Assignment 2
Naive Bayes & Text Classification (Theory)

Kuang Yu Li, st169971@stud.uni-stuttgart.de, 3440829
Ya Jen Hsu, st169013@stud.uni-stuttgart.de, 3449448
Gabriella Ilena, st169935@stud.uni-stuttgart.de, 3440942

1 Simple Bayes

1. *Box 1 contains 8 apples and 4 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability. What is the probability of choosing an apple? If an apple is chosen, what is the probability that it came from box 1?*

$$P(Apple) = \frac{1}{2} \times \frac{8}{12} + \frac{1}{2} \times \frac{10}{12} = \frac{3}{4} = 0.75$$

$$P(Box\_1 \mid Apple) = \frac{P(Box\_1 \cap Apple)}{P(Apple)} = \frac{\frac{1}{2} \times \frac{8}{12}}{\frac{3}{4}} = \frac{4}{9} = 0.44$$

2. *The blue M&M was introduced in 1995. Before then, the color mix in a bag of plain M&Ms was: 30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan. Afterward it was: 24% Blue , 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown.*
A friend of mine has two bags of M&Ms, and he tells me that one is from 1994 and one from 1996. He won't tell me which is which, but he gives me one M&M from each bag. One is yellow and one is green. What is the probability that the yellow M&M came from the 1994 bag?

$$P(1994 \mid Yellow) = \frac{P(1994 \cap Yellow)}{P(Yellow)} = \frac{\frac{1}{2} \times \frac{20}{100}}{\frac{1}{2} \times \frac{20}{100} + \frac{1}{2} \times \frac{14}{100}} = \frac{10}{17} = 0.59$$

3 kNN for Text Classification

Research and discuss how you could use a k-nearest neighbor classifier for text classification. You should at least answer these questions:

- *How do you represent the text?*
- *What distance function do you use?*
- *What decision rule do you use?*

Provide an example for your representation of the text and how your classification decision is made based on the distance function and decision rule. Explain the advantages and disadvantages of your approach.

To implement KNN classifier in text classification, we must first represent each word/term numerically to be able to calculate the distance of words (which play the role of 'attributes') in different documents ('class'). The numerical representation is often referred to as the weight which describes the relation between each word to each document. We can then create a weight matrix of dimension NxM, where N is the number of documents and M the number of words, with certain weights as the matrix value. There are explorable options of methods to determine these weight values, but some of the most common ones are: *Binary*, *Term Frequency,* and *Term Frequency-Inverse Document Frequency (TF-IDF)* methods. In our approach, we will explore the use of the *Term Frequency* method. The *Euclidean distance* method can be used to find the distance between each document vector.

An example of how we can represent the text is given as follows. Suppose we want to classify documents based on the class labels 'Spam' and 'Not Spam'. We first identify relevant words that may indicate a certain document as spam, such as 'call' and 'free', and words that may refer to personal messages i.e. not spam, such as 'sorry', and 'good'. We will assign the number 1 to class 'Spam' and 0 to class 'Not Spam'. Here, the weights are determined using the TF method, so how many times the word appears in each document.

ID	**Class**	**'call'**	**'free'**	**'sorry'**	**'good'**
1	**0**	1	0	3	3
2	**1**	2	2	0	1
3	**0**	0	0	1	2
4	**1**	2	3	0	0

Decision rule: after calculating distance of unseen data to the training data vectors, we determine the nearest vectors to the new data point. The class that has the majority vote is the class predicted for that data point.

Suppose we have a new document 5 with a feature vector: [3, 1, 0, 1]. Using K = 2, we find that the two nearest neighbours are documents 2 and 4, which are classified as Spam. Thus, document 5 will be classified as Spam.

Reference:

1. Open Genus, *Text Classification using KNN*, https://iq.opengenus.org/text-classification-using-k-nearest-neighbors/
2. Bruno Trstenjak et al., KNN with TF-IDF Based Framework for Text Categorization, Procedia Engineering, Vol. 69, 2014. https://doi.org/10.1016/j.proeng.2014.03.129

4 kNN in High Dimensional Feature Space

Research and discuss why kNN might fail for high dimensional feature spaces.
Identify and explain one approach for solving or circumventing this problem.

There are several reasons for why kNN might fail for high dimensional feature spaces.

a. If the dimension of the feature is increased, the number of samples required will increase exponentially. This is the main cause of the "curse of dimensionality".

b. When the dimension increases, the ratio of the number of other sample points within a unit distance from a sample point will decrease, which will cause us to go further distance to find nearby values. Since the selected nearby value is further and further away from the sample, it is not so valuable.
c. kNN needs to traverse the entire samples every time it runs, exponentially increasing the number of samples would cause huge calculations.

In order to solve this problem, we need to apply dimension reduction to feature spaces to extract essential features. For example, we can use the Principal Component Analysis (PCA) algorithm to compress a dataset onto a lower-dimensional feature subspace with the goal of maintaining most of the relevant information.

Reference

1. https://en.wikipedia.org/wiki/Curse_of_dimensionality
2. https://en.wikipedia.org/wiki/Dimensionality_reduction