IITM-CS5011 : Introduction to Machine Learning — Given on: Sept 28, 1am

Programming Assignment #2 — Due on : Oct 7 , 11:55pm

- The goal of this assignment is to experiment with feature extraction methods and classification methods.
- This is an individual assignment. Collaborations and discussions with others are strictly prohibited.
- You may use Matlab, Octave or Python 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.

---

1. You have been provided with training instances for an image classification problem DS2 (Same as given for Question 9 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.

   Feature Extraction Use the same instructions as given for Question 9 in PA-1.

   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. You have to use libsvm in matlab or libsvm package in python.
   2. Name the models as modelx, where x is the number of the corresponding model given above, e.g., model1
   3. Put only these models in a single .mat file, name it as your roll no.mat, and submit it, e.g., CS11S016.mat (roll no, in uppercase)

4. Please do not jumble up the r-g-b sequence while building the feature vectors or the modelx while building the classifiers.

2. Implement original back-propagation algorithm. Use DS2 for training your neural network. Report per-class precision, recall and F-measure on the test data used in Question-1. Now consider the following alternate error function for training neural networks.

$$R(\theta) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}(y_{ik} - f_k(x_i))^2 + \gamma(\sum_{k}\sum_{m}\beta_{km}^2 + \sum_{m}\sum_{l}\alpha_{ml}^2)$$

where N is the number of training instances, K is the number of output features, $f_k(x)$ is the predicted output vector, y is the original output vector, α and β are the weights and γ is a regularization parameter. Derive the gradient descent update rule for this definition of R. Now train your neural network with this new error function. Report per-class precision, recall and F-measure on the same test data. What will happen when you vary the value of γ ? Vary the value of γ from 10^{-2} to 10^2 in multiples of 10 and repeat the experiment and report the results. Can you figure out the effect of γ in the results? Look at the weights learnt using the new error function. What do you infer from them?

3. 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.
1. Convert the data into ARFF format.
2. Run J48 Decision Tree algorithm from Weka. Report precision, recall and f1- measure.
What is the effect of M inN umObj on the performance? What happens when you do reducedErrorP runing?
3. What are the important features in deciding whether a mushroom is edible or not?
4. 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.

Submission Instructions

Submit a single tarball/zip file containing the following files in the specified directory structure. Use the following naming convention: 'cs5011_a2_rollno.tar.gz'.

cs5011_a2_rollno

Dataset

DS2_train.csv
DS2_test.csv

Report

rollno-report.pdf

Code

all your code files