Section 9 Naive Bayes

1. Naive Bayes
2. Laplace Smoothing

Text Classification Example: Separate emails as Spam=0 and Not Spam=1.

Spam Email Sample:

Dear Good Friend

I am Abdoul Issouf, I work for BOA bank Ouagadougou Burkina Faso. I have a business proposal which concerns the transfer of ($13.5 Million US Dollars) into a foreign account. Everything about this transaction shall be legally done without any problem. If you are interested to help me, Please keep this transaction as a Top Secret to your self till the Money get into your account in your Country OK. and I will give you more details as soon as I receive your positive response. You will be Entitled to 50%, 50% will be for Me If you are willing to work with me send me immediately the information listed bellow.

Your Name .....

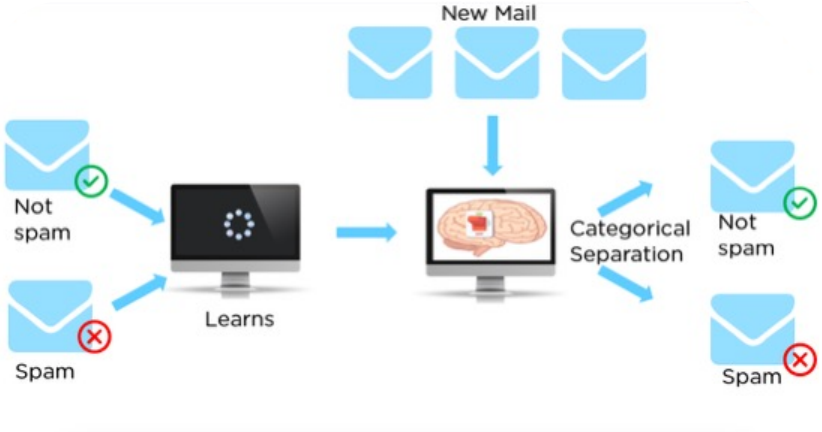
Your Nationality .....

Your Age .....

Female Or Male .....

Your Occupation .....

Your Private Telephone.....



Not Spam Email Sample

Dear Instructors,

We wanted to share an update on our paid model based on feedback we received from our community regarding the ad-supported option. This will not affect students and faculty at your institute where your school or department has or is in the process of purchasing a paid license.

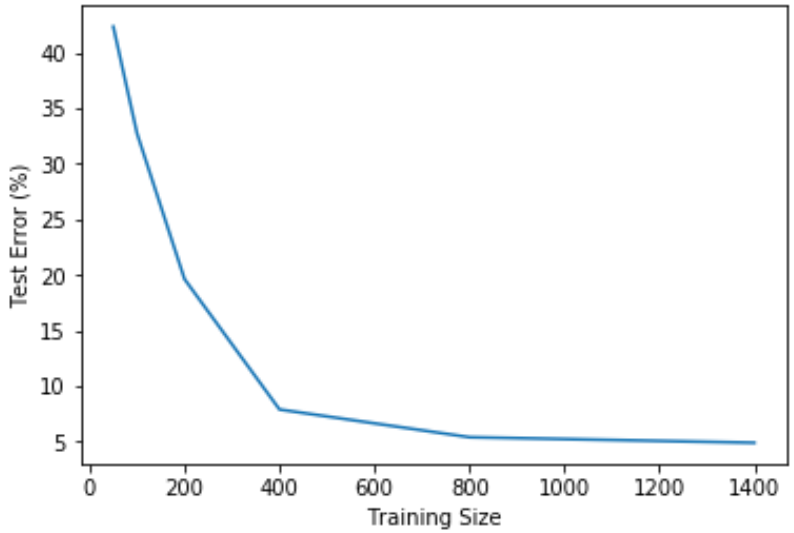
As you may know, starting 2021, Piazza is moving to a paid model so we can continue to support our users and innovate on new product features.

Originally, for schools or departments that needed additional time to get a paid license in place, we had contemplated having an unpaid ad-supported version available; Instead we are now offering a contribution-supported unpaid version of Piazza (much like how Wikipedia asks for donations). This shift to a contribution-supported model addresses the privacy concerns that we heard from faculty around the ad-supported model.

We will represent an email via a feature vector $\vec{x} \in \mathbb{Z}_2^d$, called vocabulary, whose length *d* is equal to the number of words in the dictionary, e.g., d=171,146.

In practice, we should build a dictionary with only “medium frequency” words, say d=2000. (abil absolut abus access accid .... young yourself zip)

$$\vec{x} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$


Test Error (%)
40
35
30
25
20
15
10
5
0
200
400
600
800
1000
1200
1400
Training Size

Independence

Conditional Independence

Conditional Independence Joke: A survey has pointed out a positive and significant correlation between the number of accidents and wearing heavy coats in Boston. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally, another study pointed out that people wear coats when it snows…

$$P(\text{Accident} \mid \text{Coats}, \text{Snow}) = P(\text{Accident} \mid \text{Snow})$$

$$P(\text{Accident}, \text{Coats} \mid \text{Snow}) = P(\text{Accident} \mid \text{Coats}, \text{Snow}) P(\text{Coats} \mid \text{Snow})$$

$$= P(\text{Accident} \mid \text{Snow}) P(\text{Coats} \mid \text{Snow})$$