**What is naïve bayes?**

Naïve bayes is a classification technique based on bayes theorem with an assumption that the presence of a particular feature in a class is unrelated to the presence of other feature.

Consider a document D, whose class is given by C. In the case of email spam filtering there are two classes C = S (spam) and C = H (ham). We classify D as the class which has the highest posterior probability P(C|D), which can be re-expressed using Bayes’ Theorem:



Actually, Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem**. The algorithm that we used in our model is **Bernoulli Document model.**

**I**n the Bernoulli model a document is represented by a feature vector with binary elements taking value 1 if the corresponding word is present in the document and 0 if the word is not present.

Let P(wt |C) be the probability of word wt occurring in a document of class C. The probability of wt not occurring in a document of this class is given by (1−P(wt |C)). If we make the naive Bayes assumption, that the probability of each word occurring in the document is independent of the occurrences of the other words, then we can write the document likelihood P(Di | C) in terms of the individual word likelihoods P(wt |C):

****

**How does this work?**

Same as in ppt.

**Benefits:**

* Naive Bayes model is very effective for text classification with large datasets.
* It is easy and fast to predict class of test data set. It also performs well in multi class prediction
* Naïve bayes performs better compare to other models like linear regression and logistic regression when the assumption of independence holds.
* It performs well for categorical input variables compared to numerical variables.
* It requires less training data than other models.