Pràctica 1: Regressió

Red wine quality dataset

Grup 105

Laia Rubio – NIU:

Erik Villarreal – NIU:1599119

Raúl Villar – NIU:1596830

Index

# Introducció

L’objectiu de la pràctica és analitzar una base de dades sobre la qualitat del vi vermell amb l’objectiu de poder aplicar-li models de regressió. Per a aconseguir-ho, al llarg de la memòria s’explicarà el procés aconseguit i es resoldran les preguntes de l’enunciat en forma de apartats i subapartats.

Les preguntes a resoldre son:

* Apartat C:
  + Quin és el tipus de cada atribut?
  + Quins atributs tenen una distribució Gaussiana?
  + Quin és l'atribut objectiu? Per què?
* Apartat B:
  + Quin són els atributs més importants per fer una bona predicció?
  + Amb quin atribut s'assoleix un MSE menor?
  + Quina correlació hi ha entre els atributs de la vostra base de dades?
  + Com influeix la normalització en la regressió?
  + Com millora la regressió quan es filtren aquells atributs de les mostres que no contenen informació?
  + Si s'aplica un PCA, a quants components es redueix l'espai? Per què?
* Apartat A:
  + Com influeixen tots els paràmetres en el procés de descens? Quins valors de learning rate convergeixen més ràpid a la solució òptima? Com influeix la inicialització del model en el resultat final?
  + Quines funcions polinomials (de diferent grau, de diferents combinacions d'atributs, ...) heu escollit per ser apreses amb el vostre descens del gradient? quina ha donat el millor resultat (en error i rapidesa en convergència)?
  + Utilitzeu el regularitzador en la fòrmula de funció de cost i descens del gradient i proveu polinomis de diferent grau. Com afecta el valor del regularitzador?
  + Quina diferència (quantitativa i qualitativa) hi ha entre el vostre regressor i el de la llibreria ?
  + Té sentit el model (polinomial) trobat quan es visualitza sobre les dades?
  + Ajuda la visualització a identificar aquelles mostres per a les que el regressor obté els pitjors resultats de predicció?

La base de dades sobre la que operarem tracta sobre unes 1600 mostres aproximadament, i aporta informació sobre la composició de cada vi tractat, i la seva qualificació.

# Plantejament de dades

## Inicialitzar, visualitzar i preparar les dades

La base de dades bé donada per l’arxiu “winequality-red.csv” i ens aporta un total de 12 variables i 1599 mostres. Imprimint les primeres 5 mostres, i amb la funció ‘describe()’ podem veure i aprendre tota la informació que ens donarà la base de dades al llarg de la pràctica.

|  |
| --- |
|  |
| *Taula de les primeres 5 mostres de la BD i els seus respectius valors* |

|  |
| --- |
|  |
| *Taula obtinguda amb la funció describe() de la llibreria pandas* |

Amb aquestes mostres podem crear una taula d’informació sobre las variables com aquesta:

|  |  |  |  |
| --- | --- | --- | --- |
| Nom Variable | Tipus de dada | Rang | Tipus de variable |
| fixed acidity | Float64 | (4,600 – 15,900) | Continua |
| volatile acidity | Float64 | (0,120 – 1,580) | Continua |
| citric acid | Float64 | (0 – 1) | Continua |
| residual sugar | Float64 | (0,900 – 15,500) | Continua |
| chlorides | Float64 | (0,012 – 0,611) | Continua |
| free sulfur dioxide | Int64 | (1 – 72) | Discreta |
| total sulfur dioxide | Int64 | (6 – 289) | Discreta |
| density | Float64 | (0,990 – 1,004) | Continua |
| pH | Float64 | (2,740 – 4,010) | Continua |
| sulphates | Float64 | (0,330 – 2,000) | Continua |
| alcohol | Float64 | (8,400 – 14,900) | Continua |
| quality | Int64 | (3 – 8) | Discreta |

Com no tenim ninguna variable categòrica, no haurem de fer ningun tractament específic a alguna de les variables. Abans de continuar amb l’anàlisi, ens hem assegurat de que totes les variables contenen valors i no hi ha cap mostra que pugui tenir NULLS en algun apartat de la taula.

Una vegada ja sabem de que tracta la nostra base de dades, i ja no pot tenir valors que puguin donar resultats il·lògics, ja podem començar a analitzar els atributs que la composen i la seva relació entre ells.

## Comprendre els atributs

Per a poder veure com es comporten cada atribut i com es relacionen entre ells, la millor opció és utilitzar la funció ‘plairplot()’ de la llibreria seaborn, per a crear gràfiques entre variables.

|  |
| --- |
|  |
| *Gràfica generada amb la funció ‘plairplot()’ de la llibreria seaborn* |

Aquesta gràfica ens dona informació molt important per a escollir quins atributs necessitarem per a fer la regressió lineal, però de moment ens centrarem en els histogrames generats en la diagonal de l’imatge. Aquests histogrames ens ajudaran a veure les distribucions de cada variable, i poder veure quines tenen distribucions Gaussianes. De entre totes, hem intuït que només els atributs ‘density’ i ‘pH’ tenen aquesta distribució.

Per a comprovar aquesta hipòtesi hem utilitzat la funció ‘normaltest()’ de la llibreria SciPy, la qual ens pot dir de manera estadística si de veritat son distribucions Gaussianes o no. Una vegada implementada la funció, MUCHO TEXTO QUE HAY QUE ESCRIBIR AQUIhUODGPUODWOPUHODWHODHWOPDHWOPIHDOPWHDOWHDHWOIHDOIWHDOIWHDOIDHWOIDHWOHDOWHDOWHDOHWODHWOUHDWOUHDU

A més de les gràfiques, per a poder comprendre del tot la relació entre cada atribut, hem creat dues matrius de correlació. Una, la ja explicada a classe amb els valors per defecte dels atributs, i una altre, amb tot valors absoluts per a veure relacions inversament proporcionals.

|  |
| --- |
|  |
| *Matriu de correlació generada amb la funció ‘corr()’ de la llibreria pandes i ‘heatmap()’ de la llibreria seaborn* |

|  |
| --- |
|  |
| *Matriu de correlació amb valors absoluts generada amb la funció ‘corr()’ de la llibreria pandes i ‘heatmap()’ de la llibreria seaborn* |

# Resultats

# Conclusions