

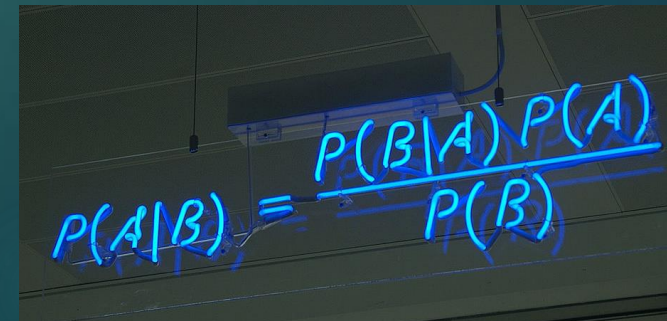
# Naïve Bayes – A Machine Learning Technique in Spam Filtering

BRANDON DRUSCHEL

COMPUTER SCIENCE, B.S.

## ABSTRACT

**Bayes' probability theorem** can be applied to spam filtering in our emails. **Naïve Bayes spam filtering** is a baseline machine learning technique for dealing with spam that can tailor itself to the email needs of individual users and is one of the oldest ways of doing spam filtering.



A photograph of a whiteboard with the Bayes' theorem formula written in blue marker. The formula is  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ . The whiteboard is illuminated by a blue light, and the background is dark.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Bayesian Classification – Bayes' Theorem

- ▶ **Bayes' Theorem** describes the probability of an event based on prior conditions. The following formula demonstrates the rule when applied to filtering email :

$$P(c | \mathbf{x}) = \frac{P(\mathbf{x} | c)P(c)}{P(\mathbf{x})} = \frac{P(\mathbf{x} | c)P(c)}{P(\mathbf{x} | S)P(S) + P(\mathbf{x} | L)P(L)}$$

- ▶ Two categories: **S** (spam) and **L** (legitimate mail)
- ▶  **$P(\mathbf{x} | c)$**  denotes the probability of obtaining a message with **feature vector  $\mathbf{x}$**  from **class  $c$** .
- ▶ What we want to know is, given a **message  $\mathbf{x}$** , what **category  $c$**  “produced” it. That is, we want to know the probability  **$P(c | \mathbf{x})$** .



Portrait (purportedly) of  
**Thomas Bayes**

# The Naïve Bayes classifier

- ▶ **Naïve Bayes** is a classification algorithm for binary (two-class) and multi-class classification problems. It was introduced into the text retrieval community in the early '60s, and remains a popular baseline method for *text categorization*
- ▶ It's called '*naïve*' Bayes because the calculation of the probabilities for each hypothesis are *simplified* to make their calculation manageable.
- ▶ Naïve Bayes classifiers typically use *bag-of-words* features to identify spam email.
  - ▶ Bag-of-words model: a simplifying representation used in natural language processing and information retrieval.
- ▶ In order to construct a Bayesian classifier for spam detection you must somehow be able to determine the probabilities  $P(\mathbf{x} | \mathbf{c})$  and  $P(\mathbf{c})$  for any  $\mathbf{x}$  and  $\mathbf{c}$ . It is clear that you can never know them exactly, but we may estimate them from the *training data*.

# Naïve Bayes spam filter

- ▶ The spam filter is **trained** by manually indicating whether a new email is spam or not.
  - ▶ For all words in each training email, the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database.
- ▶ Naïve Bayes classifiers work by correlating the use of **tokens** (typically words, sometimes other things), with spam and non-spam emails and then using **Bayes' theorem** to calculate a probability that an email is or is not spam.
- ▶ The initial training can usually be refined when wrong judgements from the software are identified (*false positives* or *false negatives*). That allows the software to dynamically adapt to the ever-evolving nature of spam.
- ▶ Email marked as spam can then be automatically moved to a "Junk" email folder, or even deleted outright.



# QUESTION: Besides false-positives, are there down sides to Bayesian spam filtering?

- ▶ Answer: Yes!
- ▶ Depending on the implementation, Bayesian spam filtering may be susceptible to **Bayesian poisoning**, which can degrade the effectiveness of spam filters that rely on Bayesian filtering.
  - ▶ A spammer practicing Bayesian poisoning will send out emails with large amounts of legitimate text (gathered from legitimate news or literary sources).
- ▶ Another technique used to try to defeat Bayesian spam filters is to replace text with pictures, either directly included or linked. The whole text of the message, or some part of it, is replaced with a picture where the same text is "drawn".
  - ▶ A solution used by Google in its Gmail email system is to perform an **OCR (Optical Character Recognition)** on every mid- to large-size image, analyzing the text inside.



# References

- ▶ *Building a Spam Filter from Scratch Using Machine Learning—  
Machine Learning Easy and Fun:*  
<https://medium.com/analytics-vidhya/building-a-spam-filter-from-scratch-using-machine-learning-fc58b178ea56>
- ▶ *Machine Learning Techniques in Spam Filtering:*  
[https://pdfs.semanticscholar.org/1ef7/d60e44998647847ca0636551eb0aaa9fa20e.pdf?\\_ga=2.18958264.1045223485.1554907544-23983939.1554907544](https://pdfs.semanticscholar.org/1ef7/d60e44998647847ca0636551eb0aaa9fa20e.pdf?_ga=2.18958264.1045223485.1554907544-23983939.1554907544)
- ▶ *Machine Learning Methods for Spam E-mail Classification:*  
<http://airccse.org/journal/jcsit/0211jcsit12.pdf>
- ▶ *Naive Bayes for Machine Learning:*  
<https://machinelearningmastery.com/naive-bayes-for-machine-learning/>