Assignment 4 (Coding Based)
CSE471 : SMAI
Spring 2017

Submission Deadline : 11.55 PM ; 25/03/2017
Total Marks : 30

General Instructions

- Assignment can be implemented in Matlab/Python/R.
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious actions including an F grade in the course is possible.
- **Compulsary:** A single zip file should be submitted. The zip file should contain a pdf file containing all the experimental findings as instructed in each problem statement. The zip file should also contain codes for different problems. Please make a separate code file for each problem in the assignment. **These will be your final codes based on which TAs will evaluate you.** Please write your codes in proper manner so that it becomes easy for TAs to evaluate. You can use Python notebook for example. Other ways are also fine. You can also include other supplementary material like JPG files etc in the zip file.
- Include the assignment number, your name and roll number at the top-left of the first page of your submission.
- Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

1 Neural Networks

Problem 1 [10 Marks] :. Training 3-layer Neural Network.

1. Implement a simple 3-layer feed-forward neural network for Multi-class classification problem. Use sigmoid activation function. Implement back-propagation for training the parameters. **Please do not use any standard Neural Network library and write your code from scratch.**

2. Finally, train this network to learn a 3-class classifier for optical character recognition for any three digits between 0 and 9.

3. **Data :** Use any three digits between 0 and 9 from the *optdigits* data set that comes from the UCI Machine Learning Repository. Link: http://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/

4. **Pre-processing :** Digitize and down-sample images (to 8×8 or so).

5. **Classifier :** Try few configurations for a 3-Layer Neural Network with varying number of hidden units and three output units and report learned weights from back-propagation done using training data.

6. **Report :** For different choices of n_H (number of hidden nodes).

   - Draw representative neural network architecture
   - Clearly report your experiments and results including the learned weights in a table and intermediate output of hidden layers for each of the three characters after learning is over.

2 Support Vector Machine

Problem 2 [10 Marks] : . Train a SVM Classifier

Use the datasets attached on Moodle. The file name is **Data_SVM.csv**.
Report the following:

1. Plot the data.

2. Train nonlinear SVM classifier using **polynomial kernel**. Vary the values of C and d (degree of polynomial) in some range. For each combination of C and d, run 10-fold cross validation 30 times and report the average cross validation accuracy and standard deviation. Make a chart for that.

   Find the best combination of C and d. Use these parameters and train SVM using complete dataset. Plot the final classifier.

3. Train nonlinear SVM classifier for each dataset using **rbf (Gaussian) kernel**. Vary the values of C and σ (width of Gaussian) in some range. For each combination of C and σ, run 10-fold cross validation 30 times and report the average cross validation accuracy and standard deviation. Make a chart for that.

   Find the best combination of C and σ. Use these parameters and train SVM using complete dataset. Plot the final classifier.

4. Compare the performances with different kernels and comment on the quality of classifiers in each case.

3 Bayes Decision Theory

Problem 3 [10 Marks] :. Implement a **Naive Bayes Classifier** for UCI Census-Income (KDD) Data Set using only the Discrete and Categorical attributes/features. Here is the link for the dataset:

http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/

If you think that some of the real attributes are useful, you can convert them to discrete feature with appropriate binning. Use log-probabilities to avoid numerical errors.

Report the following:

1. Summary of the data.
2. Explain how you choose to handle missing entries and why?
3. Run the 10-fold cross validation 30 times and then report the average cross validation accuracy and its standard deviation.