Raport - Aplicație practică (ML)

Chichirău Claudiu-Constantin, 3A2

December 29, 2023

1 Alegerea Algoritmului

Am ales algoritmul Naive Bayes pentru această problemă din următoarele motive:

- **Adecvare Teoretică**: Naive Bayes este un algoritm de clasificare probabilistică bazat pe teorema Bayes. Este cunoscut pentru eficiența sa în problemele de clasificare a textului, cum ar fi filtrarea spamului. Acesta presupune că prezența unui anumit atribut într-o clasă nu este legată de prezența altor atribute, ceea ce este o simplificare care adesea funcționează bine pentru text.
- **Eficiență**: Naive Bayes este un algoritm eficient atât din punct de vedere al timpului de antrenare, cât și al spațiului de stocare. Acesta necesită doar o singură trecere prin date pentru antrenare și stochează doar numărul de apariții ale cuvintelor pentru fiecare clasă.

2 Compararea cu alți Algoritmi

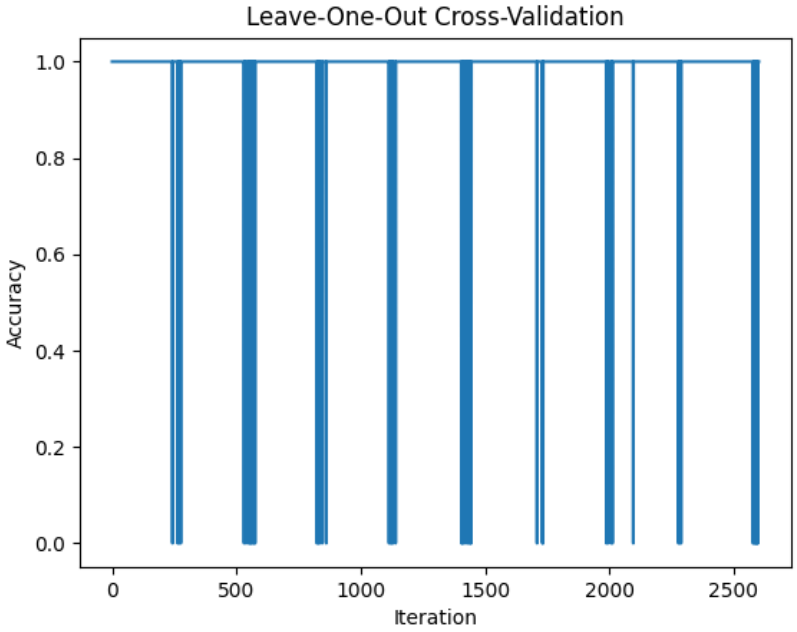
Am considerat și alți algoritmi pentru a rezolva această problemă, dar am ales Naive Bayes din următoarele motive:

- **ID3**: Deși poate fi eficient pentru unele probleme, este mai puțin adecvat pentru clasificarea textului deoarece arborele de decizie ar putea deveni foarte mare și complex datorită numărului mare de cuvinte unice din text.
- **AdaBoost**: AdaBoost este un algoritm de învățare a ansamblului care combină mai mulți clasificatori slabi pentru a crea un clasificator puternic. Deși este un algoritm puternic, este mai complex și necesită mai mult timp de antrenare comparativ cu Naive Bayes.
- **K-means**: K-means este un algoritm de clustering, nu un algoritm de clasificare. Nu este adecvat pentru această problemă deoarece avem etichete pentru datele noastre și dorim să facem predicții pe baza acestor etichete.

3 Evaluarea Performanței cu Cross-Validare Leave-One-Out

Am implementat strategia de cross-validare Leave-One-Out pentru a evalua performanța modelului nostru. În LOO, fiecare exemplu din setul de date este folosit o dată ca set de testare, în timp ce restul exemplelor sunt folosite ca set de antrenare. Acest proces este repetat pentru fiecare exemplu din setul de date.

Rezultatele evaluării cu LOO sunt ilustrate mai jos:

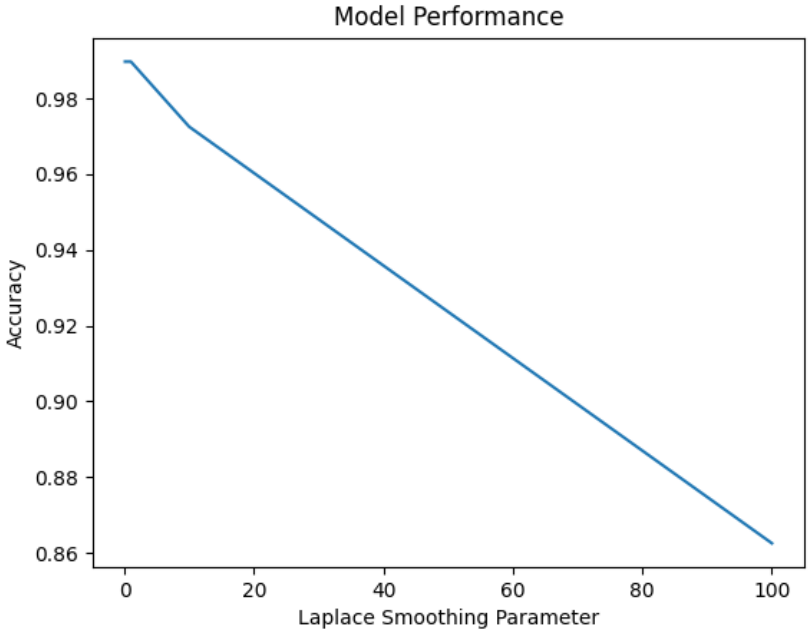


4 Performanța Algoritmului

Am evaluat performanța algoritmului **Naive Bayes** pe setul de date de testare, măsurând acuratețea obținută pentru diferite valori ale parametrului **Laplace**. Rezultatele sunt prezentate în tabelul și poza de mai jos.

După cum se poate observa, acuratețea obținută este semnificativ mai bună decât orice strategie trivială, cum ar fi alegerea mereu a aceleiași clase sau alegerea clasei la întâmplare. Acest lucru demonstrează eficacitatea algoritmului Naive Bayes pentru problema de clasificare a email-urilor spam.

Laplace	Acuratețe
0.1	0.98969
5.0	0.98969
10.0	0.97250
100.0	0.86254

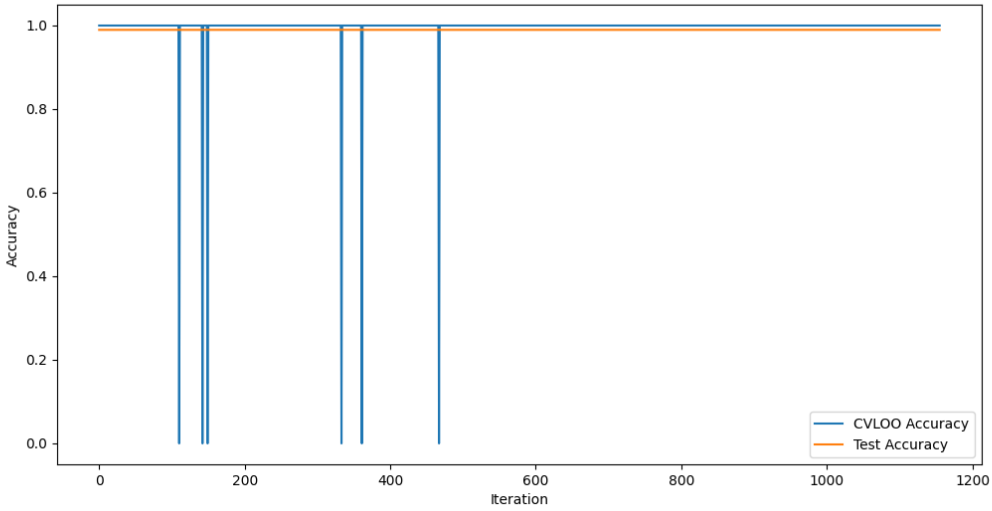


Am comparat, de asemenea, performanța algoritmului nostru cu cea a altor algoritmi pe care i-am considerat pentru această problemă. Rezultatele acestei comparații vor fi prezentate în secțiunea următoare.

5 Îmbunătățirea Performanței cu Auto-Antrenare

Am implementat o metodă de învățare semi-supervizată numită auto-antrenare pentru a îmbunătăți performanța algoritmului Naive Bayes. În auto-antrenare, modelul este mai întâi antrenat pe datele etichetate, apoi acesta este folosit pentru a prezice etichetele pentru datele neetichetate. Aceste etichete prezise sunt apoi adăugate la setul de antrenare și modelul este reantrenat pe noile date de antrenare.

Rezultatele implementarii CVLOO sunt ilustrate mai jos:


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
200
400
600
800
1000
1200
Iteration
CVLOO Accuracy
Test Accuracy