Report of the 3rd exercise sheet

Marco Adamczyk (MA) Till Brinkmann (TB)

Submission by 06th December 2019

Tutorial: Tuesday, Tutor: Riza Velioglu

1 Introduction

1.1 Datasets

Both datasets have strings of the characters [A..Z] as data, with an average length of 6 characters.

- Dataset1 consists of 5493 datapoints. There are 2 different classes. Class 0 appears 4274 times in the set while the other 1219 datapoints are in class 1.
- For Dataset2 there are 2002 datapoints in class 0, 2000 in class 1 and 1999 in the third class.

TB

1.2 Feature Extraction

As suggested, we used bigram occurences as our features. This makes the input for the classifiers a vector containing ones or zeros with the size $26^2 = 676$. TB

2 Methods/Models

2.1 Naive Bayes Classifier

As the first model we chose a Naive Bayes classifier. This model uses a probability given by a simplified Bayes formula (with P(X) always 1) for determining the most likely class for an input X:

$$h(x) = \arg\max_{c \in Y} P(X = x|Y = c) * P(Y = c)$$

We defined $P(X = x|Y = c) = 1 - \frac{|x - x_{c,avg}|}{k}$ where $x, x_{c,avg} \in \mathbb{R}^k, x_{c,avg} = (1\ if\ \frac{1}{|X|} * \sum_{\{x_i \in X|y_i = c\}} x_i > 0.5,\ 0\ otherwise)_k$. This can be described as one minus the normalized distance between x and a median of all $\{x_n \in X|y_n = c\}$. TB

3 Experiments

Which tests were done in the experiments? What was implemented? What measurement are used in the results?

3.1 Data

Which data are used? What are their characteristics? SGM

3.2 Results

4 Discussion

Short summary and future work.

SGM/TGM

Figure 1: Results

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

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