Penn CIS 5300- Speech and Language Processing - Chapter 4 Notes

Jonathon Delemos - Dr. Chris Callison Burch

June 12, 2025

This course provides an overview of the field of natural language processing. The goal of the field is to build technologies that will allow machines to understand human languages. Applications include machine translation, automatic summarization, question answering systems, and dialog systems. NLP is used in technologies like Amazon Alexa and Google Translate.

0.1 4.0 - Naive Bayes, Text Classification, and Sentiment

Classification lies at the heart of all intelligence. Deciding how to interpret symbols, words, and actions is an important step in our decision making process. In this chapter, we will discuss the Naives Bayes algorithm and how to apply it to *text categorization*. This involves determining sentinment, the positive or negative orientation of a remark. The most common form of achieving text classification in language processing is through **supervised machine learning**. This is where we have a data set, each bit associated with some correct output. The goal of the algorithm is to learn to map the new observation to the correct output.

- Y = (y_1, y_2, y_3, y_3, etc - Set of correct Inputs)
- $y \in$ Y
- c = *Class*
- d = *Document* - Think of this as our x input

We represent a document as if it were a bag of words. We only keep tracks of the frequency of the words.
Naives Bayes is a probablistic classifier, meaning that for a document d, out of all classes $c \in$ C.
In Eq. 4.1, we use the *hat* notation $\hat{c}$ to represent our estimate of the correct class. We also use the arg max operator to mean an operation that selects the argument (in this case, the class c) that maximizes a function (in this case, the probability $P(c \mid d)$):

$$\hat{c} = \arg\max_{c \in C} P(c \mid d)$$

Bayes' Rule:

$$P(x \mid y) = \frac{P(y \mid x) \cdot P(x)}{P(y)}$$

Then we substitute using bayes rule.

$$\hat{c} = \arg\max_{c \in C} \frac{P(d \mid c) \cdot P(c)}{P(d)}$$

We call Naive Bayes a generative model because we can infer an answer based off the given information. Without a loss of generality, we can represent a document d as a set of features $f_1, f_2, ...f_n$:

$$\hat{c} = \arg\max_{c \in C} \frac{P(f_1, f_2, ...f_n \mid c) \cdot P(c)}{P(d)}$$

Naive Bayes Assumption : this is conditional independence assumption that the probabilities $P(f_i \mid c)$ are independent given the class c and hence can be naively multiplied.

$$\hat{c} = \arg\max_{c \in C} P(c) \cdot \prod_{f \in F} P(f \mid c)$$

0.2 Questions?

This chapter is also going really slow.. would be a lot easier with a teacher.
In the Naive Bayes Assumption, are we creating a product vector out of the prediction given the current word? I'm a little confused.

0.3 Summary

In this chapter, we discusssed Naive Bayes theorem for classification and applied it to text categorization and sentiment analysis.

- Sentiment analysis classifies a text as reflecting the positive or negative orientation that a writer expresses.