Worksheet 1 – Probability and the Naïve Bayes Classifier

Theory

Review the videos from week 1. Answer the following questions based on those lectures

1. How is probability theory and machine learning theory related?

Answer: Probability has to do with measuring the likelihood of events occurring or not occurring. One of the two methods is to make predictions about future events based on what has happened in the past, based on data gathered about past events and their outcomes. And the first method that using past event data to develop models to help make predictions about the future is the dominant approach in machine learning.

1. What is the difference between prior probability and conditional probability? Explain your answer using two random variables X and Y.

Answer: Prior probability of a specific outcome is the number of times we observed this outcome divided by the total number of observations. And conditional probability of an outcome with a condition is the number of times that this exact event occurs divided by the number of times we observe the condition event occurs. As is seen in the formula below.

1. What is meant by event independence? What is P(X | Y) if X and Y are independent?

Answer: If a random variable X has no effect on another random variable Y, then X and Y are said to be independent. Given the condition that X and Y are independent, then the joint probability of X and Y is the product of their probabilities.

1. What is meant by conditional independence? What is P(X, Y | Z) if X and Y are conditionally independent with Z?

Answer: If our random variables are conditionally independent given a third random variable Z, conditional probability of X and Y, given Z, is just the product of the conditional probabilities of X given Z and Y given Z.

1. What is meant by the term “overfitting”?

Answer: Overfitting means that the model is overly biased towards the training distribution and does not perform well for classification or predictions of unseen data from the real-world.

1. Why is the Naïve Bayes model considered naïve? Does this matter in practice and why?

Answer: We call this naïve because of the simplifying assumptions of conditional independence between features. While this is not always strictly valid, Naïve Bayes performs very well in practice.

1. What does the “accuracy metric” measure in binomial classification problems?

Answer: These metric measure different aspects of the classifier performance. A true positive means a correct classification of the target feature as true. A false positive means an incorrect classification of the target feature as true. A true negative means a correct classification of the target feature as false. A false negative means an incorrect classification of the target feature as false.

1. What is the difference between “term-frequency” and “intra-document frequency”?

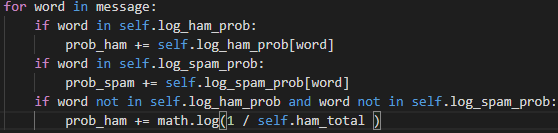
Answer: “Term-frequency” just means the idea that the term may occur more than once within the message. The higher the “term-frequency” is, the more impact it has on the result. “Intra-document” is the method to evaluate the importance of a term in a message. The higher idf, the less impact on the result.

Practice

Follow the tutorial videos from week 1 and carry out the following steps

1. Make sure you have the Anaconda environment installed from the first lab.
2. Download the code archive and extract the file from the week 1 learning materials. Make sure that you can run the examples as described in the tutorial videos
3. The basic Bag-of-Words implementation (**bow\_filter.py**) only achieves an accuracy score of 64% on average. Suggest and test changes to this basic implementation that could improve the accuracy.

At the end of the predict function, we could add a condition to judge if the word doesn’t appear at any of the previous training data. If that’s true, we consider it as a non-spam word. As is shown in the picture below.



1. Extend your improved Bag-of-Words implementation (from the previous step) to classify the multinomial **Iris** dataset (which is widely downloadable and is also available as part of sklearn). Note that you will have three classes to deal with and not just two. What accuracy can you achieve?

I achieved the accuracy of 44.4%.

1. Use the built-in sklean MultinomialNB model to classify the **Iris** dataset. Does it do better than the one you have implemented in the previous step?

The accuracy using sklearn model is 90%. A lot better than before.

Reflections

What is your initial impression of the machine learning development in the Python language?

Easy to use and analyse large datasets. And it also has a lot of third-party libraries.