Applied Machine Learning
Exercise 9

Prof. Dr. Amr Alanwar

December 2024

Datasets

1. **Sparse dataset in libsvm format:**
   - a9a D1
2. **UCI dataset:**
   - SMS Spam D2
   - Spambase D3

Exercise 1: A Spam Filter Using SVM (8 Points)

Part A: (4 Points)

Build a spam filter using a pre-processed dataset. A spam filter classifies an email as Ham or Spam, using the content of an email as features. Use dataset D3 for this task. Build a basic spam filter using SVM. Use libsvm, which accepts data in a specific format:

```
<label> <index1>:<value1> <index2>:<value2> ...
```

Convert dataset D3 into the libsvm format. Follow the README document provided at the libsvm link for guidance. Train the spam classifier on the training part of the dataset and evaluate it on the test dataset. Additionally, optimize the hyperparameter C.

Hint: Choose a diverse range for the hyperparameter, e.g., $C = \{0.1, 1, 10, 100\}$, rather than a narrow range ($C = \{1, 2, 3, 4\}$). Present results in the form of graphs and tables with detailed explanations. Choose a suitable quality criterion (e.g., classification accuracy).

Note: If you are unable to use libsvm, you may use `scikit-learn`. However, you still need to convert the data to the libsvm format.

Part B: (4 Points)

Pre-process dataset D2, which consists of labels (Ham or Spam) and SMS text content. Transform the text data into features using methods like `TFIDF` or `Count Vectorization` from `scikit-learn`. Avoid using `OneHotEncoding`, as it may not be suitable. Remove stop words (e.g., "this," "the," "is," etc.) for better results.

After preprocessing, use `scikit-learn`'s SVM implementation to train and evaluate the model. Experiment with different hyperparameters and kernels (linear and RBF). Perform 5-fold cross-validation and present results as plots and tables. You may use `sklearn.pipeline.Pipeline` to streamline the workflow.

Exercise 2: Comparison with Another Model (2 Points)

Compare the SVM-based spam filter results from one of the tasks above with another model (e.g., decision trees or logistic regression). Optimize the hyperparameters and perform 5-fold cross-validation. Compare the results and accuracy, and summarize your findings in a conclusion.