Statistical Machine Learning – Week 11

Sascha Jecklin

HS 2022/23

1 Task

In this lab we will implement a spam detector using a Random Forest Classifier. Our aim is to classify an emal text as spam or ham based on some features which we extract from the mail content. In a first step, the text has to be preprocessed which includes stop words removal, tokenization, stemming and counting the word occurence. As a result of this step we get an 1899-dimensional feature vector for each mail text.

For training and testing we use 5000 mails which are already preprocessed (i.e. a dataset of 5000 vectors of dimension 1899). At the end you an apply your Random Forest Classifier to classify your own mail as spam or ham.

- Take a look at `week11_exercise.py` and get an overview of the provided functions.
- Use sklearn's `RandomForestClassifier` and fit it to the training data. Calculate the accuracy based on the test set.
- Implement your own `MyRandomForestClassifier` class using bootstrapping and decision tree classifiers. Calculate and print the accuracy based on the test dataset.
  Hints:
  - Use sklearn's `DecisionTreeClassifier`.
  - For now, ignore the random selection of predictors which is normally done by a Random Forest Classifier.
- Test the classifier with some sample mails by calling the function `classify_mail`. You can use emails from the SPAM/HAM datest on Kaggle. You can copy complete mails including the HTML syntax.

Comments

- Take a look at sklearn's dokumentation about the RandomForestClassifier
- For the implementation of your Random Forest Classifier use sklearn's DecisionTreeClassifier.
- You might have to install the natural language toolkit: `pip3 install nltk --user`