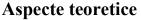
Rezolvarea unor probleme prin metode de învățare automată



Objective

Dezvoltarea sistemelor care învață singure. Probleme de tip clustering din domeniul text-mining rezolvate cu ajutorul algoritmilor de tip k-means. Evaluareaa performanței acestor metode.





Algoritmul k-means. Tehnici de pre-procesare a textelor.

Proiectarea sistemelor care învată singure.

Evaluarea sistemelor care învață singure. Metrici de performanță.



Termen de predare și evaluare

Laborator 12

Punctajele acordate:

- Implementare kMeans pt clusterizare 100 puncte
- Extragere caracteristici din texte
 - o Bag of Words / TF-IDF / Wrd2Vec 50 puncte
 - Alte caracteristici 100 puncte
- · Etichetare emotii
 - o supervizat 50 puncte
 - o nesupervizat 100 puncte
 - o hibrid 100 puncte

Cerinte



Specificați, proiectați și implementați rutine de rezolvare a unei probleme de clusterizare folosind algoritmul k-means.

Exemplu live: Ce fel de mesaje sunt?

Se doreste clusterizarea unor mesaje in doua categorii (spam si ham). Pentru fiecare mesaj se cunoaste textul aferent lui. Să se rezolve problema, implementându-se rutine pentru clusterizare cu k-means a mesajelor.

Proces:

- Se pleaca de la un set de date format din textul mesajelor precum cel din fisierul data/spam.csv
- Se imparte setul de date in date de antrenament si in date de test
- Se extrag anumite caracteristici din textul mesajelor folosind diferite reprezentari precum:
 - o Bag of Words
 - o TF-IDF
 - Word2Vec
- Se aplica algoritmul k-means pe setul de antrenament si se identifica cei doi centroizi (corespunzatori clusterilor spam si ham, respectiv)
- Fiecare mesaj din setul de test se asociaza acelui cluster pentru care distanta dintre caracteristicile mesajului si centroid este minima

Problema tema: Clasificarea textelor pe baza sentimentelor

Mai tii minte ca tocmai ti-ai inceput munca ca si software developer la Facebook si ca faci parte din echipa care se ocupa cu partea de continut a platformei?

Utilizatorii sunt foarte incantati de noul algoritm de detectie a filtrelor in poze, asadar poti sa te ocupi de o noua functionalitate care ar face platforma mai atractiva. Utilizatorii posteaza o gama larga de mesaje, iar in feed-urile lor apar de multe ori prea multe mesaje negative si prea putine pozitive. Facebook incearca o noua functionalitate prin care sa detecteze sentimentele dintr-un mesaj si sa filtreze feed-urile utilizatorilor.

Task-ul tau este sa implementezi un algoritm care poate recunoaste sentimentele dintr-un text (pozitiv, negativ, ura, rasism, etc.)

Team leaderul echipei de ML iti propune urmatorul plan de lucru

- 1. devoltarea, antrenarea si testarea unui algoritm de tip k-means folosind data de tip numeric (de ex datele cu irisi)
- 2. Considerarea unei baze cu texte etichetate cu emotii (de ex. textele din data/review_mixed.csv sau https://github.com/sarnthil/unify-emotion-datasets/tree/master/datasets)
- 3. Extragerea de caracteristici din texte folosind diferite reprezentari precum:
 - o Bag of Words
 - o TF-IDF
 - o Word2Vec
 - o N-grams, etc.
- 4. Clasificarea textelor si etichetarea lor cu emotii folosind
 - o un algoritm de invatare supervizat (folosind etichetele pt emotiile asociate fiecarui text)
 - o un algoritm de invatare nesupervizat bazat pe k-means (fara a folosi etichetele pt emotiile asociate fiecarui text)
 - o un algoritm hibrid care combina tehncile de invare cu reguli ajutatoare (de ex prin folosirea unor reguli care verifica/numara aparitiile unor cuvinte polarized words (e.g. negative words such as bad, worst, ugly, etc and positive words such as good, best, beautiful, etc.)