SAKI SS 2021 Homework 1

Author: Stefan Fischer

Program code: https://github.com/StefanFischer/SAKI-Project1

Summary

With the rapidly rising amount of data the possibility of information retrieval has also increased dramatically. Data Mining and Machine Learning methods are introduced in almost every sector of industry to get valuable insights of the market and use those for optimization. In the financial sector transaction data can be analyzed regarding their purpose and for example frauds can be identified. Furthermore, applications can be developed to help costumers keeping track of their finances. The advantage is the automatic classification of transactions, while manually categorizing those is time consuming.

In the following a pipeline for transaction classification into 6 different categories was designed. The data is consisting of 208 real-world transactions of an employee of the "Adorsys GmbH & Co. KG" and was then classified into the categories: income, living (rent, additional flat expenses, ...), private (cash, deposit, donation, presents), standard of living (food, health, children, ...), leisure (hobby, sport, vacation, shopping, ...) and finance (credit, bank costs, insurances, savings). A transaction consists of information characteristic of the ordering bank account, ordering date, valuta date, booking text, purpose, payor, account number, BLZ, payment amount and the currency. All transactions were already labeled.

For classification a gaussian naïve bayes was used with bag of word features, which are a standard feature for transaction classification. To keep the simplicity of the classifier, the characteristic payment was dropped, as it does not qualify as a bag of word feature. Furthermore, the payment amount has only an obvious link between label income and the payment amount, which has to be positive for income. For all other categories, the distribution of the characteristic amount is quite similar. Furthermore, both date characteristic were copies of each other, therefore only one of them was kept. Also the currency characteristic was dropped, as the values were for all samples identically 'eur'. Besides this, removing this characteristic led to 'eur' being a highly discriminative word. With the remaining characteristics bag of word features were calculated, after the text was concatenated. At first, punctuations are removed and all chars are interpreted as lower case. Then all words/numbers and the amount of their occurrence are counted for each transaction sample, resulting in a feature vector of around 500 words and numbers.

Additional feature selection by highest Chi-2 score did also not increase the performance, which can be explained by the relatively small word count. Besides this, also a principal component analysis dimensionality reduction on the bag of word features did not improve the results.

With the proposed methods a F1-score of about 90% was achieved, depending strongly on the assigned training data.

Evaluation

Metric: For evaluation of a classifier or regressor in machine learning different metrics are widely used to evaluate and compare the performance of systems. Here we used accuracy, precision, recall, F1-score and the confusion matrix.

1. **Accuracy:** Metric which is the fraction of predictions our model got right and does not tell anything about the frequency of the classes. This means that when working on a class-imbalanced data set the accuracy metric can be misleading.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Recall:** Recall is proportion of actual positives identified correctly and does also accounts for class-imbalanced data. For some applications the recall has to be very high, as for example for fraud detection or clinical diagnosis, where the cost for false negatives is high.

$$Recall = \frac{TP}{TP + FN}$$

3. **Precision:** Besides Recall, precision can also be used for class-imbalanced data. Precision is the proportion of positives identifications which were actually correct. This is a suiting measure if the costs for false positives is high for example in e-mail spam filtering.

$$Precision = \frac{TP}{TP + FP}$$

4. **F1-score:** The F1-score is a widely used metric and the harmonic mean of recall and precision. Furthermore precision and recall can be weighted, such that one them is more valued.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Recall, Precision and F1-score are only applicable for binary classification. Extensions for multi-class classification different weighted averages of those metrics are possible as the macro or micro weighting.

5. **Confusion-Matrix:** The confusion matrix visualises the classification result for each sample. Entry $[i, j]$ in a confusion matrix is the number of observations actually in group i but predicted to be in group j. On the diagonal are then the correctly classified samples.

Results and Discussion: The gaussian naïve bayes classifier was trained on 75% of the data set resulting in 156 samples, which were randomly selected. For evaluation the remaining 0.25% (53 samples) were used to compute weighted accuracy, recall, precision and F1-score. For the classes income, leisure and standard of living, the recall was 1.0, which means that all samples which are labeled as those, were also assigned to them. Furthermore, the precision for finance, income, living and private was also 1.0. This results for the class income for a perfect F1-score of 1.0. All other classes had an F1-score over 0.83. In total a weighted F1-score of 0.92 was achieved.

Using a feature selection of the 15 words, having the highest Chi-2 score, insights in the decision making was generated. The four highest scoring words are “kg”, “gehalt”, “adorsys” and “co”, which are all related to the class income and only occur in transactions of the income class. Another high ranking word is “gmbh”, which has a lower chi-2 score as it occurs for multiple classes, besides income from “Adorsys GmbH & Co. KG”. The word “eur” was only related to the classes leisure and private, after the whole characteristic currency was dropped. Another indicator for class private was the word “bargeld”, which was only present for this class.

For samples of the classes finance, living, and private false predictions were computed. The false predicted samples of those three classes were assigned to the highest frequent occuring classes leisure and standard of living, which can be explained by the higher prior probability of those two classes. Only the two high frequent classes have non-perfect precision scores, but both also a recall of 1.0.

The problem which needs to be mentioned here is that the classifier is trained on keywords. Those words are depending strongly on the person holding the bank account, as it is taking keywords like the name of the company the person works for. This will only work if the model is trained by data, which is actually from the same person the model is applied to. Furthermore, the classifier needs repetitions of those keywords in the input data to produce good outcomes. Problems can occur, if the person will change the job or if the employer is the supermarket the person also uses for daily grocery shopping.

Conclusion: This task showed how well a simple naïve bayes classifier can perform with a small training set and a simple feature like the bag of words. On the other side it can not be used for transaction classification in general as the keywords are different for each person. Thus, for a real-world market application, other features should be defined that find general trends in the data or a training phase for every single user has to be done.

Another suggestion is to use the multinomial naïve bayes classifier, as it is suitable for discrete features as the word counts.

Screenshots

	precision	recall	f1-score	support
finance	0.90	1.00	0.95	9
income	1.00	1.00	1.00	7
leisure	0.90	1.00	0.95	19
living	1.00	0.80	0.89	5
private	1.00	0.67	0.80	3
standardOfLiving	0.89	0.80	0.84	10
accuracy			0.92	53
macro avg	0.95	0.88	0.90	53
weighted avg	0.93	0.92	0.92	53

Classification Score