Pràctica Bloc III

9-XII-2024

Objectiu

L'objectiu d'aquest exercici avaluable és aplicar els conceptes i tècniques d'aprenentatge automàtic apreses durant el curs.

- Aquesta pràctica s'ha d'entregar no més tard del dia de l'avaluació complementària, **14 de gener de 2025 a les 23:59**.
- El resultat de la pràctica serà un document **Notebook** (`.ipynb`) a on s'ha de combinar tant el codi Python per resoldre el problema com les respectives explicacions en Markdown. També s'ha d'entregar una versió PDF d'aquest.
- La pràctica s'ha de realitzar en parella. Les parelles no tenen per què ser les mateixes que a les pràctiques anteriors i els dos membres poden ser de grups diferents, tant de teoria com de pràctica.
- Únicament es corregiran les pràctiques dels alumnes que hagin dut terme una entrevista personalitzada.
- L'entrevista es realitzarà posteriorment a la data d'entrega de la pràctica.

Materials

La pràctica es realitzarà amb Python emprant l'entorn creat per la realització d'aquesta i disponible al repositori de l'assignatura [1]. Per assegurar la reproductibilitat i evitar incompatibilitats entre llibreries, és fonamental utilitzar les versions especificades al fitxer `requirements.txt` del repositori. El conjunt de dades que emprarem és el Forest Cover Type Dataset.

[1] `https://github.com/miquelmn/ia_2024`

Descripció del problema

En aquesta pràctica, aplicarem les diferents tècniques estudiades durant la tercera part de l'assignatura al conjunt de dades *Forest Cover Type Dataset*[2].

Per llegir el conjunt de dades es pot fer ús de la llibreria `Pandas`. Podeu emprar el codi següent:

```
import pandas as pd

df = pd.read_csv("covtype.csv")
```

Aquest conjunt de dades tabulars recull informació detallada sobre les característiques del sòl i la vegetació del Bosc Nacional de Roosevelt, situat als Estats Units (vegeu la Figura 1). Consta de 581013 mostres, cadascuna descrita per 54 atributs que inclouen característiques com l'altitud, la inclinació del terreny, el tipus de sòl i el tipus d'ombra, entre d'altres.

Figura 1: Fotografia del Bosc Nacional de Roosevelt.

L'objectiu principal d'aquesta pràctica serà desenvolupar models d'aprenentatge automàtic que puguin predir amb precisió el tipus de coberta forestal associat a cada mostra. Aquest és un problema de classificació multiclasse amb 7 categories diferents, que representen tipus específics de vegetació.

Per assolir aquest objectiu, es farà ús de tècniques com la selecció de característiques, l'entrenament i validació de models, i l'anàlisi dels resultats obtinguts. A més, es compararan diversos algorismes de classificació.

Treball a fer

La tasca a realitzar en aquesta pràctica consisteix a comparar el rendiment de diversos models d'aprenentatge automàtic utilitzant el conjunt de dades proporcionat. Els models seleccionats per a aquesta comparació són els següents: el Perceptró, la Regressió

[2] `https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset/data`

Logística, les Màquines de Vectors de Suport (SVM per les seves sigles en anglès), els Arbres de Decisió i els Boscos Aleatoris. Tots aquests models estan disponibles a la llibreria `Scikit-learn`[3].

Passos a seguir

La comparació ha d'incloure els següents elements clau:

1. **Anàlisi exploratòria de dades (EDA):** Abans d'entrenar els models, serà imprescindible dur a terme una anàlisi detallada de les dades.

2. **Selecció dels millors hiperparàmetres:** Per tal d'obtenir el millor rendiment possible de cada model, s'hauran d'ajustar els seus hiperparàmetres emprant les tècniques vistes a classe. A més s'ha de justificar l'ús dels hiperparàmetres a partir del seu afecte al model.

3. **Ús correcte de les mesures d'avaluació:** Es valorarà l'aplicació adequada de mètriques per avaluar el rendiment dels models. A més, s'ha de justificar la selecció de cada mètrica segons el problema que es vol resoldre.

4. **Discussió crítica dels resultats:** Els resultats obtinguts per cada model han de ser analitzats i comparats de manera detallada. Això inclou:

    - Explicar per què certs models poden tenir un rendiment superior o inferior en aquest conjunt de dades.
    - Identificar possibles fonts de biaix o errors en les prediccions.
    - Utilització de taules per dur a terme de manera simple i directe la comparativa.
    - Ús de gràfics i tècniques de visualització per comparar els resultats i millorar l'anàlisi elaborada.

Aspectes a considerar

Les decisions preses al llarg de tot el procés han d'estar correctament justificades en el document de lliurament. Per exemple, si s'opta per una tècnica de selecció d'hiperparàmetres o si es decideix no incloure determinades variables en el model final, aquestes decisions han de ser clarament explicades.

Es valorarà positivament l'ús de tècniques de reducció de dimensionalitat. A les sessions pràctiques, hem explorat una tècnica senzilla basada en el Coeficient de Pearson per identificar variables correlacionades. Tot i això, si es fa ús de tècniques més avançades, que no s'han vist a classe, sempre que estiguin ben explicades i justificades es pot aconseguir puntuació extra. Un exemple seria l'ús de l'Anàlisi de Components Principals (PCA per les seves sigles en anglès)[4], que permet reduir la dimensionalitat tot mantenint la màxima variabilitat de les dades.

[3] `https://scikit-learn.org/stable/`

[4] `https://scikit-learn.org/1.5/modules/generated/sklearn.decomposition.PCA.html`

Finalment, el que es valora no és el resultat obtingut (la mètrica obtinguda) sinó el procés realitzat.

En resum, aquesta pràctica no només pretén ser una comparació del rendiment tècnic dels models, sinó també fomentar una anàlisi reflexiva i crítica que permeti entendre millor com i per què aquests models funcionen en el context del conjunt de dades proporcionat.

Condicions

És condició *sine qua non* per a la correcció de la pràctica que aquesta funcioni a l'ordinador del professor, l'entrega del Notebook en format `.ipynb` i la seva conversió a `PDF` i la realització d'una entrevista personalitzada.