Lab Course Machine Learning
Exercise 9

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 27.01.2019 LearnWeb 3112

January 21, 2019

1 Exercise Sheet 9

Datasets

- 1. Sparse dataset :

  (a) w8a dataset D1:
  `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#w8a`

- 2. UCI Dataset

  (a) SMS Spam D2::
  `https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection`

  (b) Spambase D3:
  `https://archive.ics.uci.edu/ml/datasets/Spambase`

Exercise 1: A spam filter using SVM(16 Points)

Part A: (8 Points): Build a spam filter using a pre-processed dataset
A spam filter classify an email to be Ham or Spam, using the content of an email as features. You have to use dataset D3 for this task. Build a basic spam filter using SVM. You have to use libsvm `https://github.com/cjlin1/libsvm/tree/master/python`. libsvm accepts data in a libsvm format. Each data row in a libsvm format is given as

$< label > < index1 >:< value1 > < index2 >:< value2 > \ldots$

Convert dataset D_3 into a libsvm format. Follow the readme document given on the libsvm link to see how you can use it to solve your problem. You have to learn a spam classifier on train part of the dataset and evaluate it on test dataset. Also optimize the hyper parameter i.e. value of C. [hint: when choosing the range of hyperparameter its always useful to check a diverse range i.e. C ={1,2,3,4} is not a good range to check for optimal value, you might want to check a broader range going from 0.1 to 100 etc.]. Present your results in form of graphs and tables, listing details. You have to choose a quality criterion according to the given problem i.e. classification. **[Note:]** If you are not able to use libsvm you can replace it with scikit learn. But you have to convert your data into libsvm format.

Part B: (8 Points): Pre-processed a dataset and learn SVM

The dataset D_2 is not preprocessed. It consists of label[ham or spam] and content of sms text. Your task in this part is to pre-process this data into a processable format. Using OneHotEnconding might not help, therefore you have to use other means of converting text data into features. You can look at scikit-learn text feature extraction utilities i.e. TFIDF or count. You might also want to get rid of the stop words i.e. This, the, is, a etc, which appear in almost all the documents. After preprocessing you have to use SVM implementation provided by scikit-learn. Here you will experiment with different hyperparameters and two kernels (linear and RBF). As usual you will perform 5-fold cross validation and present the score using plots and tables. You might also want to look at sklearn.pipeline.Pipeline utility to streamline your workflow.

Exercise 2: Compare SVM based spam filter with another model (4 Points)

You have to compare results obtained in one of the task above with another model of your choice (decision trees or logistic regression etc). Optimize the hyperparameters and perform 5-fold cross validation. You can use scikit-learn implementation. Compare the results and accuracy. Finally conclude your findings.

1.1 ANNEX

- SVM help: `https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machi`
- sklearn.model_selection, sklearn.metrics, sklearn.linear_model, sklearn.preprocessing
- Scikit Learn User Guide http://scikit-learn.org/stable/user_guide.html
- You can use matplotlib for plotting.
- sklearn.metrics http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics