Text classification using Naïve Bayes Models

Introduction

The Naïve Bayes Models is a simple probability distribution falling from Bayes formula that allows us among many things to assign a probability to a hypothesis or cause knowing the set of observable effects. The goal of this small paper is to apply that idea to text and classify in different category.

Theory of Naïve Bayes Models

According to Bayes formula, we know that:

Where is an n-dimensional vector such that with another vector. represents the hypothesis that is being tested knowing the observable effects. Also, if we recall the following:

And elaborate the expression of knowing that  **:**

Hence, we can conclude a final expression for :

If we replace the expression in (2) in (0), we get:

Applying Bayes formula to we can derive the following:

Or

Assuming

=

Which mean that the causes behind the observable effects form a partition of the universe Also, is either 0 or 1 depending on whether Cause can lead to the set of effects. We come up with the following:

With the version space of and , the final expression is the following:

Considering that does not depend on any hypothesis we can drop it:

Since we assumed uniformity of the priors, we can also drop the is the same for all the hypotheses. Hence,

Because the function log (x) is increasing for any value of R, finding the Cause that maximizes (4) is the same as finding the cause that maximizes the log []. Hence,

Or since the log of the probability varies with the probability we can drop it:

Application to text classification

Below is the statement of the problem: “Given a document that can fit in a finite number of exclusive categories, can we accurately assign that document to the proper category?”

The approach is to look at each category as a cause behind a certain document and the set of word in the document as the observable effects. Assuming that we will try to compute the expression in (3) for every single category. We assume the uniformity of the priors which means that all the elements of **Cause** are equally likely to produce the set of observables. We assume that the words occur independently from each other. Hence the presence of one word in a document does not at all impact the presence of another (which is not true obviously). For the causes, we assume that no two hypotheses can be true at the same time which is also not true since we can have articles that are about politics and sport at the same time for example.