TEMA 2 TIA

(Test de personalitate cu predictii bazate pe KNN)

**Prezentarea temei**

Aproape jumatate din toti studentii inmatriculati la o facultate din Romania in 2015 nu au terminat-o pana in 2021. Am considerat ca o tema ce abordeaza aceasta problema este interesanta. In acest sens, am creat un formular (<https://forms.gle/LaEqY1GpvACmtagD7> ) cu 11 intrebari cu raspuns de tip da/nu. Prima reprezinta concluzia la care vrem sa ajungem intrepretandu-le pe celelalte 10.

**Prezentarea setului de date si a etapei de preprocesare**

Setul de date consta in raspunsurile oferite de aproximativ 10 colegi + 10 oferite de mine avand diferite scenarii in minte +- duplicate. Am considerat potrivit sa inserez cateva duplicate deoarece este posibil ca unele raspunsuri sa se potriveasca, iar raspunsurile date erau deja destul de variate. Acestea sunt puse intr-un fisier data.csv si transformate in matrici.

Pentru ambii algoritmi am incercat sa pastrez o structura a datelor asemanatoare cu cea din cursurile in care algoritmii sunt prezentati.

*Preprocesarea pentru KNN*

-am transformat fisierul data.csv care consta in da/nu intr-o matrice in care da=1 si nu=0

O imagine care conține text, captură de ecran

Descriere generată automat

-am schimbat valorile pentru o intrebare astfel incat raspunsurile pentru toate intrebarile sa aiba corespondentul 1 daca raspunsul respectiv denota faptul ca facultatea a fost aleasa corespunzator: (este vorba despre Sunt mai multe zilele in care te simti demotivat decat motivat sa inveti pentru admitere/ pentru temele de la facultate? ). Daca se raspunde cu Nu, acesta va avea corespondentul 1 in loc de 0 deoarece este un raspuns care duce in cazul fericit.

- am clasificat cele 10 intrebari in 2 categorii: Factori Externi (intrebarile 2,5,6,7,11, de exemplu Tu ai fost cel care a ales facultatea? (si nu parintii)) si Factori Interni (restul intrebarilor).

O imagine care conține captură de ecran, text, tipografie, proiectare

Descriere generată automat

- am adunat punctajele de 1 si 0 din fiecare categorie si le-am pus in matricea ce urmeaza a fi folosita pentru antrenarea si testarea modelului

O imagine care conține text, captură de ecran, Font

Descriere generată automat

*Preprocesarea pentru Naïve Bayes*

-am transformat fisierul data.csv care consta in da/nu intr-o matrice in care da=1 si nu=0

**Prezentarea aplicabilității algoritmilor propuși pentru problematica aleasă**

*KNN*

Deoarece acesta funtioneaza prin calcularea unor distante intre intrari , am interpretat acest lucru in felul urmator: cu cat cele 2 trasaturi sunt mai mari, cu atat e mai probabil ca in final clasa sa fie 1 (da, facultatea e buna). Asadar, cu o fereastra de 3 (numar impar + proportional cu marimea setului de date) pentru a marca primele 3 cele mai apropiate puncte, consider ca se poate face o predictie satisfacatoare.

Naïve Bayes

Aici am putut pastra fiecare intrebare ca pe o trasatura in sine, si, mergand pe principiul ca indivizii carora li se potriveste facultatea vor raspunde statistic la fel la mai mult de jumatate din intrebari si viceversa, pare un algoritm potrivit. De asemenea setul de date este mic si variabilele sunt discrete, deci problema intra in tiparul celor din curs.

**Descrierea procesului de antrenare, validare, testare**

Pentru ambele am ales sa impart datele in 80-20 la suta pentru antrenare si testare. Pentru Naïve Bayes, am introdus si posibilitatea de a raspunde la intrebari din consola iar modelul sa faca o predictie. O imbunatatire ar fi sa retina informatiile din acest quiz si sa se si antreneze cu ele (nu am implementat aceasta functionalitate).

**Avantaje, dezavantaje si concluzii**

Ambii algoritmi sunt usor de inteles .Ambii accepta seturi mici de date. Acuratetea este una satisfacatoare. Pentru 30+ intrari avem peste 85% pentru ambii, pentru 22 avem 80% pentru ambii. Pastrand datele nereduntante, avem 100% pentru ambii. Faptul ca sunt identice o pun pe seama setului mic de antrenament si de testare (practic pentru testare raman 3,4,5 scenarii). Pe masura ce cresc setul de date cu intrari care se repeta, acuratetea scade pana la aproximativ 70%. Cel mai relevanta mi se pare acuratetea cand intrarile sunt unice, desi este de 100%. Cred ca este asa deoarece raspunsurile sunt destul de polarizate in general la chestionar si cred ca acest lucru se repeta si la esantioane mai mari de date.

Consider ca KNN este mai potrivit deoarece scorurile create se aseamana cu testele reale de personalitate. Ca si rezultate, sunt asemanatoare ambele variante. Mai jos avem quiz-ul cu Naïve Bayes.

O imagine care conține text, captură de ecran, Font

Descriere generată automat