Exercise for Machine Learning (SS 20)

Assignment 2: Naive Bayes and Text Classification

Prof. Dr. Steffen Staab, steffen.staab@ipvs.uni-stuttgart.de
Alex Baier, alex.baier@ipvs.uni-stuttgart.de
Janik Hager, janik-manel.hager@ipvs.uni-stuttgart.de
Ramin Hedeshy, ramin.hedeshy@ipvs.uni-stuttgart.de
Analytic Computing, IPVS, University of Stuttgart

Submit your solution in Ilias as either PDF for theory assignments or Jupyter notebook for practical assignments.
Mention the names of all group members and their immatriculation numbers in the file.
Submission is possible until the following Monday, 11.05.2020, at 14:00.

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?

   Solution:

$$
\begin{aligned}
P(B=1) = P(B=2) &= \frac{1}{2} \\
P(X=A|B=1) &= \frac{8}{12} \\
P(X=A|B=2) &= \frac{10}{12} \\
P(X=A) &= P(X=A|B=1)\cdot P(B=1) + P(X=A|B=2)\cdot P(B=2) \\
&= \frac{8}{12}\cdot\frac{1}{2} + \frac{10}{12}\cdot\frac{1}{2} = \frac{18}{24} = \frac{3}{4} \\
P(B=1|X=A) &= \frac{P(X=A|B=1)\cdot P(B=1)}{P(X=A)} \\
&= \frac{\frac{8}{12}\cdot\frac{1}{2}}{\frac{3}{4}} = \frac{4}{9} \approx 0.44
\end{aligned}
$$

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?

Solution:

There are two possible combinations: Either bag 1 is from 1994 and bag 2 is from 1996 ($C = 1$) or vice versa ($C = 2$). Furthermore, we need to compute the probability that a yellow M&M is taken from a 1994 bag and a green M&M is taken from a 1996 bag ($X = (Y, G)|C = 1$) and vice versa ($X = (Y, G)|C = 2$).

$$
\begin{aligned}
P(C = 1) = P(C = 2) &= \frac{1}{2} \\
P(X = (Y, G)|C = 1) &= \frac{20}{100} \cdot \frac{20}{100} = \frac{400}{1000} \\
P(X = (Y, G)|C = 2) &= \frac{14}{100} \cdot \frac{10}{100} = \frac{140}{1000} \\
P(C = 1|X = (Y, G)) &= \frac{P(X = (Y, G)|C = 1) \cdot P(C = 1)}{P(X = (Y, G))} \\
&= \frac{P(X = (Y, G)|C = 1) \cdot P(C = 1)}{\sum_{i=1}^{2} P(X = (Y, G)|C = i) \cdot P(C = i)} \\
&= \frac{\frac{400}{1000} \cdot \frac{1}{2}}{\frac{400}{1000} \cdot \frac{1}{2} + \frac{140}{1000} \cdot \frac{1}{2}} = \frac{20}{27} \approx 0.74
\end{aligned}
$$

2 Spam Classification with Naive Bayes

Please download the Jupyter notebook *assignment2.ipynb* and the dataset *spamham.txt*. Follow the instructions in the Jupyter notebook.

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.

Solution:

We treat each document as a set of words. The frequency of words is not relevant. To measure the difference or similarity between two documents is now a question of

measuring the similarity between two sets. A common measure for set similarity is the Jaccard similarity J defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Given the following labeled example document corpus for sentiment classification

- d_1 = (the, changes, are, the, worst) ⇒ negative
- d_2 = (i, like, the, changes) ⇒ positive
- d_3 = (i, do, not, like, any, of, that) ⇒ negative

we can construct the following sets:

- d_1 = {the, changes, are, worst}
- d_2 = {i, like, the, changes}
- d_3 = {i, do, not, like, any, of, that}

Given a new document d_4 = {i, do, like, the, changes} with $k = 3$, we first compute the pairwise similarity between d_4 and each training document:

- $J(d_4, d_1) = \frac{2}{7}$
- $J(d_4, d_2) = \frac{4}{5}$
- $J(d_4, d_3) = \frac{3}{9}$

As decision rule, we can weight the votes of each neighbor according to their similarity. Note that we compute the similarity and not the distance. If distance between two documents is low, the similarity will be high. Therefore, we can just sum up the similarities per class to compute the respective votes. d_1 and d_3 vote for class "negative" with a total weight of $\frac{18+21}{63} \approx 0.62$. d_2 votes for class "positive" with a total weight of $\frac{4}{5} = 0.8$. Consequently, we classifiy the new document d_4 as positive sentiment, which is intuitively also correct.
This proposed approach is very limited. We ignore a large amount of valuable of knowledge with our document representation, such as the word frequency, the frequency of words per document and the order of words. From information retrieval, we know more effective methods for representing documents. A traditional method from information retrieval is TF-IDF (term frequency-inverse document frequency). It includes information about the frequency of words and their rarity over all documents. For example, words that are frequent in all documents are typically not useful for classification, while words that occur only in very few documents are important for classification. However, TF-IDF also ignores the order of words.

4 kNN in High-Dimensional Feature Spaces

For all students other than B.Sc. Data Science:
Research and discuss why kNN might fail for high dimensional feature spaces. Identify and explain one approach for solving or circumventing this problem.

Solution:
Read Chapter 2.5 in "Elements of Statistical Learning" by Hastie. In high-dimensional feature spaces, common distance measures, e.g. Euclidean distance, lose meaning, since most points have similar distances to each other. We can circumvent this issue by reducing the dimensionality of our feature space with various feature reduction and selection techniques, such as:

- Principal Component Analysis (PCA)
- L1-regularization
- Neural word embeddings

We will discuss some of these techniques in future lectures and assignments.