Tema 3 TIA

(Evaluare model de clasificare a imaginilor- KNN vs Naïve Bayes)

Modificarile aduse de mine programului din laboratorul 4:

* Am adaugat matricile de confuzie
* Am adaugat curba ROC (receiver operating characteristic curve) pentru fiecare clasa (pentru ambii algoritmi)
* Am comparat AUC (Area under the ROC Curve) pentru cei 2 algoritmi

**Matricile de confuzie**

O imagine care conține text, captură de ecran, diagramă, număr

Descriere generată automat

O imagine care conține text, captură de ecran, diagramă, Dreptunghi

Descriere generată automat

O imagine care conține text, captură de ecran, diagramă, număr

Descriere generată automat

O imagine care conține text, captură de ecran, diagramă, Dreptunghi

Descriere generată automat

Sunt afisate atat pentru setul de antrenare cat si de testare. Voi pune mai mult accentul pe curba ROC deoarece sumarizeaza matricile de confuzie.

Pentru KNN, acuratetea pe setul de antrenare este de 0.85 iar pe cel de test este de 0.87, ceea ce pare rezonabil, insa pentru Naïve Bayes, setul de antrenare are acuratete mai mare decat cel de testare. (0.79 vs 0.70). O posibila cauza ar fi fenomenul de underfitting (modelul nu primeste suficiente date sau nu reuseste sa surprinda complexitatea lor). Ca si cauze s-ar putea numara faptul ca acest algoritm nu raspunde la fel de bine ca si KNN la setul de date primit, poate setul nu este suficient de mare, sau algoritmul este prea simplu pentru date. O posibila solutie ar fi cresterea numarului de epoci. (informatii preluate de pe: <https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/> )

**Curba ROC**

Atunci cand avem o clasificare, in functie de natura problemei, putem pune un obiect intr-o clasa atunci cand suntem, sa zicem pest 80% siguri, sau 50% ca apartine clasei etc. De asemenea, pot avea prioritate cazurile TP fata de FP- daca vrem sa detectam o boala mortala, este mai important sa avem in vedere si cazurile cu suspiciuni mici decat sa nu catalogam ca bolnava o persoana sanatoasa. Daca luam exemplul unui drug test, poate fi pe dos: ne intereseaza mai mult cazurile unde este foarte probabil ca persoana in cauza sa fi consumat droguri, si nu dorim sa adaugam la “dosarul” unei persoanei un potential consum decat daca motivele sunt intemeiate. Curba ROC ofera o reprezentare grafica pentru aceasta problema, mai convenabila decat a evalua toate matricile de confuzie. Graficul ROC porneste din coltul din stanga jos care semnifica un threshold de 1 (100%) iar apoi, odata cu scaderea acestuia, se indreapta catre dreapta. In modelul nostru, nu pot spune daca una din cele doua componente este mai importanta decat cealalta. Linia diagonala punctata reprezinta ca proporita de clasificari TP = FP. Un model bun, ar trebui sa se tina cat mai deasupra acelei diagonale, asa cum vom vedea evaluand si AUC.



Ambii algoritmi in comun tendinta generala. As spune ca cel mai convenabil threshold pentru problema noastra este undeva la mijloc, acolo unde True Pozitive Rate este de 0.7 si False Pozitive Rate este 0.3, adica rosul se clasifica ca rosu/albastru in 70% din cazuri si 30 la suta din albastu/rosu este clasificat gresit (am marcat cu o bulina rosie).

Ce au diferit cele doua curbe este faptul ca Naïve Bayes este mult mai zimtata la mijloc. Aceasta se datoreaza faptului ca performantele modelului nu sunt consistente in jurul thresholdului de 50%. Acei zimti apar doar in zona de mijloc si, personal, ma duc cu gandul la natura probabilistica a algoritmului. Nu stiu daca este corect, insa cred ca atunci cand oferim o limita de 50% peste care consideram ceva adevarat(sau in jurul acestei valori), algoritmul incepe sa oscileze intre rosu si albastru, sa incadreze la nimereala anumite poze.

KNN are o granita de decizie mai flexibila decat Naïve Bayes.

**Scenariu**

O posibila aplicatie a algoritmului ar fi diagnosticarea degeraturilor. Expusa la frig pielea umana devine mai intai rosie, apoi albastuie, apoi, cand devine neagra este deja ireveribil. In acest caz am vrea sa reperam cat mai clar TP pentru albastru, deci sa avem o rata TP cat mai mare, in detrimentul cazurilor mai putin grave diagnosticate gresit. Asadar am vrea sa ne aflam in zona marcata cu mov.

O imagine care conține linie, Interval, diagramă, pantă

Descriere generată automat



**O imagine care conține text, linie, diagramă, Interval

Descriere generată automat**



**AUC**

Faptul ca AUC este mai mare pentru KNN, implica faptul ca pentru problema noastra de clasificare, KNN se potriveste mai bine, desi celalat algoritm este pe aproape. Poate cu o atentie sporita la threshold sau cu o prelucrare mai atenta a setului de date, rolurile se vor inversa. Aceasta metrica acopera toate scenariile, insa pentru unul singular, poate Naïve Bayes este mai bun.

**Concluzii**

* Folosirea Naïve Bayes cu threshold de 50% nu cred ca este recomandata
* KNN are performante mai bune, atat datorita acuratetii crescute, cat si pentru ca performeaza mai bine pe setul de testare