Data Mining: Practical Assignment #7

Due on Thu & Fri, July 11-12 2019,

Task 1

Course Evaluation ≈ 10-15 min.
For Anton's group, this part will be swapped with the lab demo part.

Task 2

Text Categorization & sentiment analysis, ≈ 120 min. Similar to the DAMI 4 tutorial, we are going to classify today real text data, the newsgroup 20 corpus. As the name suggests, the dataset comprises 20 news categories, e.g. science, sports, etc. Use the `texprocessing_basic.ipynb` file and answer the following questions:

takes a collection of text and creates a matrix of tokens and the count of each token in the text

- What is the purpose of the `countVectorizer`?
- Why it is sensible to convert the data using the *td-idf* measure?

words that are in the text more often get weighted less than words that are in it less often. simply because words like 'it' usually don't have a greater meaning.

- Do you spot any pecularities in the classifier evaluation (the classification report)?

nö

- **optional**: if you used more than 1 classifier, which one produced the best performance?

As you are now familiar with the structure and function calls of the `scikit learn` library, choose a classfier and perform experimental trials. Check out especially the arguments for the `countVectorizer` function. As classifier you can use `logistic regression` or `random forests`. A brief description is given below. Baseline: your classifier should achieve around 70%-80% accuracy.

Another problem we address today is called sentiment analysis. In the corresponding notebook `sentiment_analysis.ipynb` a movie review dataset will be downloaded (just run the cell), which categorizes the reviews into good or bad. Use the logistic regression classifier and create a little test set (an initial example is given). What are your results?

Task 3

Tutorial Topics Summary + Lab Demo, ≈ 45 min.
Reminder
1. Klausur: 23.07.2019, 9:30-11:30, ESA A, 2. Klausur: 13.09.2019, 9:30-11:30, ESA C

Conceptual Explanations

Logistic regression

Linear regression has continuous outputs (e.g. test scores 0-100) and, thus, cannot be used for classification as classes are categorical/nominal (e.g. 0=No, 1=Yes). Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function (extension to multiclass: multinomial logistic regression). The coefficients of the independent variables for linear regression are learnt by minimizing the error of squared distances, while for logistic regression the maximum likelihood estimation (MLE) is used.

Good source: `https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python`

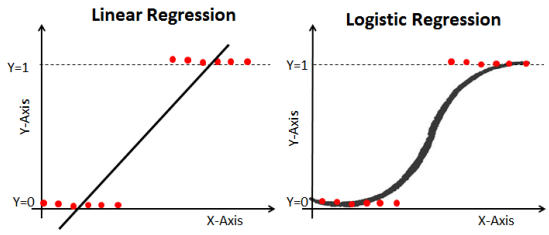

Figure 1: Linear vs. logistic regression.

Random forests

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. It relates to *bagging* as the individual trees sample the data randomly (with replacement), therefore different trees can be generated, forming eventually a forest.
Good source: `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`