Învățare Automată

Tema 1 - 2024

1. Descriere generală

În practica de zi cu zi a unui inginer sau cercetător în domeniul învățării automate intră frecvent următoarele trei aspecte:

- Vizualizarea și “explorarea” datelor unei probleme (Exploratory Data Analysis)
- Încercarea de a extrage atribute ale datelor problemei pentru a fi utilizate în obiectivul de analiză ales (e.g. clasificare, regresie, detectie de anomalii)
- Evaluarea mai multor modele pentru găsirea soluției celei mai bune pentru problema dată

Sarcinile voastre de lucru vor solicita utilizarea de biblioteci de **vizualizare a datelor (crearea de diagrame)**, **extragerea de atribute (feature extraction)** pentru folosirea algoritmilor de clasificare discutați la curs, precum și **utilizarea unor modele** simple de machine learning.

2. Descrierea Seturilor de Date

Aveți la dispoziție un set de date adecvat învățării supervizate care conține informații cu pacienți pentru care au fost măsurate 18 atribute. Acești pacienți au fost diagnosticați conform coloanei *Diagnostic*. Scopul este să antrenați modele de învățare care să reușească să prezică acest diagnostic cu performanțe cât mai bune. Găsiți setul de date pe Moodle la Assginment-ul acestei teme.

3. Cerințe

3.1. Explorarea Datelor (Exploratory Data Analysis) [4p]

Primul pas recomandat în rezolvarea unei probleme de clasificare este obținerea unor informații asupra caracteristicilor principale ale problemei. De regulă, foarte folositoare în această etapă este aplicarea unor metode de ***vizualizare a datelor*** și de ***raportare a distribuțiilor de valori*** pe fiecare variabila folosită în predicție.

<u>Analize recomandate</u>

1. **Analiza echilibrului de clase**

Realizați un grafic al frecvenței de apariție a fiecărei etichete (clase) în setul de date de antrenare / test, folosind **bar plot** / **count plot**.
Pentru realizarea unor astfel de bar plots puteți folosi mai multe biblioteci:

- Folosind biblioteca seaborn pentru barplot sau countplot
- Direct dintr-un DataFrame Pandas folosind pandas.DataFrame.plot.bar

2. Vizualizarea datelor

Analizați atributele cu care lucrați în funcție de tipul lor.

Sugestii de valori statistice de extras pentru atribute numerice:

- Medie
- Abaterea standard
- Abaterea medie absolută
- Valoare minimă
- Valoare maximă
- Diferența de valori maxime și minime
- Mediană
- Abaterea mediană absolută
- Intervalul intercuartil

Sugestii de valori statistice de extras pentru atribute discrete / nominal / ordinale:

- Valori unice
- Histogramă

Realizați analize de covarianță, atât între atribute, cât și între atribute și clasă.

Notă: Diagramele din această secțiune sunt cele ***minimal cerute*: NU** sunt singurele pe care le puteți face :-)

3.2. Extragerea manuală a atributelor și utilizarea algoritmilor clasici de Învățare Automată [6p]

Pentru analiza setului de date veți folosi următorii algoritmi:

- RandomForest - folosiți implementarea din scikit-learn
- ExtraTrees - folosiți implementarea din scikit-learn
- GradientBoosted Trees - folosiți implementarea din biblioteca xgboost
- SVM - folosiți implementarea din scikit-learn

Folosiți înțelegerea datelor câștigată la pasul 3.1 pentru a determina dacă este necesară **standardizarea datelor.** Acest pas este unul des întâlnit în etapa de pre-procesare a datelor înainte de antrenarea unui clasificator, în vederea uniformizării valorilor numerice aferente fiecărui tip de atribut (e.g. nu este dorit ca unele atribute sa aibă valori de ordinul miilor, iar altele de ordinul unităților).

Tratați valorile lipsă pentru atributul *Weight* (marcate cu valoarea -1) prin diverse metode (vezi https://scikit-learn.org/stable/modules/impute.html pentru metode precum ***SimpleImputer*** și ***IterativeImputer***).

Frecvent se întâmplă ca nu toate atributele să aibă o contribuție importantă în cadrul predicției. Ca atare, investigați aplicarea tehnicilor de **selectare a atributelor (eng. Feature**

selection) oferite în scikit-learn. Folosiți cel puțin una din metodele ***Variance Threshold*** sau ***Select Percentile***. Explicați (măcar intuitiv) diferențele dintre cele 2 seturi (cel inițial și cel redus prin selecție).

Fiecare algoritm din cei propuși are o serie de **hiper-parametri** care influențează funcționarea acestuia. Pentru a găsi valorile potrivite pentru aceștia veți folosi o procedură de **căutare a hiper-parametrilor** pe bază **de Grid Search cu Cross Validation.**
Setul minim de hiper-parametri de căutat este:

- SVM: tipul de kernel, parametru *C* de regularizare
- RandomForest: numărul de arbori, adâncimea maximă a unui arbore, procentul din input folosit la antrenarea fiecărui arbore
- ExtraTrees: numărul de arbori, adâncimea maximă a unui arbore, procentul din input folosit la antrenarea fiecărui arbore
- GradientBoostedTrees: numărul de arbori, adâncimea maximă a unui arbore, learning rate

Evaluarea algoritmilor
În raportul vostru trebuie să prezentați următoarele:

- Rezultatul procedurii de feature selection: numărul total de feature-uri considerate și numărul total de feature-uri utilizate la antrenare (ca urmare a procedurii de feature selection). În cazul în care acestea diferă, explicați (intuitiv) de ce, cu referire concretă la coloanele în cauză.
- Pentru fiecare algoritm, realizați un tabel în care să prezentați **media și varianța** pentru ***acuratețea generală de clasificare***, ***precizie / recall / F1*** **la nivelul fiecărei clase în parte**
    - Pe linii va fi indexată configurația de hiper-parametri rezultată din procedura de GridSearch.
    - Pe coloane vor fi prezentate metricile cerute
    - **Relevați prin bolduire** valorile maxime pentru fiecare metrică
- Pentru **cea mai bună variantă a hiper-parametrilor,** pentru **fiecare algoritm**, realizați o **matrice de confuzie** peste clase.

4. Predarea temei

Tema va fi încărcată pe Moodle însoțită de un raport sub formă de fișier PDF, care include:

- **Cerința 3.1** - cuprinde toate vizualizările și statisticile cerute. **Este obligatorie** prezența în text a **unei interpretări / analize** a diagramelor rezultate.
- **Cerința 3.2** - include raportarea extragerii de atribute și a evaluării algoritmilor de clasificare pentru cele două tipuri de seturi de date propuse. **Este obligatorie** prezența în text a **unei interpretări / analize** a rezultatelor obținute (e.g. care atribute sunt cele mai predictive, cât de puternic este impactul hiper-parametrilor asupra performanței fiecărui algoritm considerat, care sunt clasele cu cele mai bune predicții).

Rezultatele temei vor fi prezentate în cadrul laboratoarelor de Învățare Automată, **exclusiv pe baza rapoartelor încărcate.**