IITM-CS4011 : Principles of Machine Learning — Release date: 12 Sept, 2017, 11:55 pm

Programming Assignment #2 — Deadline: 10 Oct, 2017, 11:55 pm

- The goal of this assignment is to experiment with different feature extraction techniques, SVM kernels and Decision Trees.
- This is an individual assignment. Collaborations and discussions with others are strictly prohibited.
- You may use Python (recommended), R or Matlab for your implementation. If you are using any other languages, please contact the TAs before you proceed.
- You have to turn in the well documented code along with a detailed report of the results of the experiment electronically in Moodle. Typeset your report in LaTeX.
- Be precise for your explanations in the report. Unnecessary verbosity will be penalized.
- You have to check the Moodle discussion forum regularly for updates regarding the assignment.
- **Follow the submission instructions** given on last page, strict **penalties** on not following.
- **Please start early.**

Feature Extraction

1. You have been provided with a 3-dimensional dataset (DS3) which contains 2 classes.
    - Perform PCA on the dataset and extract 1 feature. Use the data in this projected space to train linear regression with indicator random variables.
    - Use the learnt model to classify the test instances. Report per-class precision, recall and f-measure.
    - Report the 3-D plot of the dataset and the plot of the dataset in the projected space along with the classifier boundary.
2. Use the same DS3 dataset for this question.
    - Perform LDA on the dataset and project the dataset to the derived feature space. Report per-class precision, recall and f-measure.
    - Report the 3-D plot of the dataset and the plot of the dataset in the projected space along with the classifier boundary.
    - What do you infer from these two experiments? Which feature extraction technique performs better for this scenario? Why?

3. We have discussed about Linear Discriminant Analysis(LDA) in the class. We will see how different variants of this technique works. For this experiment, you have to use Iris Dataset (http://archive.ics.uci.edu/ml/datasets/Iris).

   - Use only petal width and petal length features and perform LDA. Visualize the boundaries learnt.
   - Read about Quadratic Discriminant Analysis (QDA) and Regularized Discriminant Analysis (RDA) from the text book [1]. Do QDA and RDA on the same data set and visualize the boundaries.

SVM

4. You have been provided with training instances for an image classification problem DS2 (Same as given for Question 7 in PA-1). You have to train an SVM to classify the test images into either of the following four categories: coast, forest, inside-city, mountain.

   Use the training data to build classification models using the following kernels.

   1. Linear kernel
   2. Polynomial kernel
   3. Gaussian kernel
   4. Sigmoid kernel

   Come up with the kernel parameters for the various models. You can use a fraction of data supplied to do a n-fold cross validation to find the best model parameters.

   Important Notes:

   1. **Feature Extraction:** Follow the instructions given in *instructions* file in the DS2 folder. Please do not jumble up the r-g-b sequence while building the feature vectors.
   2. You have to use libsvm in matlab, or libsvm package in python.
   3. If you are using Python, store the model as '*svm_modelx.model*', where x is the index of the corresponding model given above, e.g. '*svm_model1.model*' for linear. You can use libsvm's `svm_save_model` method.
   4. If you are using MATLAB:
      - Name the models as 'modelx', where x is the index of the corresponding model given above, e.g., 'model1' for linear
      - Put these 4 models in a single .mat file, name it as '*svm_model.mat*'.

[1] Section 4.3 of Elements of Statistical Learning

Decision Trees

5. You need to use Weka for this question. We will use Mushroom dataset from UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Mushroom). This is a 2-class problem with 8124 instances. Use the last 1124 instances as test data and the rest as training data.

    - Convert the data into ARFF format.
    - Run J48 Decision Tree algorithm from Weka. Report precision, recall and f1-measure.
    - What is the effect of MinNumObj on the performance? What happens when you do reducedErrorPruning?
    - What are the important features in deciding whether a mushroom is edible or not?
    - Turn in the Decision Tree learnt by the model (the decision tree with the best performance).

Using external libraries

- Use LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) for SVM.
- If you are using Python, then you can use PCA, LDA, QDA, and libsvm in sklearn package. You can use pickle to save sklearn's libsvm model.

Submission Instructions

Submit a single tar/zip file containing the following files in the specified directory structure. Use the following naming convention: ’rollno_PA2.tar.gz’.

Note: ’*run.py*’ script in each code folder should run everything asked and display the results.

rollno_PA2

- **Dataset**
  - **DS2**
  - **DS3**
  - **iris**
  - **mushroom**
- **Code**
  - **q1**
    - *run.py*
    - other code, model and result files
  - ⋮
  - **q5**
    - *run.py*
    - other code, model and result files
- **Rollno-report.pdf**
- **README**