Raport Tema Practică Machine Learning

Al shrafat Maroan, Aron Constantin-Robert

January 2024

1 Înțelegerea datelor

Primul pas în abordarea acestei teme, a fost procesarea setului de date primit. Pentru algoritmul nostru ales, am hotărât să pregătim datele în urmatorul mod:

1. Am eliminat toate numerele existente din conținutul email-urilor;
2. Am eliminat cuvintele de stop și semnele de punctuație;
3. Am tokenizat conținutul email-urilor;

Această procesare a fost necesară pentru a obtine o listă de cuvinte care vor folosi la construirea unui dicționar de frecvență, necesar pentru algoritmul Naive Bayes.

2 Alegerea algoritmului

Am ales să implementăm Naive Bayes deoarece este un algoritm rapid și usor de implementat. Față de alte concepte precum Arbori de decizie, AdaBoost sau Clusterizare, Naive Bayes nu necesită mulți parametri de configurat. Naive Bayes există în mai multe tipuri: Multinomial NB, Gaussian NB și Bernoulli NB. Noi am ales varianta Multinomial, deoarece din setul de date primit, am putut cu ușurință să construim un dicționar de frecvență, care să servească drept input pentru algoritm. În implementarea algoritmului, am folosit diferite tehnici care să îmbunatațească calitatea acestuia. De exemplu, am adăugat o aplatizare a probabilităților, astfel încât dacă un cuvânt nu se gasește în dicționarul generat, să

nu influențeze restul probabilităților și în final să ajungem la probabilitate 0. De asemenea, pentru a evita lucrul cu numere mici, am aplicat operația logaritm pe probabilități.

3 Grafic acuratețe

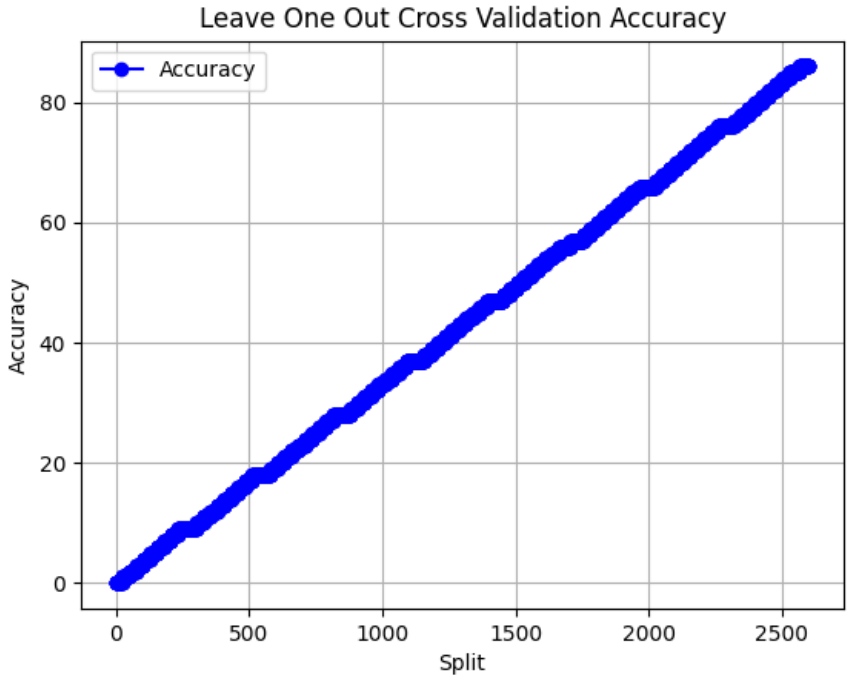

Figure 1: LOOCV Accuracy