Mathematical Model: Multinomial Naive Bayes Text Classifier

Prosie Radu-Teodor

November 15, 2025

Introduction

Given a text document, the goal is to classify it into one of a finite set of classes (e.g., AI-generated vs human-written) using only the counts of words in the document.

Notation

- $\mathcal{C} = \{c_1, \ldots, c_K\}$: set of classes
- $\mathcal{V} = \{w_1, \ldots, w_V\}$: vocabulary
- For a given document d, $\mathbf{x} = (x_1, \ldots, x_V)$ where $x_j =$ count of word w_j in document d
- N_c: number of training documents in class c
- $n_{c,j}$: total count of word w_j in class c
- $T_c = \sum_j n_{c,j}$: total word count in class c
- $\alpha > 0$: smoothing parameter

Model

Assumptions:

1. Bag-of-words: word order is ignored
2. Conditional independence: words occur independently given the class.

Under these assumptions, the likelihood of a document to have class c is:

$$P(\mathbf{x} \mid c) = \prod_{j=1}^{V} P(w_j \mid c)^{x_j}$$

where $P(w_j \mid c) = \frac{n_{c,j}}{T_c}$. By inverting the probabilities using Bayes' rule, we now have:

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c)P(c)}{P(\mathbf{x})}$$

Now, we want to find a c such that $P(c \mid \mathbf{x})$ is maximized. Since $P(\mathbf{x})$ is the same for all classes, it can be ignored, and we just have to look for the class c that maximizes $P(\mathbf{x} \mid c)P(c)$.

Caveats and computation

To avoid underflow during the computation of probabilities, we shall maximize the logarithm of the entire expression. Therefore, the predicted class for a document $\mathbf{x}$ is:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} \left[\log P(c) + \sum_{j=1}^{V} x_j \log P(w_j \mid c) \right]$$

Additionally, to avoid having to deal with events our model assumes have a 0 probability of happening, such as seeing a new word in the document we're testing the model on (after training), we will use Laplace smoothing. To be specific, we will change the definition of $P(w_j \mid c) = \frac{n_{c,j}}{T_c}$ to $P(w_j \mid c) = \frac{n_{c,j} + \alpha}{T_c + \alpha \cdot V}$. With these changes, we can calculate the predicted class for a document in $O(CV)$ operations.