MODELS DE REGRESSIÓ

Actualment vivim envoltats d'una gran quantitat de dades, computadores potents i intel·ligència artificial. Això només és el començament. La ciència de dades i l'aprenentatge automàtic impulsen el reconeixement d'imatges, el desenvolupament de vehicles autònoms, decisions en els sectors financer i energètic, avenços en medicina, augment de les xarxes socials i molt més. Els models de regressió juguen un paper molt important en aquest camp. Predir el comportament d'un sistema ens permetrà prendre decisions més encertades o, fins i tot, dissenyar nous productes que tinguin èxit en els mercats.

Un model predictiu de regressió es fa servir, principalment, per predir la resposta o resultat d'una variable, segons els canvis que assoleixen unes variables que estan relacionades amb aquesta. Posem pel cas que volem estimar el creixement de les vendes d'una empresa en funció de les condicions econòmiques actuals. Tenim les dades recents de l'empresa que indiquen que el creixement de les vendes és al voltant de dues vegades i mitja el creixement de l'economia. Amb aquesta perspectiva, podem predir les vendes futures de l'empresa, a partir d'informació actual i anterior.

La regressió lineal és la tècnica de regressió més senzilla i la que ofereix la interpretació de resultats més senzilla. Anem a descobrir els seus fonaments matemàtics.

Quan implementem una regressió lineal d'alguna variable dependent **y** del conjunt de variables independents $\mathbf{x} = (x_1, ..., x_r)$, on r és el nombre de predictors, assumim una relació lineal entre **y** i **x**: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. Aquesta equació és la de regressió. $\beta_0, \beta_1, ..., \beta_r$ són els coeficients de regressió i ε és l'error aleatori.

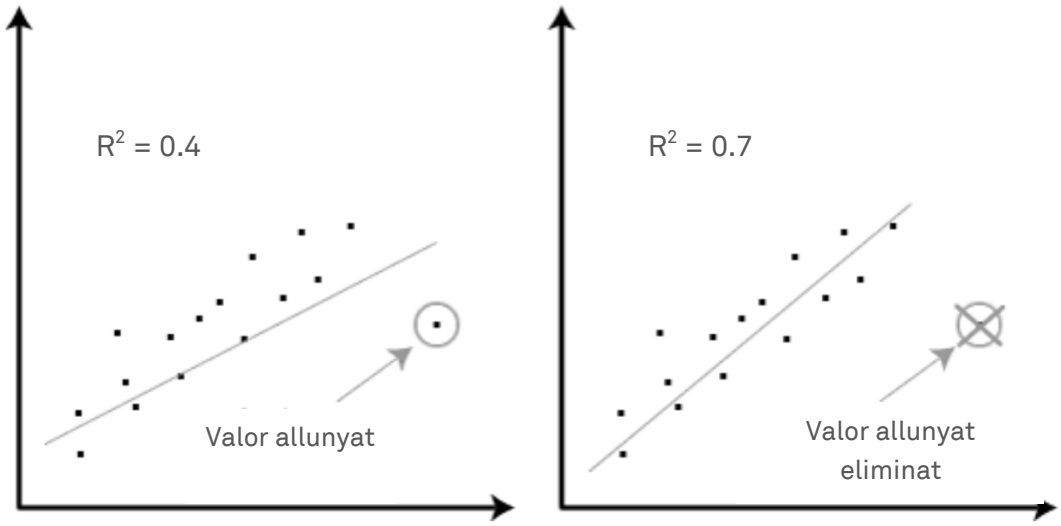
Y es correspondria a l'atribut del qual volem preveure el resultat. Els valors d'**x** corresponen amb els atributs dels quals se suposa una certa relació amb **y**.

La variació de les respostes reals $\mathbf{y}_i$, $i = 1, ..., n$, es produeix, en part, a causa de la dependència de l'atribut $\mathbf{x}_i$. Tot i això, també hi ha una diferència inherent a la sortida. El coeficient de determinació, conegut com a $\boldsymbol{R^2}$, indica quina quantitat de variació en **y** es pot explicar per a la dependència d'**x** mitjançant el model de regressió. El major R^2 indica un millor ajustament i significa que el model pot explicar millor la variació de la sortida amb diferents entrades. El valor $R^2 = 1$ s'adapta perfectament, ja que els valors de les respostes previstes i reals s'ajusten i coincideixen perfectament.

La regressió lineal més simple és aquella en què la variable **y** només depèn d'una altra variable **x**. En els casos pràctics, en ciència de dades, aquest cas no s'acostuma a donar. Els casos reals són més complexos i normalment tenen més d'una variable **x** relacionada amb un atribut **y**. En aquests casos reals, estarem parlant, doncs, de regressió lineal múltiple.

Si només hi ha dues variables independents, la funció de regressió estimada és $f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2$. L'objectiu de la regressió és determinar els valors dels coeficients b_0, b_1 i b_2, de tal manera que els resultats estiguin el més a prop possible de les respostes reals.

Quan generem els models, hem de veure si aquests poden tenir un coeficient de correlació massa baix o massa alt (molt a prop d'1). En el primer cas, estarem davant d'un model que genera uns resultats massa allunyats dels reals. Estarem davant d'una situació d'*underfitting.* En el segon cas, estarem davant d'un sistema d'*overfitting*. Això voldrà dir que el model encaixa perfectament amb les dades donades, però corre el risc de preveure molt malament quan s'aplica a altres dades amb les quals no ha estat entrenat. Com que l'error és la distància quadrada entre el punt de dades i la línia de regressió, les grans distàncies tenen errors desproporcionadament grans, i fan que l'anàlisi de regressió esdevingui una solució amb un coeficient de correlació baix. Com a tal, aquells valors que estiguin significativament allunyats de la recta, haurien de ser eliminats del conjunt de dades. Aquests valors són aquells que es podrien descobrir durant l'exploració i depuració de dades.



En les dues gràfiques que es mostren just a dalt, es veu com una vegada hem eliminat el valor allunyat, el nostre model predictiu ha millorat considerablement. Per facilitar-nos la interpretació dels resultats, en ciència de dades és molt útil comptar amb gràfiques que mostrin les dades de manera visual.

Descobreix tot el que Barcelona Activa pot fer per a tu

Acompanyament durant tot el procés de recerca de feina

barcelonactiva.cat/treball

Suport per posar en marxa la teva idea de negoci

barcelonactiva.cat/emprenedoria

Serveis a les empreses i iniciatives socioempresarials

barcelonactiva.cat/empreses

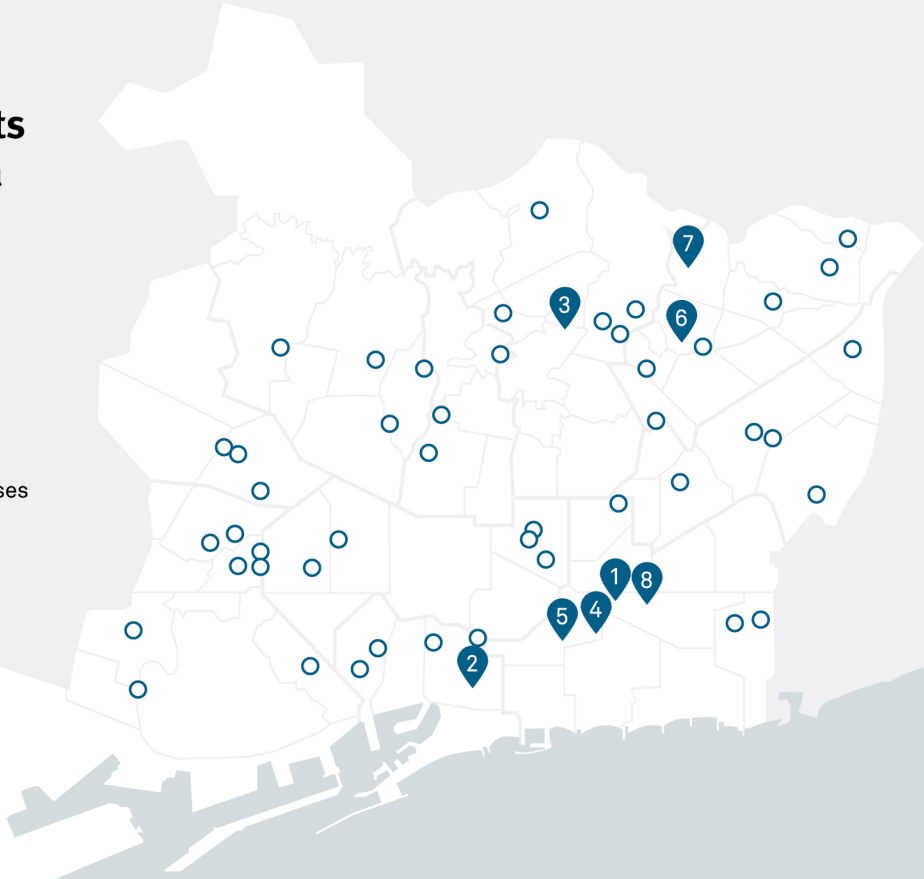
Formació tecnològica i gratuïta per a la ciutadania

barcelonactiva.cat/cibernarium

Xarxa d'equipaments de Barcelona Activa

1. Seu Central Barcelona Activa
   Porta 22
   Centre per a la Iniciativa Emprenedora Glòries
   Incubadora Glòries
2. Convent de Sant Agustí
3. Ca n'Andalet
4. Oficina d'Atenció a les Empreses
   Cibernàrium
   Incubadora MediaTIC
5. Incubadora Almogàvers
6. Parc Tecnològic
7. Nou Barris Activa
8. innoBA

○ Punts d'atenció a la ciutat



Barcelona Activa

Ajuntament de Barcelona

Darrera actualització 2019

Cofinançat per:

UNIÓ EUROPEA
Fons Europeu de Desenvolupament Regional

Segueix-nos a les xarxes socials:

barcelonactiva.cat/cibernarium
barcelonactiva
barcelonactiva
company/barcelona-activa