**Naïve Bayes for Sentiment Analysis Notes**

* Generative model: assumes features are generated by some model, we combine the posterior probability for each label, then we compare to find the highest posterior probability
* Can also look at probabilities for each sentence
* Multinomial distribution = the probability of observing **word counts/presences** across various categories. Finding a best-fit multinomial distribution. Categories are **word types**.
* How to convert string to numbers?
* Word types = words in the vocabulary
* Tokens = individual instances of words in the texts.

**Pros of Naïve Bayes** (Vanderplas)

* p. 382
* Extremely fast and simple, suitable for **high-dimensional datasets**, few tunable parameters, useful as “a quick-and-dirty baseline for a classification problem”, p. 385: the *predict\_proba* method allows you to analyze probability of a set of features having each label, good to analyze uncertainty.
* P. 389: fast for training and testing, straightforward probabilistic prediction, easy to interpret, few tunable parameters, good initial baseline classification before exploring more sophisticated methods. Perform well for **well-separated categories**, very **high-dimensional data where model complexity is less important**. As the dimensions of a dataset grow, it is much less likely for 2 points to be found close together, as they must be close *in every single dimension* to be close overall. So clusters in high dimensions tend to be more separate, assuming the dimensions actually add information SUCH AS TEXT DATA 🡺 often used on tasks such as spam detection, both commercial and open-source applications, which typically deal with unbalanced datasets (spam is a rarer event) “Naive Bayes is very popular in commercial and open-source anti-spam e-mail filters. “ 🡪 [Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, January). Spam filtering with Naive Bayes - Which Naive Bayes? In CEAS.] <https://www.researchgate.net/publication/221650814_Spam_Filtering_with_Naive_Bayes_-_Which_Naive_Bayes>
* Good for multi-class prediction (<https://www.upgrad.com/blog/naive-bayes-classifier/>) , good for text (high-dimensional) inputs compared to a few numerical features (<https://www.upgrad.com/blog/naive-bayes-classifier/> ). “Its assumption of feature independence, and its effectiveness in solving multi-class problems, makes it perfect for performing Sentiment Analysis. Sentiment Analysis refers to the identification of positive or negative sentiments of a target group (customers, audience, etc.)
* “
* Bayes’ theorem: describes the “relationship of conditional probabilities of statistical quantities”, finding the probability of a label *given some observed features* (i.e. word counts), P(L | features).
* “Generative model”: it specifies the “hypothetical random process that generates the data”, specifies the generative model for each label (main part of training).
* Multinomial = data is drawn from a **multinomial distribution**, p. 386: 🡪 “the features are assumed to be generated from a simple multinomial distribution”.
* Cons: simplifying assumptions (independence and equal weight of each feature)
* “Its accuracy before was 86.4%, so we actually got a very slight decrease. In general, support vector machine and logistic regression-based algorithms will benefit less, or perhaps even be harmed, by pre-filtering the training features. This is because these algorithms are able to learn feature weights that correspond to the significance of each feature, whereas Naive Bayes algorithms do not.” 🡪 Naïve Bayes do not weight features by importance unlike SVMs, therefore are more sensitive to filtering and removing less important features than other classification algorithms. 🡺  [Python 3 Text Processing with NLTK 3 Cookbook](https://learning.oreilly.com/library/view/python-3-text/9781782167853/) by [Jacob Perkins](https://learning.oreilly.com/search/?query=author%3A%22Jacob%20Perkins%22&sort=relevance&highlight=true)
* **Good list of pros and cons:** <https://medium.com/@bayramorkunor/naive-bayes-algorithm-a-simple-and-powerful-classification-method-7f58a1eb975c>
* <https://www.geeksforgeeks.org/naive-bayes-classifiers/>