Proiect practic la Machine Learning

Mocanu Octavian si Robu Victor

January 11, 2024

1 Introducere

În era digitală contemporană, gestionarea și clasificarea eficientă a emailurilor reprezintă o provocare semnificativă, atât pentru utilizatorii individuali, cât și pentru organizații. Cu un volum imens de emailuri care sunt trimise și primite zilnic, identificarea și filtrarea emailurilor nedorite sau spam devine o necesitate critică pentru menținerea eficienței și securității comunicațiilor digitale. În acest context, algoritmii de învățare automată oferă soluții promițătoare pentru automatizarea și îmbunătățirea procesului de clasificare a emailurilor.

Pentru această problemă, am ales să explorăm și să comparăm doi algoritmi consacrați de învățare automată: Bayes Naiv și ID3. Motivația noastră pentru alegerea acestor doi algoritmi se bazează pe natura și caracteristicile specifice ale fiecăruia, care sunt potrivite pentru clasificarea textului.

1.1 Motivatie pentru alegerea algoritmului Bayes Naiv

Bayes Naiv este un algoritm bazat pe teorema lui Bayes și este cunoscut pentru eficiența sa în lucrul cu seturi mari de date. Simplitatea sa computationala, alături de performanța robustă în condiții de incertitudine și zgomot din date, îl face o alegere excelentă pentru identificarea spamului în emailuri. Capacitatea sa de a învăța rapid din date și de a face predicții eficiente, chiar și cu o cantitate relativ mică de date de antrenament, este de asemenea un avantaj semnificativ în contextul clasificării emailurilor.

1.2 Motivatie pentru alegerea algoritmului ID3

Pe de altă parte, ID3, un algoritm bazat pe arbori de decizie, a fost ales datorită abordării sale intuitive și transparente în clasificarea datelor. ID3 utilizează entropia și câștigul de informație pentru a construi un arbore de decizie care clasifică datele în categorii distincte. În cazul clasificării emailurilor, acest algoritm oferă o reprezentare vizuală clară a deciziilor luate în procesul de clasificare, facilitând înțelegerea și interpretarea rezultatelor.

Prin compararea acestor doi algoritmi în contextul clasificării emailurilor, dorim să determinăm care dintre ei oferă performanțe superioare, atât în termeni de acuratețe, cât și de eficiență, în abordarea acestei probleme complexe și dinamice.

1.3 Metode

Setul de date pe care s-a realizat acest experiment este format din 10 foldere, fiecare continand atat email-uri de tip spam cat si email-uri normale. Antrenarea a fost realizata folosind primele 9 foldere, ultimul fiind utilizat pentru testare. Acest experiment a fost repetat pentru 4 categorii de email-uri diferite (fiecare avand 10 foldere): lemm, bare, stop si lemm_stop.

Pentru a evalua robustețea modelului, a fost implementată validarea Leave-One-Out Cross-Validation (LOOCV). În această metodă, unul dintre cele 9 foldere de antrenament a fost folosit ca set de validare pentru fiecare iterație, în timp ce restul datelor din celelalte 8 foldere au fost folosite pentru antrenament. Aceasta a permis evaluarea modelului într-un mod care reduce impactul aleatoriu al selecției datelor de antrenament și de testare, oferind o imagine mai clară asupra performanței reale a modelului.

2 Implementarea algoritmului Bayes Naiv

Inainte de antrenare, datele din cele 9 foldere au fost grupate in doi vectori: features si labels, unde $features[i]$ contine email-ul cu indicele i, iar $labels[i]$ va lua valori din $0, 1$, 0 insemnand email normal, iar 1 email de tip spam.
In functia de antrenare, fiecare email va fi impartit in cuvintele componente si asociat unui vector ce va retine doar email-urile spam, respectiv normale, creeand in acelasi timp si un "bag of words" ce va contine toate cuvintele si frecventa lor din toate email-urile.
In continuare se vor calcula argumentele din urmatoarele formule:

$$P(\text{Spam}|X) = P(x_1|\text{Spam}) \cdot P(x_2|\text{Spam}) \cdot \ldots \cdot P(x_n|\text{Spam}) \cdot P(\text{Spam})$$

, unde $P(\text{Spam})$ este probabilitatea totala ca un email sa fie spam, iar $P(x_i|\text{Spam})$ este probabilitatea ca un cuvant x_i din email-ul X sa apara stiind ca email-ul este spam, fiind calculata in felul urmator:

$P(x_i|\text{Spam}) = \frac{\text{Count of } x_i \text{ in Spam}+\alpha}{\text{Total word count in Spam}+\alpha \cdot \text{Vocabulary size}}$, unde α este o constanta ce ia valoarea 1, care, impreuna cu vocabulary size, ajuta la uniformizarea valorilor si la evitarea probabilitatilor egale cu 0 in cazul cuvintelor care nu apar in email-urile de tip spam. Aceasta operatie poarta denumirea de "Laplace Smoothing".
In mod similar se calculeaza si $P(\text{Normal}|X)$, aceste doua valori urmand sa fie comparate in functia de predictie pentru a decide carei clase va apartine un email oarecare. De mentionat este faptul ca au fost folosite valorile in logaritm natural in cazul tuturor probabilitatilor pentru a evita eventualele probleme legate de reprezentarea numerica.

2.1 Implementarea aLgoritmului ID3

Implementarea algoritmului ID3 pentru clasificarea email-urilor în categoriile spam sau normale a fost realizată într-un mod structurat și metodic. Inițial, datele din cele 9 foldere au fost încărcate și etichetate corespunzător ca spam sau normale, pe baza titlurilor fișierelor. Fiecare email a fost apoi procesat pentru a construi un "bag of words", reprezentând totalitatea cuvintelor unice din toate email-urile, alături de frecvența lor de apariție.

În etapa următoare, a fost aplicat un proces de pruning pentru a reduce dimensiunea și complexitatea setului de date. În loc să folosim întregul "bag of words", am selectat un număr limitat de cuvinte, bazându-ne pe cele mai frecvente cuvinte întâlnite în email-uri. Acest pas de reducere a dimensiunii caracteristicilor a avut un dublu scop: în primul rând, a simplificat modelul, făcându-l mai ușor de analizat și interpretat, și în al doilea rând, a contribuit la îmbunătățirea performanței de clasificare prin eliminarea cuvintelor mai puțin semnificative care ar fi putut adăuga zgomot și ambiguitate.

Prin selecția acestor cuvinte-cheie, am putut focaliza analiza pe termenii cei mai relevanți și mai probabili de a diferenția între email-urile spam și cele normale. Acest proces de pruning a avut ca rezultat un set de date mai curat și mai relevant pentru antrenarea arborelui de decizie ID3.

În cadrul procesului de antrenare, s-au calculat parametrii necesari pentru construirea arborelui ID3. Entropia fiecărui atribut a fost determinată pentru a măsura gradul de impredictibilitate sau "dezordine" din date. Aceasta a fost esențială pentru calculul câștigului de informație al fiecărui atribut, care indică cât de bine un atribut poate separa setul de date în clase corespunzătoare. Algoritmul a selectat atributul cu cel mai mare câștig de informație pentru a împărți setul de date, procedând astfel până când toate datele dintr-un nod au fost perfect clasificate sau nu au mai rămas atribute pentru continuarea împărțirii.

Pentru fiecare nod al arborelui, algoritmul a verificat dacă datele din acesta aparțin unei singure clase. În caz afirmativ, nodul a devenit un nod frunză cu eticheta clasei respective. În caz contrar, s-a ales un nou atribut pentru împărțirea ulterioară a datelor. Acest proces de împărțire a fost repetat recursiv pentru fiecare ramură nou creată a arborelui.

După construirea arborelui, s-a realizat testarea performanței acestuia pe setul de date de testare. Fiecare email din setul de testare a fost transformat într-un vector de caracteristici conform "bag of words" și apoi a fost clasificat folosind arborele ID3. Acuratețea modelului a fost evaluată comparând predicțiile modelului cu etichetele reale ale email-urilor.

2.2 Rezultate experiment

	BN	ID3
antrenare	0.995	0.999
testare	0.989	0.945
LOOCV	0.992	0.952

Table 1: Acuratete pentru categoria bare

	BN	ID3
antrenare	0.995	0.999
testare	0.989	0.945
LOOCV	0.990	0.959

Table 2: Acuratete pentru categoria lemm

	BN	ID3
antrenare	0.998	0.999
testare	0.989	0.945
LOOCV	0.992	0.950

Table 3: Acuratete pentru categoria lemm_stop

	BN	ID3
antrenare	0.998	0.999
testare	0.989	0.945
LOOCV	0.993	0.962

Table 4: Acuratete pentru categoria stop

3 Concluzii

Experimentul nostru a avut ca scop evaluarea și compararea performanțelor a doi algoritmi populari de învățare automată, Bayes Naiv și ID3, în contextul filtrării emailurilor. Rezultatele obținute indică faptul că algoritmul Bayes Naiv a excelat în această sarcină, depășind performanța algoritmului ID3 în toate testele efectuate.

Unul dintre motivele principale pentru această superioritate a fost capacitatea algoritmului Bayes Naiv de a gestiona eficient și precis caracteristicile de înaltă dimensiune, tipice datelor textuale din emailuri. Acesta a demonstrat o adaptabilitate notabilă în diferite condiții de testare, inclusiv în scenarii cu date zgomotoase și neuniforme, care sunt comune în fluxurile reale de emailuri.

Pe de altă parte, algoritmul ID3, deși a oferit o înțelegere valoroasă asupra procesului de clasificare prin structura sa arborelui de decizie, nu a atins nivelul de acuratețe necesar pentru a fi considerat practic viabil în această aplicație. Limitările sale în manipularea atributelor textuale și sensibilitatea la datele zgomotoase au fost factori contribuitori la performanța sa relativ inferioară comparativ cu Bayes Naiv.

În plus, Bayes Naiv s-a dovedit a fi mai ușor de implementat și a necesitat mai puțină ajustare a parametrilor, făcându-l o soluție mai eficientă din punct de vedere al timpului și resurselor pentru filtrarea emailurilor. Acest avantaj este deosebit de important în contextul în care companiile și utilizatorii individuali caută soluții rapide și eficiente pentru gestionarea emailurilor nedorite.

În concluzie, experimentul nostru a demonstrat că, deși ambele algoritmi au punctele lor forte, Bayes Naiv este mai adecvat pentru sarcina de filtrare a emailurilor, oferind o combinație optimă de acuratețe, eficiență și ușurință în implementare. Aceste descoperiri subliniază importanța selectării unui algoritm potrivit pentru specificul problemei și a setului de date, o decizie crucială în domeniul învățării automate.