Enunciat

Objectiu

Aplicar distints classificadors (regressió logística, perceptró, i random forest) a un problema real. Analitzar les característiques del problema, estudiar i ajustar els paràmetres dels classificadors.

Materials

El problema es realitzarà amb Python utilitzant l'entorn explicat a classe. L'entrega serà un PDF resultant d'un o varis Python Notebooks.

package	version
python	3.9.7
numpy	1.21.2
matplotlib	3.4.3
pandas	1.3.4
scikit-learn	1.0.1
jupiterlab	3.2.1

Puntuació

El problema s'organitzarà en tres nivells de dificultat: A, (sobre 10), dificultat alta; B, (sobre 8), dificultat mitja i C, (sobre 6), dificultat baixa. Para aprovar es un requisit necessari completar satisfactòriament la part de dificultat (C), demostrant així una comprensió fonamental de la matèria.

Entrega

Dues dates límit d'entrega (deadlines).

- 21 de gener del 2022 – Avaluació ordinària – Nota màxima 10
- 16 de febrer del 2022 – Avaluació extraordinària – Nota màxima 7

El problema

Predicció de la qualitat del vi utilitzant distints algoritmes de classificació.

Per a dur a terme aquesta tasca utilitzarem el dataset proporcionat per "UCI Machine Learning Repository" anomenat " Wine Quality Data Set" [7]. Aquest dataset està format per dos subconjunts de dades. Un associat a dades de vi blanc i un altre associat a dades de vi tinto. Ambdós datasets disposen de les següents característiques per a cada mostra:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Finalment, el dataset disposa d'una columna extra amb la variable objectiu (Basada amb dades sensorials):

12. quality (score between 0 and 10)

Amb aquest dataset volem crear un sistema capaç de predir la qualitat d'una nova mostra de vi en base a les 11 característiques d'entrada. Solucionarem el problema de dues maneres:

- La sortida ha de ser un valor entre 0 i 10 (els mateixos valors que a la columna "quality")
- La sortida ha de ser un dels següents tres valors de qualitat: baixa (valor de "quality" menor que 6), mitjana (valor de "quality" igual a 6) i alta (valor de "quality" major que 6).

Preguntes

1. Dificultat C

Realitzar una comparació del rendiment dels següents models: la regressió logística [3], el perceptró[4], i el Random Forest [5]. Realitzar l'aprenentatge amb els valors dels hiperparàmetres "per defecte" de les implementacions de `scikit-learn`.

Heu d'utilitzar els dos datasets de vins (vi blanc i tinto) disponibles a l'Aula Digital en format CSV com un únic dataset. Per a fer-ho haureu de fusionar els dos datasets per a

que siguin una única taula i afegir una columna a les dades indicant el tipus de vi de cada mostra (conegut pel fitxer d'on prové la mostra).

La comparació dels models es realitzarà amb el rendiment de cada model en relació a un conjunt de validació que heu d'escollir (explicar clarament perquè heu escollit aquesta partició del conjunt de dades i com l'heu construït).

Per acabar, heu de donar una explicació dels resultats obtinguts. Teniu en compte, que haureu de preparar bé les dades ja que inicialment estan en format "brut".

Nota: Aquest dataset és àmpliament usat per a la comunitat. Podreu trobar molta informació i casos d'ús a la plataforma Kaggle [2, 1].

2. Dificultat B

La part de dificultat B consisteix en la realització de la part d'enginyeria de característiques. Aquesta tasca consisteix en seleccionar les característiques més adequades per a resoldre el problema o crear-ne de noves a partir de les existents. Explicar i comparar els resultats obtinguts utilitzant totes les característiques (dificultat C) o el subconjunt seleccionat.

3. Dificultat A

Per acabar, la part de dificultat A consisteix a estodiar els paràmetres dels millors models predictius utilitzant la funció GridSearchCV [6]. Extreure una conclusió global de quin és el millor model (i els seus paràmetres).

References

[1] Kaggle competició "wine quality". `https://www.kaggle.com/c/wine-quality/overview`. Accedit el 08-12-2021.

[2] Kaggle dataset "red wine quality". `https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009`. Accedit el 08-12-2021.

[3] Scikit-learn documentació de l'implementació de l'algorisme "logistic regression". `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`. Accedit el 08-12-2021.

[4] Scikit-learn documentació de l'implementació de l'algorisme "perceptrón". `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Perceptron.html`. Accedit el 08-12-2021.

[5] Scikit-learn documentació de l'implementació de l'algorisme "random forest". `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`. Accedit el 08-12-2021.

[6] Scikit-learn documentació de l'implementació del "grid search". `https://scikit-learn.org/stable/modules/grid_search.html`. Accedit el 08-12-2021.

[7] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. Smart Business Networks: Concepts and Empirical Evidence.