Report of the 3rd exercise sheet

Marco Adamczyk (MA) Till Brinkmann (TB)

Submission by 06th December 2019

Tutorial: Tuesday, Tutor: Riza Velioglu

1 Introduction

This is an template for a report. You may use the outline or change it freely. There are no directives.

1.1 Datasets

"dataset1"

2 Methods/Models

Which classifiers are used? How do they classify/differ?

2.1 k-nearest-Neighbor classifier

The knn classifier represent the data on a n-dimational space and in this space the distance of dataset element to the new element is calculated by a specific metric. In the given case, the p-Norm with p = 1 is used, because the dataset vector entries are zero or one, so the relative distance will stay the same. After the distance calculation, the k closest dataset points to the new element are chosen. Now the new data has to be classified by those selected dataset points. There are multiple approaches. The naive apporach would be to select the class that appears on the selected data most often. More ideas are disscussed below. MA

2.2 Classification tree

The classfier structures the trainingsdata as a tree, so the MA

3 Bayes Experiment

4 kNN Experiment

The knn-algorithm was implemented as a class in python. A method was saving the training set and its outputs in two variables. With the second method "predict(element,k)" the estimated class for a new element was returned. In short, it calculated the distance in a help method for every data in the trainings set. Then it iterates through that list and selects the k-nearest elements. Form the selected elements the class is chosen that appears the most in these elements. MA

4.1 Data

The two given datasets were used to test the classifier. The datasets were split in two parts. The testset contains 20 elements from the data set. The other data is used as the trainingset. Every data from the testset will be classified for a k = 1..10. As seen in the graphics bellow, there is not significant difference between the balanced and unbalanced dataset. The kNN only classifies with selected element not the whole dataset, so it is only the dependent how the classes in the dataset are distributed. MA

4.2 Results

Figure 1: Results

You can refer to Figure 1. You can also refer to Table 6.2.

Table 1: An Example of a Table

Data	Method 1	Method 2	Method 3
data 1	0.54	0.6	0.98
data 2	0.74	0.54	0.48
data 3	0.82	0.71	0.67

FGM

5 Discussion

As seen in the results some classifiers are more suited for specific tasks.

TB/MA

6 Further Features

6.1 Data Features

Treegramms TB

6.2 kNN Features

The kNNs selection of the class only the depends on the number of occurences of a certain class. It could be reasonable to include the distance, because the following problem could occur: Assume k = 4. 3 elements are from the class 1 and 1 element is from the class 0. Normally, the data would be classified as 1, but if the distances of the class 1 elements are far away and the element with class 0 is a lot closer it could be more reasonable to classify the new data as class 0. To include the distance as a new parameter to the calculation, the average distance from the new data and a specific class could be calculated and the compared. With this change, we get the following result for the same setup presented in the results above:

Figure 2: Results

You can refer to Figure 1.

Table 2: An Example of a Table

Data	Method 1	Method 2	Method 3
data 1	0.54	0.6	0.98
data 2	0.74	0.54	0.48
data 3	0.82	0.71	0.67

MA