mining*), proposada per Daimler-Chrysler i SPSS, que van ser pioners en l'ús de la mineria de dades en els seus processos de negoci. Una altra metodologia coneguda és SEMMA (*Sample, Explore, Modify, Model, Assess*), proposada pel SAS Institute i definida com el procés de selecció, exploració i modelatge de grans quantitats de dades per a descobrir patrons desconeguts. Altres metodologies són les proposades per Zumel i Mount (2014), i Jain i Sharma (2015). Aquests últims introdueixen el *framework* BADIR (*Business question, Analysis plan, Data collection, Insights, Recommendation*). Tot i que cada metodologia té algunes particularitats, totes tenen molts aspectes en comú, ja que recullen els passos bàsics que un analista ha de seguir per a desenvolupar projectes d'analítica. Totes elles presenten un conjunt de fases que es mouen des de la definició del problema, contextualitzat en l'entorn de negoci o científic, fins a la implantació i presa de decisions. Els processos solen ser interactius, en diàleg continu amb l'analista i agents clau (*stakeholders*), i iteratius, ja que progressen en iteracions successives, i amb possibles realimentacions, en què el resultat d'una fase pot provocar el replantejament de fases prèvies.

Per al propòsit d'aquest mòdul, prendrem com a referència el procés d'analítica de dades proposat per Zumel i Mount, en el qual s'enumeren les fases següents:

1) Definir l'objectiu.
2) Recollir i gestionar les dades.
3) Construir el model.
4) Avaluar críticament el model.
5) Presentar els resultats i documentar-los.
6) Implantar el model.

En la primera fase, es plantegen els **objectius de l'anàlisi**. Aquests poden estar contextualitzats en un entorn científic, amb interessos de recerca, o un entorn aplicat o de presa de decisions, com per exemple en l'àmbit d'un negoci. En qualsevol cas, hi ha una necessitat d'anàlisi de dades que s'ha de formular de manera explícita i que marcarà la definició de la resta de les fases.

En la **recol·lecció** de dades s'identifiquen i s'obtenen les dades d'interès. La font de dades pot provenir de dades prèviament emmagatzemades o bé es dissenya un procés de recollida de dades expressament. En aquesta fase s'ha de tenir en compte el mostreig de les dades per a aconseguir mostres representatives de l'objecte d'estudi, ja que les conclusions que s'extreguin de les dades depenen en gran mesura d'aquesta representativitat. Les tècniques de

mostreig descrites en el mòdul «Introducció a l'Estadística» (com el mostreig probabilístic i, dins seu, el mostreig estratificat simple) promouen aquesta representativitat tan necessària. Quan es disposa de la mostra, és freqüent ferne un preprocés, especialment si prové de repositoris de dades sobre les quals l'analista no ha tingut possibilitat d'intervenir. Així doncs, es revisen les dades i es procedeix a la neteja de dades, com ara detecció d'errors, tractament de valors perduts i transformació de dades per al seu modelatge posterior.

Un cop es disposa de la mostra preparada per a l'anàlisi, es **construeix el model** de les dades. Aquesta és la fase central de l'anàlisi, en què s'extreuen els patrons de les dades (moltes vegades denominats *insights*). En aquesta fase s'ha de decidir quins models són més apropiats per a extreure la informació en funció de l'objectiu de l'anàlisi i els recursos disponibles. Aquesta fase s'anomena *fase d'entrenament* del sistema (*training*), que es distingeix de la *fase de test* que ve a continuació.

La fase 4 de la metodologia consisteix a **avaluar críticament el model**, la qual cosa vol dir que s'han de proporcionar mesures de la qualitat del model o, el que és el mateix, de la qualitat de la informació extreta de les dades. En aquest pas no ens referim encara al fet de si els *insights* són útils per al context, la qual cosa s'ha d'analitzar en un moment posterior, sinó més aviat amb quina precisió o confiança podem afirmar aquests *insights*. Per exemple, si un *insight* extret és que els clients que compren al supermercat els dilluns gasten un 50% menys que els que hi compren els dissabtes, hauríem de poder acompanyar aquesta afirmació amb la seva precisió. Podríem dir que això es compleix en un 85% dels casos. Si s'obtenen models que no són prou precisos, es poden replantejar algunes de les fases anteriors: el mètode usat per a extreure els *insights*, la seva configuració, la representativitat de la mostra o el mateix enfocament de l'anàlisi.

Així mateix, en aquesta fase és interessant situar la informació extreta dins el context de definició del problema. S'han de revisar qüestions com si la informació extreta és útil per als objectius de la investigació o de l'àmbit de decisió. Potser la informació és precisa (per exemple, el comportament extret pels clients del supermercat podria tenir una precisió del 95%) però no aportar res de nou en relació amb els objectius d'anàlisi. Si és així, s'han de revisar les fases prèvies per a abordar altres enfocaments.

En les fases 5 i 6 es **presenten els resultats** davant els agents clau i es **desenvolupa i s'implanta** el sistema, o es prenen les decisions oportunes. Com correspon a les fases finals d'un procés, també es revisa si cal fer noves etapes en el projecte que facilitin nova informació amb més dades o mitjançant nous enfocaments.

L'aspecte central d'aquest mòdul i que configura una gran part del procés d'analítica de dades és la fase 4, l'avaluació del model. Partim de la suposició que l'analista ja ha realitzat el mostreig i la neteja de dades, i es disposa a

construir el model. Abans de construir-lo, cal dissenyar com es construirà, és a dir, en quines condicions i com se n'avaluarà la qualitat. Precisament, cal mostrejar de nou les dades per a escollir les que formaran part de la construcció del model i les que es reservaran per contrastar la qualitat del model. Cal també configurar amb quines mètriques s'avaluarà el model, de manera que es puguin recollir aquestes mesures tant durant la construcció del model com en finalitzar el procés. És per tot això que diem que l'avaluació de la qualitat del model determina en gran mesura el procés d'analítica de dades. En l'apartat següent s'introdueixen breument les diferents mètriques de qualitat dels mètodes principals de l'anàlisi de dades. A continuació, es descriuen els processos de mostreig necessaris vinculats a l'estimació de la qualitat del model. Finalment, es tracta la validació estadística que permet mesurar en quin grau les conclusions obtingudes de l'anàlisi són vàlides estadísticament.

2. Mètriques

En aquesta secció es presenten les mesures habituals d'avaluació de la qualitat dels models. El tipus de mètrica depèn del model que s'extreu de les dades i aquest, alhora, depèn de la pregunta d'investigació o necessitat d'informació plantejada en els objectius del projecte d'analítica. Així, si la pregunta és quines tipologies de clients compren en un supermercat, el model serà d'agrupació o *clustering* i la seva avaluació haurà de mesurar en quin grau els grups identificats compleixen determinats criteris desitjables per a una bona agrupació. Si l'objectiu del projecte és detectar abocaments de petroli a l'oceà, el model construït és de classificació o categorització (una taca sospitosa s'ha de classificar com a abocament o no) i la seva avaluació és el grau de precisió, o dit d'una altra manera, la capacitat de detecció d'aquests abocaments.

A la taula 1 es resumeixen les mètriques corresponents als models més freqüents d'anàlisi de dades. Concretament, ens centrem en els models de classificació, regressió, associació i agrupació, els quals es descriuen a continuació.

Taula 1. Models i mètriques de qualitat d'anàlisi de dades

Exemple	Model	Mètriques
Detecció d'abocaments de petroli a partir d'imatges de satèl·lit	Classificació	• Precisió (*accuracy*) • Kappa • Sensitivitat • Especificitat
Predicció de l'àrea cremada de bosc	Regressió	• Error absolut mitjà • Error quadràtic mitjà • Arrel de l'error quadràtic mitjà • Error quadràtic mitjà relatiu • Arrel de l'error quadràtic mitjà relatiu • Coeficient de correlació
Agrupació de tipologies de clients per a una campanya de màrqueting	Agrupació o segmentació (*clustering*)	• Grau de cohesió dins dels grups en relació amb la distància entre grups • Índex de Davies-Bouldin
Identificar llibres que es compren conjuntament en un comerç electrònic	Associació	• Suport i confiança • *Lift*

2.1. Classificació

La mètrica més comuna en models de classificació és la precisió (*accuracy*). La precisió mesura el percentatge d'encerts, és a dir, el nombre de classificacions fetes correctament en relació amb el total de classificacions efectuades.

Prenguem un exemple senzill del conjunt de dades iris disponible al repositori UCI (Lichman, 2013). En aquest problema, s'han de classificar instàncies referides a flors en tres tipus, *Iris setosa*, *Iris versicolor* o *Iris virginica*, a partir dels atributs de longitud i amplada del pètal i del sèpal. La precisió és el nombre de flors classificades correctament en relació amb el total. Si es detalla la predicció per cada classe, s'obté la matriu de confusió que es mostra a la taula 2.

Taula 2. Matriu de confusió per al conjunt de dades iris

		Classe real			
		Iris setosa	*Iris versicolor*	*Iris virginica*	Total
Predicció	*Iris setosa*	7	1	0	8
	Iris versicolor	1	8	0	9
	Iris virginica	2	1	10	13
	Total	10	10	10	30

La precisió es correspon amb el nombre de classificacions correctes (suma de valors de la diagonal) dividit pel total d'instàncies. Segons els valors de la taula, la precisió és: (7 + 8 + 10) / 30, que correspon al percentatge 83,3%. També és interessant analitzar la classificació per cada classe o *true positive rate* (TPR). En la classe *Iris setosa*, TPR = 7 / 10, és a dir, es prediuen correctament 7 de les 10 flors de tipus *Iris setosa* del conjunt de dades.

És freqüent utilitzar una mètrica menys optimista de la precisió del model, anomenada Kappa (Cohen, 1960). Aquesta mètrica calcula el percentatge d'encerts més enllà del que es podria aconseguir fent prediccions merament a l'atzar. La seva formulació és:

$$Kappa = \frac{precisió - precisió\ \ a\ \ l'atzar}{1 - precisió\ \ a\ \ l'atzar}$$

Es pot mesurar la precisió del classificador si hagués fet les classificacions a l'atzar de la forma següent. Per exemple, per a la flor *Iris setosa*, la probabilitat que una flor sigui d'aquest tipus és 10 / 30 i la probabilitat que el classificador esculli *Iris setosa* és 8 / 30. Si el classificador és a l'atzar, la probabilitat d'encert s'obté a partir de la multiplicació de les dues probabilitats: $10 \cdot 8 / 30^2$. Per tant, fent el càlcul per a cada tipus de flor, la precisió d'un classificador a l'atzar seria:

$$\frac{8 \cdot 10 + 9 \cdot 10 + 13 \cdot 10}{30^2} = \frac{300}{900} = 0,333$$

El valor Kappa és:

$$Kappa = \frac{0,833 - 0,333}{1 - 0,333} = 0,75$$

Com s'observa, si restem la contribució de l'atzar en la classificació, el classificador passa d'un percentatge d'encert aparent del 83,3% a un percentatge del 75%.

A partir de la matriu de confusió es poden extreure altres mètriques, com la sensitivitat i l'especificitat, quan el problema de classificació és binari (dues classes). Hi ha altres consideracions que cal tenir en compte quan es fan servir mètriques de precisió. Per exemple, si el problema està molt desequilibrat, la precisió no és una mètrica adequada de la qualitat del model.

Predicció de frau

Prenent com a exemple el problema de predicció de frau en targetes de crèdit, aquest pot estar desequilibrat, ja que les transaccions no fraudulentes superen en nombre les fraudulentes. Si el percentatge de transaccions fraudulentes és del 99,5%, un model que sempre predigui la classe majoritària tindria una precisió del 99,5%.

Nota

Per a més detalls sobre aquests aspectes, podeu consultar el treball de Stapor (2017).

2.2. Regressió

En models de regressió es prediu un valor numèric en lloc d'una classe que pertany a un conjunt finit de categories. En aquest cas, no té sentit parlar de precisió, ja que no es classifica la instància com a pertanyent a una categoria o una altra. Alguns exemples de regressió són la predicció del volum de vendes d'un producte o la predicció de l'àrea cremada de bosc (Lichman, 2013).

La qualitat del model és el grau d'aproximació del valor predit al valor real. Es disposa d'un conjunt de valors predits p_1, p_2 ... p_n i els seus corresponents valors reals a_1, a_2 ... a_n. La mètrica més bàsica és calcular la suma de les diferències entre p_i y a_i en valor absolut i dividir-les per n (el nombre de prediccions realitzades). Aquesta és la mesura de l'error absolut mitjà. No obstant això, la mesura més habitual és l'arrel de l'error quadràtic mitjà (*Root Mean-Squared Error*), que es calcula segons la fórmula següent:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

L'inconvenient de la mesura és que tendeix a exagerar els efectes dels valors extrems. A més, un valor determinat no té significat de forma aïllada i s'ha de contextualitzar en relació amb els valors de predicció. No és el mateix un error de valor 50 en relació amb prediccions al voltant de 100 que sobre prediccions al voltant de 10.000. L'arrel de l'error quadràtic mitjà relatiu divideix el valor en relació amb els valors que es manegen. També es poden fer servir altres

mètriques, com el coeficient de correlació de Pearson entre els valors predits i els valors reals. Per a més detalls sobre mètriques de regressió, podeu consultar el llibre de Witten, Frank i Hall (2011).

2.3. Agrupació

L'avaluació dels mètodes d'agrupació o segmentació és més complexa que en classificació o regressió perquè no es disposa d'un valor de referència amb el qual comparar. Un criteri clàssic és mesurar la qualitat de l'agrupació sobre la base del grau de cohesió dins del grup en relació amb les distàncies entre grups.

L'índex de Davies-Bouldin (1979) presenta una mesura de cohesió enfront de distància:

$$DB = \frac{1}{k}\sum_{\substack{i=1 \\ i\neq j}}^{k} \max\left(\frac{\sigma_i + \sigma_j}{d\ (C_i, C_j)}\right)$$

on k és el nombre de grups, C_x denota el centroide de cada grup, σ_x és la distància mitjana de tots els elements del clúster al seu centroide respectiu C_x i $d\ (C_i, C_j)$ és la distància entre els centroides C_i y C_j. Com més baix sigui el valor d'aquest índex, millor serà l'agrupació.

2.4. Associació

La identificació de patrons d'associació entre els atributs d'un conjunt de dades és, igual que en l'agrupació, un cas en què no hi ha un valor de referència. Les mesures de qualitat es basen en criteris de freqüència, representativitat i novetat de les associacions identificades.

Per exemple, suposem que s'analitzen les regles d'associació en la compra de llibres en una llibreria per internet. La sintaxi típica de les regles d'associació consisteix en un antecedent X i un consegüent Y. Així, la regla següent:

Statistics, *Data analytics* → *Statistics for big data*

es llegeix com: «Si el client compra el llibre *Statistics* i el llibre *Data analytics*, llavors també comprarà *Statistics for big data*».

Es fan servir dues mètriques habituals per a mesurar la qualitat de l'associació: el suport i la confiança. El suport és el percentatge de compres en què aquests tres llibres es compren conjuntament:

$$\text{Suport}\ \ (X \rightarrow Y) \ \ = \ \ \text{suport}\ \ (\text{X}\cup\text{Y})$$

on suport(X∪Y) es refereix a la unió dels elements que formen part de l'antecedent i del consegüent. Per exemple, si entre 1.000 compres, 75 contenen aquests tres llibres alhora, llavors el suport és 75 / 1000 = 0,075 (7,5 %).

La confiança es calcula:

$$\text{Confiança}\ \ (X \rightarrow Y) \ \ = \ \ \frac{\text{suport}\ \ (X \cup Y)}{\text{suport}\ \ (X)}$$

En l'exemple que ens ocupa, hem vist que es compren 75 vegades entre 1.000 els tres llibres esmentats. Suposem que 120 vegades es compren els llibres *Statistics* i *Data analytics*. Així doncs, la seva confiança és 75 / 120 = 0,625.

La confiança es pot interpretar com l'estimació de la probabilitat de trobar el consegüent si es produeix l'antecedent. És a dir, la probabilitat condicionada p(Y|X).

Una altra mesura és el *lift* que compara la freqüència d'un patró observat en relació amb l'expectativa de trobar aquest patró per atzar. Es calcula de la manera següent:

$$\text{Lift} = \frac{\text{suport}\ \ (X \cup Y)}{\textit{suport}\ \ (X)\ \textit{suport}\ \ (Y)}$$

Valors de *lift* propers a 1 impliquen una alta probabilitat de trobar el patró per atzar en el conjunt de dades.

3. Avaluació del model

Quan s'avalua un model d'anàlisi de dades, l'objectiu és tenir una mesura de la qualitat del model quan es faci servir en explotació o producció. Per exemple, si un model prediu el nombre de vendes d'un determinat producte cada dia, s'espera que quan el sistema estigui implantat, la predicció que faci sigui precisa. En general, el model no serà perfecte, i per això, cal conèixer quina precisió (o marge d'error) tindrà el sistema quan s'utilitzi en la seva operació diària. A més, hi ha un segon motiu pel qual cal avaluar la qualitat del model: quan és necessari ajustar-lo per a un màxim rendiment, parametritzant adequadament el model triat o escollint el millor model entre diversos de disponibles.

En el cicle de vida de l'analítica de dades, hi ha dues fases ben diferenciades:

- l'**entrenament** del model
- les **proves**

En la fase d'entrenament es construeix el model (correspon a la fase 3 de la metodologia presentada en l'apartat 1). En la fase de proves, s'avalua el model (fase 4). La construcció del model es fa sobre una mostra de dades. Però, com s'han de realitzar les proves? Una opció seria fer-les sobre la mateixa mostra de dades. Però, com endevinareu, aquest mètode donarà com a resultat una estimació esbiaixada positivament, és a dir, l'avaluació serà optimista. I la causa d'això és que mentre avaluem la qualitat del model en el mateix conjunt de dades que s'ha fet servir per a construir el model, aquest model es farà servir en dades desconegudes en el moment de la producció. Per consegüent, el model serà menys precís quan operi amb dades reals. Val a dir que la diferenciació entrenament-proves existeix principalment en models de tipus classificació i regressió, i són poc habituals en els d'agrupació i associació, ja que en els dos últims no hi ha un valor de referència amb el qual comparar.

En aquest apartat es defineix el problema de l'estimació de l'avaluació del model i es revisen els principals estimadors.

3.1. Estimació de la precisió d'un model

L'objectiu de provar un model és conèixer-ne la precisió o error quan el model operi en dades reals. La precisió o error són *desconeguts a priori*, ja que no es disposen de les dades amb les quals el sistema operarà realment. D'ara endavant, farem servir el terme *precisió* per a referir-nos a la mètrica d'interès que es vulgui avaluar. Aquesta pot ser qualsevol de les mètriques introduïdes en l'apartat 2. Atès que la precisió real es desconeix *a priori*, tan sols se'n pot fer una estimació a partir del conjunt de dades disponibles durant la construcció

del model. Anomenem aquests dos valors *precisió calculada en la mostra* i *precisió real*. Atès que l'estimació es fa sobre una mostra de la població, hi haurà un *biaix* i una *variància*.

3.2. Biaix i variància de l'estimador

L'estimador es pot considerar una variable aleatòria *Y* que estima un paràmetre *p* d'una població. El biaix de l'estimador es defineix com la diferència entre l'esperança de la variable Y i el paràmetre que cal estimar:

$$E\ (Y) - p$$

La variància de la variable aleatòria es defineix com:

$$Var(Y) = E((Y\text{-}p)^2)$$

La variància es pot interpretar com l'amplitud o dispersió de la distribució al voltant de la mitjana.

3.3. Estimador basat en la mostra d'entrenament

La fórmula més senzilla per a estimar la precisió d'un model és calcular-la sobre la mostra de dades disponibles. És a dir, s'estima la precisió real a partir de la precisió obtinguda pel model sobre el conjunt de dades d'entrenament, això és, les mateixes dades amb què s'ha entrenat.

Com que el conjunt de dades de la mostra sol ser petit, la generalització de la precisió calculada sobre la mostra a la precisió de la població serà esbiaixada positivament. És a dir, la precisió en el conjunt d'entrenament serà més gran que la que el sistema tindrà en explotació (quan s'apliqui a instàncies o casos no usats directament en l'entrenament). Per tant, diem que la precisió d'entrenament és un estimador esbiaixat de manera optimista. Estimar la precisió en un conjunt de dades no utilitzades en l'entrenament és un estimador més bo de la precisió real.

3.4. Mètode de retenció

El mètode de retenció (*holdout*) divideix el conjunt de dades disponibles en dos subconjunts: el conjunt d'entrenament i el conjunt de prova (test). S'assumeix que els dos subconjunts de dades són mostres representatives de la població d'interès. El model es construeix amb el conjunt de dades d'entrenament i després se n'estima la qualitat en el conjunt de test. D'aquesta manera, el model es prova en un conjunt de dades diferent del de l'entrenament. Aquest estimador no està esbiaixat de manera optimista com succeeix amb l'estimador basat en les dades d'entrenament.

Habitualment, la mida recomanada per a aquests subconjunts és del 70% de les dades en entrenament i el 30% de dades per al test. El codi R següent fa servir la funció *createDataPartition* del paquet *caret* per a crear dos subconjunts de dades a partir del conjunt original:

```
intrain<-createDataPartition(y=data$class,  p=0.70)
train <- data [intrain,]
test <-  data [-intrain,]
```

L'exemple està pensat per a separar dades segons *holdout* assumint que pertanyen a un problema de classificació. La variable *data* és la mostra de dades i *class* és la classe associada als exemples de la mostra. En el subapartat 3.6 veureu per què s'introdueix la variable *class* a la partició. La funció *createDataPartition* retorna els índexs de les dades seleccionades, corresponents a una mostra aleatòria del 70% de les dades. Amb això, es construeix la mostra *train* i amb les dades restants (les que no són indexades per *intrain*) es construeix la mostra *test*.

L'estimador *holdout* presenta diversos inconvenients. En primer lloc, si es disposa d'una mostra de dades petita, reservar algunes dades per a les proves implica que les dades d'entrenament es redueixen. Conseqüentment, la qualitat del model construït serà pitjor en no poder entrenar amb totes les dades disponibles. L'estimador estarà esbiaixat negativament. Un altre inconvenient del mètode és que té una elevada variància. Si es repeteix l'experiment amb un mostreig diferent de dades en l'entrenament i en el test, l'error obtingut diferirà de manera significativa entre una prova i una altra.

Si la mida de les dades disponibles és prou gran, l'estimador presenta menor biaix i variància que si les dades són escasses i el seu ús podria ser acceptable. Per a mostres petites, una alternativa per a minimitzar el biaix és fer diverses proves repetides de *holdout*. El mètode de retenció repetida (*repeated holdout*) realitza diversos mostreigs de conjunts entrenament-test i per a cada un entrena el model i n'avalua la precisió. La precisió final estimada serà la mitjana de la precisió comesa en cada mostreig.

El mètode de retenció repetida ens introdueix en el món del remostreig (*resampling*) que expliquem a continuació.

3.5. Remostreig

Les tècniques de remostreig (*resampling*) consisteixen a mostrejar múltiples vegades el conjunt de dades original per a obtenir diferents parells de subconjunts d'entrenament i test amb l'objectiu d'obtenir millors estimadors.

3.5.1. Validació creuada

La validació creuada (*cross-validation*) consisteix a dividir el conjunt de dades en k particions disjuntes denominades *folds*. El mètode s'entrena k vegades. Per a la iteració i on $1 \leq i \leq k$, es construeix el model a partir del subconjunt de dades format per totes les particions excepte la i, i es prova el model en el subconjunt i. La precisió final és la mitjana de la precisió obtinguda en les k proves. Amb això s'aconsegueix que el mètode es provi amb tots els exemples disponibles i alhora s'estreni amb instàncies independents de les proves. Un valor habitual de k és 10 (*10-fold cross-validation*), amb la qual cosa s'entrena cada vegada amb el 90% de la mostra original. La taula 3 especifica un exemple d'execució d'aquest mètode de remostreig per al problema de classificació de flors iris (Lichman, 2013).

Tabla 3. Estimació de la precisió mitjançant el mètode de validació creuada (k=10)

	Subconjunt d'entrenament	**Subconjunt de test**	**Precisió (acc)**
1	{2-10}	1	0,92
2	{1, 3-10}	2	0,99
3	{1-2, 4-10}	3	0,98
4	{1-3, 5-10}	4	0,89
5	{1-4, 6-10}	5	0,94
6	{1-5, 7-10}	6	0,96
7	{1-6, 8-10}	7	0,98
8	{1-7, 9-10}	8	0,95
9	{1-8, 10}	9	0,93
10	{1-9}	10	0,97
Precisió mitjana ($\overline{acc}$)			0,951
Desviació estàndard (SD)			0,031

El codi següent il·lustra com executar el mètode de veïns més propers k-NN amb validació creuada 10-CV:

```
ctrl<-trainControl(method="repeatedcv", repeats=1)
knn<-train(Species~., data=iris, method="knn",
preProc=c("center","scale"),
trControl=ctrl)
knn
## k-Nearest Neighbors
##
## 150 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## Pre-processing: centered (4), scaled (4)
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 135, 135, 135, 135, 135, 135, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.9600000  0.94
##   7  0.9533333  0.93
##   9  0.9533333  0.93
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```

Com es pot observar, es fan servir 10 particions i cadascuna d'elles té 135 dades (de les 150 de la mostra original) en el conjunt d'entrenament. El codi prova tres valors del paràmetre *k* (nombre de veïns) del mètode k-NN, executant una validació creuada per a cada valor del paràmetre.

Nota

k-NN és el nom que rep el mètode de mineria de dades anomenat veïns més propers. En aquest, el paràmetre k configura el nombre de veïns que el mètode usa per a calcular el resultat. No us confoneu amb el paràmetre k del mètode de validació creuada, que es refereix al nombre de particions de la mostra de dades. Ambdós mètodes usen el mateix nom per als seus respectius paràmetres.

Per a un nombre de particions *k* elevat, el biaix del mètode és inferior al mètode de retenció, ja que el model s'entrena amb un percentatge elevat de les dades (per a $k = 10$, s'entrena amb un 90% de les dades). No obstant això, continua essent un estimador esbiaixat negativament, ja que el model s'entrena amb un $(k - 1) / k$ 100 % de les dades, i s'espera que serà pitjor que un model entrenat amb tot el conjunt de dades. El biaix es pot minimitzar amb el mètode LOO que s'explica a continuació.

A més del biaix, els resultats de la validació creuada tenen una variància elevada. Si s'executen dues proves de validació creuada amb particions diferents, els resultats de cadascuna d'elles seran diferents. Per a minimitzar la variància, es pot repetir la validació creuada diverses vegades amb diferents submostreigs. En aquest cas, el mètode s'anomena validació creuada repetida (*repeated cross-validation*). L'inconvenient és l'alt cost computacional que representa entrenar el model repetides vegades (concretament $n \cdot k$ vegades, on n són les vegades que es repeteix i k és el nombre de particions).

3.5.2. LOO

Un cas particular del mètode de validació creuada és l'anomenat *Leave-One-Out* (LOO). En aquest mètode, cada mostra de test es correspon amb un únic exemple. Així, el model es construeix n vegades, on n és el nombre d'instàncies de la mostra original. A cada iteració, el model s'entrena amb n-1 exemples i es prova amb l'exemple restant. L'avantatge del LOO és que el model es construeix amb pràcticament tots els exemples disponibles (n-1), alhora que es prova amb tots els exemples de test, igual que les proves en la validació creuada. No

obstant això, l'alt cost computacional fa que sigui un mètode poc aplicable excepte en els casos en què es disposi d'un conjunt de dades molt reduït. El codi següent mostra l'aplicació del mètode LOO fent servir R:

```
ctrl<-trainControl(method="LOOCV", repeats=1)
knn<-train(Species~., data=iris, method="knn",
preProc=c("center","scale"),
trControl=ctrl)
knn
## k-Nearest Neighbors
##
## 150 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## Pre-processing: centered (4), scaled (4)
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 149, 149, 149, 149, 149, 149, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.9466667  0.92
##   7  0.9600000  0.94
##   9  0.9533333  0.93
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was k = 7.
```

3.5.3. *Bootstrap*

El mètode *bootstrap* es basa en el remostreig amb reemplaçament. Els anteriors remostreigs es realitzaven sense reemplaçament: així que un exemple era seleccionat per a un subconjunt, no podia ser seleccionat de nou. Això generava conjunts disjunts d'entrenament i test. A *bootstrap*, cada dada pot ser seleccionada diverses vegades per a formar part del subconjunt d'entrenament. En l'anomenat mètode *0,632 bootstrap*, se seleccionen n instàncies d'un conjunt de dades de mida n per a formar part d'un conjunt d'entrenament, que serà de la mateixa mida que l'original. Com que algunes instàncies no sortiran seleccionades, es fan servir aquestes per a formar part del conjunt de prova. El nom del mètode prové de la probabilitat que una instància no formi part del conjunt d'entrenament, que es calcula com:

$$\left(1-\frac{1}{n}\right)^n \approx e^{-1} = 0{,}368$$

Així doncs, la mida del conjunt de test és del 36,8% i el conjunt d'entrenament, el 62,3% de la mostra original aproximadament.

L'estimador és pessimista perquè el conjunt d'entrenament només conté el 62,3% de les instàncies. Per a compensar-ho, la precisió final es calcula com:

$$\text{acc} = 0{,}632 \cdot \text{acc}_{test} + 0{,}368 \cdot \text{acc}_{train}$$

El codi en R següent executa el mètode k-NN fent servir *bootstrap*:

```
ctrl<-trainControl(method="boot632")
knn<-train(Species~., data=iris, method="knn",
preProc=c("center","scale"))
knn
## k-Nearest Neighbors
##
## 150 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## Pre-processing: centered (4), scaled (4)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 150, 150, 150, 150, 150, 150, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.9427889  0.9136946
##   7  0.9359550  0.9033683
##   9  0.9410505  0.9110650
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```

3.6. Estratificació

En introduir el mètode de retenció (*holdout*) s'ha especificat que les mostres d'entrenament i prova han de ser representatives de la població d'interès. En primer lloc, la mostra original de dades hauria de ser representativa. Si no ho és, en fer el remostreig en els dos subconjunts d'entrenament i proves es preservarà la mateixa falta de representativitat. En el mòdul 1, s'ha esmentat el mostreig probabilístic per a seleccionar una mostra de dades representativa. Però tot i disposar d'una mostra original representativa, pot succeir que a l'hora de remostrejar en els dos subconjunts es produeixin biaixos. Prenent com a exemple el problema de classificació iris, podria succeir que el conjunt d'entrenament només contingui mostres de dues de les tres flors, la qual cosa impossibilitaria extreure un model que classifiqui la flor que no està representada. A més, la flor que no està present en el conjunt d'entrenament estaria sobrerepresentada en el conjunt de proves, amb la qual cosa l'error de classificació seria elevat.

El problema es presenta especialment amb mostres petites. Tot i que el mètode de remostreig garanteixi que cada instància del conjunt de dades té la mateixa probabilitat de ser seleccionada en el subconjunt d'entrenament, hi pot haver biaixos inevitables. La solució passa, igual que en el mostreig de la mostra original, per realitzar un mostreig estratificat. En aquest, els estrats són cadascuna de les classes. En el cas del problema iris, se separen les instàncies en tres grups, un per a cada classe. Després, es fa un mostreig aleatori dins de cada classe. Així, es garanteix que en els subconjunts d'entrenament i de test es preserva la proporció de classes del conjunt de dades original. I amb això donem també resposta a la línia de codi especificada anteriorment, quan s'ha introduït el mètode *holdout*:

```
createDataPartition( y=data$class, p=0,70 )
```

Com es pot deduir, s'indica al mètode de partició *holdout* que consideri la classe associada als exemples del conjunt *data* per a aplicar estratificació. L'estratificació es pot aplicar en tots els mètodes de remostreig.

4. Validació estadística

En la secció anterior, s'ha tractat l'estimació de la qualitat o rendiment dels mètodes d'anàlisi de dades mesurada sobre una mostra de dades. Atès que habitualment el nombre de dades per a entrenar l'algoritme i provar-ne la qualitat és limitat, s'han introduït els mètodes de remostreig de dades amb l'objectiu d'obtenir una estimació precisa de la qualitat del model. Un cop es disposa d'aquesta estimació, es planteja la qüestió de com procedir a partir d'aquest moment i això dependrà del context i l'objectiu del disseny experimental. Els escenaris possibles són:

- Generalització dels resultats
- Comparació entre diversos models de dades en un problema determinat
- Comparació entre diversos models de dades en múltiples problemes

1) Generalització dels resultats. El primer punt es refereix a estudiar com es generalitzen els resultats a noves dades. És a dir, consisteix a inferir l'estimació de la qualitat del model obtinguda sobre la mostra a la qualitat que el model tindria en explotació, quan es fa servir en dades futures, que pertanyen a la població d'interès. La formulació general seria:

> Si tenim un conjunt de valors de precisió acc_1, acc_2 ... acc_L, cada un d'ells referit a un subconjunt de test, quina seria la precisió *acc* del model quan s'aplica a dades reals?

En termes estadístics, consisteix a inferir el valor del paràmetre poblacional a partir del paràmetre mostral. Així, es considera cada valor acc_i com el valor d'un individu d'una mostra. Es calcula la mitjana i desviació mostral i, tot seguit, es dedueix l'**interval de confiança** de la mitjana de la població. El mètode no està exempt de debat, ja que si els valors acc_i provenen de mostres no independents, com el cas de la validació creuada repetida, no seria apropiat. Es recomana llavors calcular els intervals de confiança a partir de *bootstrap*. L'aprofundiment en aquest debat està fora de l'abast d'aquest mòdul. Podeu consultar el treball de Vanwinckelen i Blockeel (2012) per a aprofundir en aquest aspecte.

2) Comparació entre models en un problema determinat. El segon escenari és molt habitual en experimentació científica de dades i en àmbits aplicats. Es tracta de comparar el rendiment de dos o més mètodes en un problema determinat. La formulació seria:

> Si tenim *K* models diferents, quin d'ells ofereix un millor rendiment per al problema determinat?

Aquests models diferents poden representar diferents algoritmes d'anàlisi de dades, com el k-NN i els arbres de decisió, o poden provenir d'un mateix algoritme amb diferents paràmetres de configuració (per exemple, valors dife-

rents del nombre de veïns *k* per a l'algoritme k-NN). En termes estadístics, la pregunta formulada es tradueix en un **contrast d'hipòtesis per a dues o més mostres**. Si els diferents models es comparen sota les mateixes condicions (el mateix remostreig de les dades, les mateixes particions i les mateixes mètriques), podem considerar que les mostres estan aparellades. En aquest cas, s'apliquen **contrastos d'hipòtesis per a dues o més mostres aparellades**. Cal fer la distinció segons si comparen dos algoritmes o més de dos. L'elecció dels mètodes de validació estadística també diferirà segons aquests dos casos. Una última consideració és decidir entre tests paramètrics i tests no paramètrics. La taula 4 resumeix els principals tests estadístics.

Taula 4. Proves estadístiques per a la comparació de models d'anàlisi de dades

Mostres	Comparació	Tests paramètrics	Tests no paramètrics
Aparellades	Dos models	• Test t per a mostres aparellades • Test z	• Prova de la suma de rangs de Wilcoxon • Test de signe • Test de McNemar
	Més de dos models	• Anàlisi de variància (ANOVA) • Test z	• Prova de Kruskall-Wallis • Prova de Friedman

3) Comparació entre models en múltiples problemes. Aquest escenari és la generalització del segon escenari a múltiples problemes. Pertany bàsicament al domini dels estudis científics més que a dominis aplicats. L'interès inherent és trobar models d'anàlisi de dades que ofereixin rendiments superiors als existents. Per a això, l'experimentador prova el model en relació amb altres models estàndard d'anàlisi de dades i extreu conclusions generals sobre el rendiment del model en comparació amb els altres. Un altre objectiu podria ser comparar el rendiment de models existents, sense un interès destacat en cap d'ells *a priori* (Lim, Loh i Shih, 2000). En canvi, en l'àmbit aplicat, l'interès més comú és desenvolupar el model òptim per a un problema determinat.

La formulació general de la comparació de múltiples models en múltiples problemes és:

> Si tenim *N* problemes i *K* models, quin dels models ofereix un millor rendiment?

La formulació requereix alguns matisos. En primer lloc, pot ser que no hi hagi un mètode que tingui un rendiment general millor. Per aquest motiu, primer es formula si hi ha diferències entre tots els mètodes. *A posteriori*, si hi ha evidència de diferències entre ells, s'identifica quin dels mètodes és el que destaca sobre els altres. Aquest tipus de formulació parteix d'una hipòtesi que un determinat mètode és millor que els altres. I això sorgeix d'un estudi experimental de ciència de dades, en què l'investigador ha proposat un nou mètode que es contrasta amb altres d'existents.

En traslladar aquesta formulació en termes de validació estadística, trobem el mateix tipus d'aproximacions que per a la comparació de mètodes en un problema determinat. La distinció rau en si es comparen dos o més mètodes, i si es poden aplicar tests paramètrics o no paramètrics, tal com es mostra a la taula 4. Les diferències metodològiques entre tots dos escenaris es refereixen a l'obtenció dels valors de qualitat de cada model. Per a un problema determinat (escenari 2), els valors que cal comparar provenen dels diferents submostreigs de dades (*cross-validation*, *bootstrap*...), tal com s'ha detallat en l'apartat anterior. Per a múltiples problemes, els valors que es comparen provenen de la precisió mitjana del model en diferents problemes. Aquesta diferència pot determinar l'elecció dels tests estadístics apropiats (Demsar, 2006).

4.1. Comparació entre dos models

En aquesta secció es descriuen els principals mètodes de validació estadística aplicables a la comparació de dos models de dades. S'assumeix que els models presenten respectivament unes mesures de precisió x_1, x_2, ... x_N e y_1, y_2, ... y_N. Aquestes dades conformen les dues mostres que cal comparar, assumint que són dues mostres aparellades. Els mètodes que veurem a continuació són aplicables a la comparació entre dos models per a un problema determinat o per a múltiples problemes. Com s'ha comentat, la diferència rau en el fet que les dades de precisió provenen dels diferents submostreigs de dades (per a un problema determinat) o de diferents problemes. Al final d'aquesta secció, es detallen els matisos existents segons si la comparació entre tots dos models es produeix sobre un únic problema o sobre diversos problemes.

4.1.1. Test *t* per a dues mostres aparellades

Probablement ja coneixeu el test t aparellat. És un test molt habitual en dissenys experimentals. Aplicat al disseny experimental de models de dades, assumim que es disposa dels resultats de dos models en un conjunt d'*N* problemes: x_1, x_2, ... x_N e y_1, y_2, ...y_N. El test *t* verifica si la diferència en el rendiment sobre els conjunts de test és significativament diferent de zero.

Es pot assumir que les mostres estan aparellades si els resultats de cada model s'han computat sobre els mateixos conjunts de dades. Així, es calculen les diferències entre tots dos models: $d_i = x_i - y_i$. La mitjana de les diferències és la diferència de les mitjanes: $\overline{d} = \overline{x} - \overline{y}$ i es comporta com una distribució *t* amb *N*-1 graus de llibertat. La hipòtesi nul·la estableix que la mitjana de les diferències és zero. L'estadístic *t* es calcula segons:

$$t = \frac{\overline{d}}{\sqrt{S_d^2 / N}}$$

on $\overline{d}$ és la mitjana de les diferències, S_d^2 és la variància de les diferències i N, la mida de la mostra.

El test t presenta dues restriccions. La primera és que assumeix una distribució normal de les dades, la qual cosa habitualment no es compleix. D'altra banda, el test requereix que les dades siguin independents. Si s'aplica per a la comparació de dos mètodes en un únic problema, aquesta assumpció no es compleix (Salzberg, 1997). El motiu és que els tests d'entrenament i de prova no són totalment independents, com passa, per exemple, si es fa servir una validació creuada repetida. L'alternativa és usar tests no paramètrics com els que es descriuen a continuació.

4.1.2. Prova de suma de rangs de Wilcoxon

Demsar (2006) recomana utilitzar la prova de suma de rangs de Wilcoxon (1945) o el test de signe. Aquests són els equivalents no paramètrics del test t aparellat, tal com es mostra a la taula 4. A continuació, s'explica l'aplicació de la prova de Wilcoxon, ja que el seu ús és més estès.

S'aplica de la manera següent:

1) Es calculen les diferències entre els dos mètodes que cal comparar.

2) Aquestes diferències s'ordenen en ordre creixent i en valor absolut. S'assigna un rang a cada valor, en funció de la seva posició. En cas d'empat, es calcula el rang mitjà.

3) Es calcula *R*+ com la suma dels rangs en què el primer model millora el segon. De manera inversa, es calcula R-. Si una distància és igual a 0, es reparteix entre *R*+ i *R*-.

4) Es calcula T = mínim (*R*+ ,*R*-). El seu valor crític es pot calcular per mitjà de taules (fins a N igual a 25). Per a valors grans d'N, es fa servir l'estadístic z que es distribueix normalment:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

En la taula 5 es detalla un exemple per a dos conjunts de mostres.

Taula 5. Exemple de càlcul del test de suma de rangs de Wilcoxon

x	y	d	rang
10,54	12,04	-1,50	-10
10,70	11,75	-1,05	-8
10,23	11,22	-0,99	-7

x	y	d	rang
10,43	10,18	0,25	3
10,53	11,34	-0,81	-6
10,98	9,73	1,25	9
10,62	10,67	-0,05	-1
10,81	11,11	-0,30	-4
10,40	10,24	0,16	2
10,50	10,87	0,37	-5

La suma de rangs positius *R+* és igual a 14 i la de rangs negatius *R-* és igual a 41. Per tant, $T = min\ (R+\ ,R-) = 14$. El valor de *z* és -1,38 i el valor *p* corresponent de les taules de distribució normal en una prova bilateral és 0,1688, amb la qual cosa no es pot rebutjar la hipòtesi nul·la que no hi ha diferències entre tots dos models.

El codi següent mostra el càlcul de la prova de Wilcoxon en *R*:

```
wilcoxsign_test(x1~x2)
##
##  Asymptotic Wilcoxon-Pratt Signed-Rank Test
##
## data:  y by x (pos, neg)
##   stratified by block
## Z = -1.376, p-value = 0.1688
## alternative hypothesis: true mu is not equal to 0
```

Per a valors petits d'*N* (mida de la mostra) és més recomanable consultar taules de càlcul directe del valor *p* en lloc del valor *z*, ja que ofereix una aproximació millor.

4.1.3. Prova de McNemar

La prova de McNemar és de tipus no paramètric i s'aplica comparant el resultat dels dos models per a cada instància del conjunt de dades. Es construeix una taula com la següent:

Taula 6. Càlcul del test de McNemar

n00 Nombre d'instàncies mal classificades per tots dos models	n01 Nombre d'instàncies classificades incorrectament pel mètode X, però no per Y
n10 Nombre d'instàncies classificades incorrectament pel mètode Y, però no per X	n11 Nombre d'instàncies classificades correctament per tots dos models

on $n = n_{00} + n_{01} + n_{10} + n_{11}$ és el nombre total d'exemples del conjunt de dades. El test està pensat per a mostreigs del tipus *holdout*. Sota la hipòtesi nul·la, tots dos models haurien de tenir el mateix error ($n_{01} = n_{10}$). Per a això, es compara el nombre d'errors sota la hipòtesi nul·la en relació amb els valors observats:

$$\frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Aquest valor es distribueix segons χ^2 amb 1 grau de llibertat. Si la hipòtesi nul·la és certa, la probabilitat que aquest valor sigui més gran que $\chi^2_{1,0.95} = 3{,}84$ és menor que 0,05. Per tant, es rebutjaria la hipòtesi nul·la a favor de la hipòtesi que tots dos models tenen un rendiment diferent.

Segons Dietterich (1998), aquest test presenta baix error tipus I. L'error tipus I és la capacitat de detectar diferències quan no n'hi ha. Com a inconvenient, el test aplicat amb un mostreig *holdout* no té en compte la variabilitat deguda a l'elecció del conjunt d'entrenament ni la variabilitat deguda a l'algoritme. Dietterich recomana fer servir aquest test només quan aquestes variabilitats són petites.

4.1.4. Comparació de dos models en un o diversos problemes

Demsar (2006) argumenta que la comparació de dos models en un únic problema és insegura a causa de la manca d'independència entre els valors de precisió obtinguts, especialment en els casos en què els conjunts d'entrenament i prova provenen de remostreigs iteratius. En aquests casos, Demsar recomana l'ús de la prova de McNemar, ja que no realitza assumpcions de distribució de les dades. No obstant això, com s'ha comentat, no té en compte la variabilitat en l'elecció del conjunt d'entrenament ni la variabilitat interna de l'algoritme d'aprenentatge que extreu el model de les dades.

Si es comparen dos models en múltiples conjunts de dades, les mesures són difícilment comparables, ja que les dades provenen de fonts diverses. Per a això, Demsar recomana el test de suma de rangs de Wilcoxon.

4.2. Comparació de múltiples models

En la comparació múltiple de models, es volen comparar múltiples mètodes K en N conjunts de dades diferents. Per a cada mètode i cada problema es disposa d'una estimació de la precisió o qualitat del model. Es descarta de moment l'anàlisi de variància sobre el remostreig i s'assumeix que el valor de rendiment de cada mètode és una estimació precisa. Aquesta estimació pot provenir d'un mètode de validació creuada o un altre mètode de remostreig.

En executar algoritmes en múltiples problemes s'obtenen mesures independents, a diferència de la comparació sobre un únic problema. Per tant, les comparacions són més simples que en la comparació amb el mateix conjunt de dades (Demsar, 2006).

4.2.1. Correcció de Bonferroni

Un error habitual en comparar múltiples models en diversos conjunts de dades és utilitzar el test t de mostres aparellades (Kuzon, Urbancheck i McCabe, 1996). El problema que es deriva de fer servir el test t aparellat en comparació múltiple és que l'error de tipus I que es comet es propaga per a cada comparació per parells i, per tant, s'hauria d'ajustar el nivell de significació α a un valor més restrictiu (Salzberg, 1997).

Aquest és el mètode que segueix la correcció de Bonferroni. Si es disposen d'm hipòtesis que cal contrastar, on m és el nombre de parells d'algoritmes que es comparen entre si $m = K\ (K - 1) / 2$, el valor d'α corregit seria el valor original d'α dividit per m. Per exemple, per a quatre models que cal comparar amb un valor de significació α igual a 0,05, el valor corregit d'α és 0,05 / 6 que correspon a 0,0083. Un cop corregit, s'aplica el test t a cada parell de mètodes.

La correcció de Bonferroni és molt restrictiva, la qual cosa deriva en una prova amb poca potència. És a dir, la capacitat per a detectar diferències significatives quan n'hi ha és molt baixa.

4.2.2. ANOVA

El mètode estadístic habitual per a comparacions múltiples és ANOVA, o anàlisi de variància. Les mostres aparellades corresponen a la qualitat de cada model sobre els mateixos conjunts de dades, preferentment fent servir les mateixes particions dels subconjunts d'entrenament i test. La hipòtesi nul·la és que no hi ha diferències entre els diferents mètodes que es comparen.

ANOVA divideix la variabilitat total en la variabilitat entre els models de dades, la variabilitat entre els conjunts de dades i la variabilitat residual. Si la variabilitat entre models és prou gran, rebutgem la hipòtesi nul·la a favor de la hipòtesi alternativa que hi ha diferències entre els models. Us remetem a textos d'estadística bàsica per a conèixer els detalls del test.

Si es rebutja la hipòtesi nul·la, es pot dir que hi ha diferències significatives entre els models, però no es coneix quins d'ells són diferents als altres. Per a això, s'apliquen proves anomenades *post-hoc* (*a posteriori*), o proves de comparacions múltiples. Entre aquests mètodes hi ha el test de Tukey per a la comparació entre tots els models, o el test de Dunnett per a comparar els models

en relació amb un d'ells. Les dues proves són similars al test t, amb la correcció que els valors crítics són més elevats per a assegurar que es manté el valor d'α determinat (com succeeix amb la correcció de Bonferroni).

L'inconvenient d'ANOVA és que s'assumeix que les mostres es distribueixen normalment. Un altre requisit és l'esfericitat (homogeneïtat de variància). Si aquests requisits no es compleixen, el test de Friedman, que és l'equivalent no paramètric, és més aconsellable. A continuació, es descriu amb més detall el test de Friedman, ja que la seva aplicació en el context del disseny experimental en analítica de dades és més estesa.

4.2.3. Test de Friedman

El test de Friedman ordena cada model en funció del seu rendiment en cada problema i calcula la mitjana dels rangs de cada model. L'estadístic es calcula segons la fórmula següent:

$$\chi_F^2 = \frac{12N}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)^2}{4}\right]$$

on R_j^2 és la mitjana dels rangs de cada algoritme, N és el nombre de problemes i K, el nombre de models que es comparen. Per al càlcul de R_j, s'ordena el resultat de cada model per a cada problema, de manera que el model amb millor resultat obté el rang 1, el següent rang 2, etc. El rang total R_j de cada model és la mitjana dels rangs del model en tots els problemes. Sota la hipòtesi nul·la, tots els models són equivalents, és a dir, tots els rangs R_j són iguals. L'estadístic es distribueix segons χ_F^2 amb (K-1) graus de llibertat.

Ivan-Davenport proposen una correcció sobre l'estadístic per a millorar la potència del test:

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}$$

que es distribueix segons F amb K-1 i (K-1)(N-1) graus de llibertat.

El test de Friedman té menys potència que el test ANOVA quan les condicions d'ANOVA se satisfan. Quan es rebutja la hipòtesi nul·la que tots els algoritmes són equivalents, es passa a realitzar un test *post-hoc*.

4.2.4. Proves de comparacions múltiples o proves *a posteriori*

Les proves (*a posteriori*) (*post-hoc*) realitzen comparacions per parells entre els models per a identificar quins d'ells tenen millor rendiment. Les comparacions per parells s'associen dins d'una família d'hipòtesis i, per tant, cal controlar el valor de significació α. L'estadístic per a comparar dos models és:

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{K(K+1)}{6N}}}$$

on R_i i R_j són els rangs respectius de cada mètode, K és el nombre de models que es comparen i N, el nombre de problemes o conjunts de dades.

Test de Nemenyi

El test de Nemenyi es fa servir per a comparar tots els models entre si, de manera anàloga al test Tukey per a ANOVA. Per a això, ajusta el valor d'α dividint pel nombre de comparacions que es realitzen $m = K\,(K - 1) / 2$. Aquest procediment és el més simple, però també el que té menys potència.

Una formulació alternativa del test proposada per Demsar (2006) consisteix a calcular la distància crítica DC:

$$DC = q_\alpha \sqrt{\frac{K(K+1)}{6N}}$$

Per a calcular q_α es fa servir l'estadístic de rangs Student per a infinits graus de llibertat dividit per $\sqrt{2}$. La taula 7 presenta els valors crítics per a α = 0,05, en funció del nombre de models que es comparen entre si.

Taula 7. Distàncies crítiques per al test de Nemenyi i test de Bonferroni-Dunn bilaterals

Nombre de models	2	3	4	5	6	7	8	9	10
Nemenyi $q_{0,05}$	1,960	2,343	2,569	2,728	2,850	2,949	3,031	3,102	3,164
Bonferroni – Dunn $q_{0,05}$ (α/(K-1))	1,960	2,241	2,394	2,498	2,576	2,638	2,690	2,724	2,773

Ajust d'α per a la comparació amb un model de referència

Quan tots els models es comparen amb un model de referència, es poden fer servir altres procediments menys conservadors que s'ajusten amb K-1 en comparar amb un únic algoritme de control en lloc de l'ajust $K\,(K\text{-}1) / 2$. El procediment consisteix a calcular el valor z a partir dels rangs entre els models i el

model de referència, tal com s'ha descrit anteriorment. A continuació, es fa servir la taula de distribució normal per a obtenir el valor de probabilitat *p*. Aquest valor es compara amb el valor d'α que es vol.

Hi ha diverses proves que ajusten el valor d'α de maneres diferents. A continuació, es descriuen les més habituals.

Prova de Bonferroni-Dunn

Aquest test ajusta l'error dividint α amb el nombre de comparacions (*K*-1) o alternativament en el càlcul de distància crítica (DC) amb valors crítics α/(*K*-1). La correcció Bonferroni-Dunn és menys conservadora que el test de Nemenyi.

Prova de Holm

En aquest test s'ordenen les hipòtesis per ordre del valor de *p* corresponent: $p_1<p_2< ... <p_{K-1}$. Llavors es compara cada p_i de la manera següent. Si p_1 és superior a α/(*K*-1), es rebutja la hipòtesi corresponent i es passa a comparar p_2 amb α/(*K*-2). Així successivament fins que un valor p_i no permet rebutjar més hipòtesis, moment en què no es realitzen més comparacions.

Prova de Hochberg

El test de Hochberg ordena els valors $p_1 < p_2 < ... < p_{K-1}$ i comença amb el valor més gran de *p* (p_{K-1}). Es compara el valor p_{K-1} amb α, el següent valor p_2 amb α/2 i així successivament fins que troba una hipòtesi que pugui rebutjar. Llavors, la resta d'hipòtesis amb valors de *p* més petits són rebutjades també.

4.2.5. Exemple de comparació múltiple

Per a il·lustrar l'aplicació de les proves de comparació múltiple, es realitzarà una comparació de $K = 4$ mètodes amb $N = 15$ conjunts de dades. S'assumeix que no es compleixen les condicions d'aplicació de les proves paramètriques, amb la qual cosa s'han d'aplicar les proves no paramètriques equivalents.

A la taula 8 es presenten els resultats de la precisió dels quatre models estudiats en els 15 problemes:

Taula 8. Exemple de comparació múltiple. Resultats de 4 models en 15 problemes

Problema	M1	M2	M3	M4
1	94,97	90,82	97,91	97,22
2	97,66	113,04	94,01	88,33
3	98,41	107,01	97,73	92,19
4	91,50	103,82	96,17	84,75

Problema	M1	M2	M3	M4
5	88,69	83,59	92,39	94,21
6	91,45	97,94	96,52	93,99
7	92,61	91,10	96,19	93,72
8	93,70	93,85	96,82	94,35
9	91,36	99,17	100,09	93,26
10	88,65	90,67	94,48	89,50
11	93,46	91,88	94,22	89,05
12	90,51	98,68	94,91	92,50
13	90,50	99,2	99,67	90,19
14	92,77	95,51	96,25	99,16
15	92,95	96,97	97,84	78,98

En primer lloc, es calculen els rangs de cada model per a cada problema. A la taula següent es resumeixen aquests rangs:

Taula 9. Rangs de la comparació dels 4 models en els 15 problemes

Problema	M1	M2	M3	M4
1	3	4	1	2
2	2	1	3	4
3	2	1	3	4
4	3	1	2	4
5	3	4	2	1
6	4	1	2	3
7	3	4	1	2
8	4	3	1	2
9	4	2	1	3
10	4	2	1	3
11	2	3	1	4
12	4	1	2	3
13	3	2	1	4
14	4	3	2	1
15	3	2	1	4
Rang mitjà	**3,20**	**2,27**	**1,60**	**2,93**

Es calcula el test de Friedman:

$$\chi_F^2 = \frac{12\cdot 15}{4\cdot 5}\left[3{,}20^2 + 2{,}27^2 + 1{,}60^2 + 2{,}93^2 + \frac{4\cdot 5^2}{4}\right] = 13{,}88$$

$$F_F = \frac{(15-1)\cdot 13{,}88}{15\cdot(4-1) - 13{,}88} = 6{,}24$$

F_F es distribueix amb (4-1) i (4-1)*(15-1) = 42 graus de llibertat. El valor crític de $F_{3,42}$ per a α = 0,05 es pot consultar a les taules de distribució *F* o alternativament mitjançant *R*:

```
qf(0.05, k-1, (k-1)*(N-1), lower.tail=FALSE)
## [1] 2.827049
```

El valor crític és 2,83. Com que el valor obtingut *F* = 6,24 és major que el valor crític, es conclou que hi ha diferències significatives. Alternativament, es pot consultar el valor *p* associat a l'estadístic de contrast obtingut:

```
pf(F, k-1, (k-1)*(N-1), lower.tail=FALSE)
## [1] 0.001326882
```

Atès que el valor *p* és inferior a 0,05, es descarta la hipòtesi nul·la que totes les diferències són iguals. A continuació, es calcula la prova de Nemenyi per comparar els models entre si. La forma més senzilla és calcular la distància crítica DC consultant el valor $q_{0,05}$ de la taula 7:

$$DC = 2{,}569\sqrt{\frac{4\cdot 5}{6\cdot 15}} = 1{,}21$$

Perquè hi hagi diferències entre dos parells de mètodes, els seus rangs han de diferir almenys en 1,21. A la taula següent es mostren les diferències de rangs entre cada parell de mètodes.

Taula 10. Diferències entre rangs de l'exemple de comparació múltiple

	1	2	3	4
1	0	0,93	**1,6**	0,27
2	-0,93	0	0,67	-0,67
3	**-1,6**	-0,67	0	**-1,33**
4	-0,27	0,67	**1,33**	0

Com s'observa a la taula, les diferències entre rangs que superen el valor de distància crítica 1,21 són les existents entre el model 1 i 3 i entre el model 3 i 4. La resta de comparacions entre parells de models dona com a resultat el no rebuig de la hipòtesi nul·la.

A continuació, es realitzen les comparacions dels models en relació amb el model 3. Seguint el test Bonferroni-Dunn, la distància crítica és:

$$DC = 2{,}394\sqrt{\frac{4\cdot 5}{6\cdot 15}} = 1{,}13$$

El valor crític d'aquesta prova és menys restrictiu que la prova de Nemenyi. No obstant això, no s'afegeixen noves hipòtesis que es puguin rebutjar.

A continuació, es comparen els models amb el model M3 per contrastar la hipòtesi de si aquest model dona resultats significativament diferents als altres tres. Per a realitzar els càlculs de *z*, es calcula l'error típic:

$$ET = \sqrt{\frac{4\cdot 5}{6\cdot 15}} = 0{,}47$$

La taula següent conté els càlculs necessaris:

Taula 11. Resultats del contrast d'hipòtesis segons el test de Holm

	R_3-R_j	Z	Valor p	i	α / (K-i)	H0
M3-M1	1,60 - 3,20 = -1,60	-3,40	0,0006738	1	0,017	Rebuig H0
M3-M2	1,60 - 2,27 = -0,67	-1,42	0,1556077	3	0,050	No es rebutja H0
M3-M4	1,60 - 2,93 = -1,33	-2,82	0,0048024	2	0,025	Rebuig H0

Segons el mètode de Holm, s'inicien els contrastos amb el valor de *p* més significatiu: p_1. Es compara $p_1 < \alpha / 3$ que correspon a 0,00067 < 0,017. Es rebutja la hipòtesi nul·la que M3 i M1 són equivalents. A continuació, es compara $p_2 < \alpha / 2$, que correspon a 0,0048 < 0,025. Es rebutja també la hipòtesi nul·la que M3 i M4 donen resultats equivalents. Finalment, es compara M3 amb M2: $p_3 <^? \alpha$ amb els valors corresponents: $0{,}1556 <^? 0{,}05$. Com que la condició no es compleix, no es pot rebutjar la hipòtesi nul·la que els mètodes M3 i M2 són equivalents.

El mètode de Hochberg comença el contrast d'hipòtesis amb el valor més gran de *p*. En aquest cas, compara $p_3 <^? \alpha$ que no es compleix $(0{,}155 > 0{,}05)$. No es rebutja la hipòtesi nul·la. A continuació, es contrasta el valor de p_2 amb

$\alpha / 2$. Com que ara es compleix la condició ($0{,}0048 < 0{,}025$), es rebutja la hipòtesi i, per tant, es rebutgen les altres hipòtesis, i es conclou que M3 és significativament diferent a M1 i M4.

Els resultats de la comparació múltiple segons Nemenyi han donat els mateixos resultats que amb el contrast fent servir les proves de Bonferroni-Dunn, Holm i Hochberg en relació amb el model M3. No necessàriament és així en general. Les proves que impliquen comparacions múltiples són més conservadores que les que comparen un model en relació amb els altres. En aquest exemple, el contrast segons Bonferroni-Dunn, Holm i Hochberg no ha revelat noves diferències significatives, la qual cosa reflecteix una diferència significativa entre M3 i M4, i entre M3 i M1, mentre que M3 és bastant similar a M2 en els resultats obtinguts.

4.2.6. Reflexió sobre la comparació múltiple de classificadors

En aquest apartat s'han tractat les proves més habituals en la comparació de resultats de models d'anàlisi de dades. No obstant això, alguns autors van més enllà amb la proposta de proves més adequades. Podeu consultar els treballs de Salzberg (1997) i García i Herrera (2008) per a més detalls.

Resum

En aquest mòdul s'ha tractat el disseny experimental dels mètodes d'analítica de dades. El primer pas consisteix a fer ús d'una metodologia d'analítica de dades que especifiqui els passos que cal seguir en funció de l'objectiu de l'anàlisi. Aquest objectiu fixarà així mateix el tipus de model que cal extreure de les dades (els més comuns són els models de classificació, regressió, associació i agrupació), la qual cosa determinarà l'elecció d'un conjunt d'algoritmes d'analítica possibles. S'han revisat les principals mètriques d'avaluació de la qualitat del model, com la precisió en models de classificació o l'arrel de l'error quadràtic mitjà per a problemes de regressió. L'avaluació del model consisteix a estimar la precisió (o una altra mètrica adequada) real del model, és a dir, la precisió del model quan s'aplica a la població d'interès. Habitualment, l'estimació es fa sobre una mostra petita d'aquesta població i, per això, són necessaris els mètodes de remostreig de la mostra de dades original.

Com que és ben sabut que no hi ha un algoritme òptim per a tot tipus de problemes ni una configuració òptima per defecte d'un algoritme, cal estimar el millor model per al problema determinat. Per a això, s'ha aprofundit en l'estimació de la qualitat del model i la comparació de models. S'han distingit diversos escenaris, segons si es comparen dos models o múltiples, sobre un mateix problema o en múltiples problemes, i mitjançant proves paramètriques o no paramètriques.

El disseny experimental aplicat a l'obtenció de resultats de models d'anàlisi de dades és anàleg al disseny experimental que es pot aplicar a altres àmbits. Així, la comparació entre els resultats de dos models pot equiparar-se a la resposta fisiològica de dos tractaments mèdics o al rendiment d'un grup d'estudiants abans i després de l'ensenyament d'un nou mètode d'estudi. La complexitat de l'avaluació dels resultats dels models d'anàlisi de dades resideix en la dificultat d'estimar la precisió del model, i la minimització del biaix i variància mitjançant mètodes de remostreig. A part d'això, els mètodes de validació estadística tractats en aquest mòdul es poden aplicar a qualsevol altre àmbit que compleixi les mateixes condicions.

Bibliografia

Cohen, J. (1960). «A coefficient of agreement for nominal scales». A: *Educational and Psychological Measurement* (núm. 20, pàg. 37–46).

Davies, D. L.; Bouldin, D. W. (1979). «A cluster separation measure». A: *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1* (núm. 2, pàg. 224–227). Disponible a: <10.1109/TPAMI.1979.4766909>.

Demsar, J. (2006). «Statistical Comparisons of Classifiers over Multiple Data Sets». A: *Journal of Machine Learning Research* (núm. 7, pàg. 1-30).

Dietterich, T. G. (1998). «Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms». A: *Neural Computation* (vol. 10, núm. 7, pàg. 1895-1923).

García, S.; Herrera, F. (2008). «An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons». A: *Journal of Machine Learning Research* (núm. 9, pàg. 2677-2694).

Jain, P.; Sharma, P. (2015). *Behind Every Good Decision: How Anyone Can Use Business Analytics to Turn Data into Profitable Insight*. Nova York: Amacom.

Kuzon, W.; Urbancheck, M.; McCabe, J. (1996). «The Seven Deadly Sins of Statistical Analysis». A: *Annals of Plastic Surgery* (pàg. 265-272).

Lim, T.; Loh, W.; Shih, Y. (2000). «A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms». A: *Machine Learning* (núm. 40, pàg. 203–229).

Lichman, M. (2013). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. Disponible a: <http://archive.ics.uci.edu/m1>.

Salzberg, S. L. (1997). «On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach». A: *Data Mining and Knowledge Discovery* (núm. 1, pàg. 317–327).

Stapor, K. (2017). «Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations». A: Kurzynski, M.; Wozniak, M.; Burduk, R. (eds.). «Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017». A: *Advances in Intelligent Systems and Computing* (núm. 578). Cham (Suïssa): Springer.

Vanwinckelen, G.; Blockeel, H. (2012). «On Estimating Model Accuracy with Repeated Cross-Validation». A: *21st Belgian-Dutch Conference on Machine Learning* (pàg. 39-44).

Witten, I. H.; Frank, E.; Hall, M. A. (2011). *Data Mining. Practical Machine Learning Tools and Techniques* (3a edició). Burlington: Morgan Kaufmann.

Zumel, N.; Mount, J. (2014). *Practical Data Science with R*. Shelter Island (Nova York): Manning.