Aprenentatge Automàtic i Mineria de Dades (AAMD)
GEI (2024-25)

Pràctica 1: Aprenentatge supervisat

Objectiu

Ser capaç de fer prediccions i classificacions sobre un conjunt de dades reals, seguint tots els passos necessaris per a una correcta aplicació de la teoria.

Descripció

Donades unes dades reals, caldrà fer el seu pre-processament, aplicar diferents tècniques de predicció i classificació, i utilitzar la cross-validation per a ajustar els paràmetres i avaluar la capacitat de predicció de cadascuna de les tècniques aplicades.

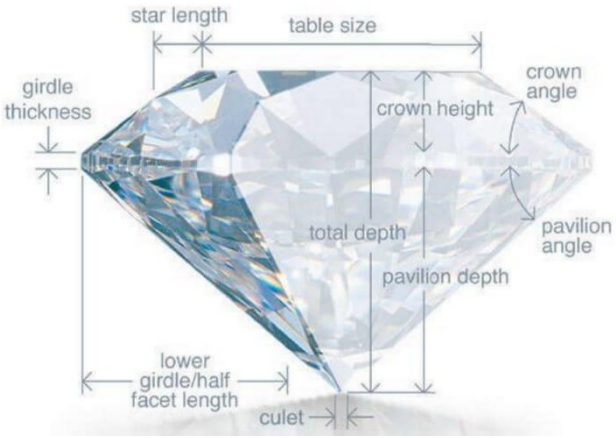
Dades

Utilitzareu dades sobre les característiques físiques i el preu d'uns 54000 diamants. S'han dividit en dos arxius:

- `diamonds-train.csv`: conté dades de 44000 diamants, que s'utilitzaran per a preparar el sistema de predicció.
- `diamonds-test.csv`: conté dades de 9940 diamants, que només s'utilitzaran per a avaluar la qualitat de les prediccions.

Els atributs que tenim de les dades i el seus rangs de variació són els següents:

- *id*: identificador del diamant (1 – 53940)
- *price*: preu en dòlars USA ($326 – $18,823)
- *carat*: pes del diamant (0.2 – 5.01)
- *cut*: qualitat del tall, entre Fair (pitjor), Good, Very Good, Premium, Ideal (millor)
- *color*: color del diamant, entre J (pitjor) i D (millor)
- *clarity*: mesura de la claredat del diamant, entre I1 (pitjor), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (millor)
- *x*: longitud en mm (0 – 10.74)
- *y*: amplada en mm (0 – 58.9)
- *z*: profunditat en mm (0 – 31.8)
- *depth*: percentatge de profunditat total = z / mean(x, y) = 2 * z / (x + y) (43 – 79)
- *table*: amplada del cim del diamant relatiu a la màxima amplada (43 – 95)



Desenvolupament i lliurament

Aquesta pràctica es fa en grups de dos.

Heu de lliurar un únic "Jupyter notebook" que contingui tot el que es demana a continuació, estructurat en seccions i incloent totes les explicacions que siguin necessàries per a entendre el que s'està fent. El codi ha d'estar correctament comentat (evitar comentaris que siguin "obvis"). Podeu utilitzar Python, Julia o R, com preferiu, i us heu d'assegurar que funcionen correctament en els contenidors Docker que vam explicar a classe. El nom del notebook ha de ser del tipus:

- `A1-Group_X-Nom1_Cognom1-Nom2_Cognom2.ipynb`

Treball a fer

1. Lectura de les dades.

2. Pre-processament de les dades:
    a. Identificar els patrons amb dades "no físiques".
    b. Identificar outliers a partir de la distribució de cada atribut.
    c. Identificar outliers a partir de scatter plots atribut/preu.
    d. Fer una taula amb els identificadors dels outliers, els seus atributs, i les raons per la qual cadascú es considera outlier.
    e. Decidir què fer amb els outliers i aplicar-ho als conjunts de dades.
    f. Convertir els atributs categòrics en numèrics. Observeu que, per la seva semàntica, els valors de cada atribut categòric estan ordenats, i per tant una representació numèrica té més sentit que un one-hot encoding.
    g. Estandaritzar tots els atributs per separat, excepte x, y, z i price.
    h. Escala x, y i z entre -1 i 1, però de forma conjunta, no cada atribut per separat. És dir, fer les correspondències min(x, y, z) → -1, max(x, y, z) → 1.
    i. Com el preu té un rang de variació que s'estén tres ordres de magnitud, en lloc de treballar amb el preu és millor fer-ho amb el seu logaritme. Per tant, calculeu el logaritme dels preus, guardeu-ho en un nou atribut *log_price*, i apliqueu un escalament lineal entre 0.1 i 0.9.
    j. Construir un nou atribut *high_price* que valgui 0 si el preu està estrictament per sota dels $2,500 i 1 en cas contrari.

    Observació: l'atribut id serveix només per a identificar els diamants, no s'ha d'utilitzar com a atribut d'entrenament ni de test.

3. Predicció del preu:
    a. Es vol predir el preu dels diamants (de fet, el logaritme del preu escalat).
    b. S'han d'utilitzar els següents mètodes de predicció:
        i. Regressió multilineal (MLR).
        ii. K-Nearest Neighbors (k-NN).
        iii. Multilayer Neural Network amb Back-Propagation (MLNN-BP).
    c. Utilitzeu cross-validation per a ajustar els paràmetres dels models, i un cop decidits els paràmetres, mesurar la qualitat prevista de la predicció. Tot això es fa sobre el conjunt d'entrenament proporcionat, que es dividirà en diferents subconjunts d'entrenament i validació en el procés de cross-validation.

d. Un cop decidits els paràmetres i avaluada la qualitat prevista de la predicció, utilitzeu tot el conjunt d'entrenament per a ajustar el model predictiu definitiu que s'aplicarà al conjunt de test.
e. La mesura de qualitat de la predicció que s'ha d'utilitzar és el Mean Absolute Percentage Error (MAPE), que es calcula així:

$$\mathrm{MAPE}(\boldsymbol{y}, \boldsymbol{z}) = 100 \times \frac{1}{p} \sum_{\mu=1}^{p} \left| \frac{y^{\mu} - z^{\mu}}{z^{\mu}} \right|$$

on y^{μ} és la predicció i z^{μ} el valor real corresponents al patró μ-èssim. Aquests valors han de representar preus en USD ($), no els seus transformats utilitzats durant l'entrenament.
f. Fer prediccions dels preus (en USD) sobre el conjunt de test per a cada model de predicció i analitzeu els resultats:
    i. Calculeu el MAPE en el conjunt de test per a cada model.
    ii. Genereu un dataframe amb els atributs originals més les prediccions dels tres models, amb noms *pred_price_mlr*, *pred_price_knn*, *pred_price_bp*.
    iii. Feu scatter plots entre el valor real del preu i la seva predicció per a cada model.
    iv. Feu scatter plots entre el valor real del preu i la seva predicció per a cada model.
g. Discutiu i elaboreu conclusions a partir dels resultats obtinguts.

4. Classificació del preu:
    a. Es vol classificar els patrons en funció de si tenen un preu alt o no, segons la variable *high_price*.
    b. S'utilitzaran els següents mètodes de classificació:
        i. Logistic Regression (LR).
        ii. Support Vector Machines (SVM).
        iii. Multilayer Neural Network amb Back-Propagation (MLNN-BP).
    c. Utilitzeu cross-validation per a ajustar els paràmetres dels models, i un cop decidits els paràmetres, mesurar la qualitat prevista de la classificació. Tot això es fa sobre el conjunt d'entrenament proporcionat, que es dividirà en diferents subconjunts d'entrenament i validació en el procés de cross-validation.
    d. Un cop decidits els paràmetres i avaluada la qualitat prevista de la classificació, utilitzeu tot el conjunt d'entrenament per a ajustar el model classificador definitiu que s'aplicarà al conjunt de test.
    e. Les mesures de qualitat de la classificació que es demanen són: accuracy, sensitivity i specificity.
    f. Pels models LR i MLNN-BP el resultat és una estimació de la probabilitat de pertànyer a la classe *high_price*, mentre que per SVM és directament la classe (0 o 1). En els dos primers cassos, es pot seleccionar la classe de major probabilitat.

g. Aplicar els tres mètodes sobre el conjunt de test i analitzeu els resultats:
    i. Calculeu les taules de contingència per a les prediccions de cada model.
    ii. Calculeu accuracy, sensitivity i specificity en el conjunt de test per a cada model.
    iii. Genereu un dataframe amb els atributs originals més les classificacions dels tres models, amb noms *class_high_price_lr*, *class_high_price_svm*, *class_high_price_bp*. Afegiu també les probabilitats de pertànyer a la classe *high_price* en els atributs *prob_high_price_lr* i *prob_high_price_bp*.
    iv. Pels models LR i MLNN-BP, dibuixeu la ROC curve i calculeu l'AUC.
h. Discutiu i elaboreu conclusions a partir dels resultats obtinguts.