E1 213 pattern Recognition and neural Networks
Home Work Assignment: 2
Due on May 7, 2021

In this assignment there are 4 problems. As in the previous assignment, you are given some data on which you are supposed to learn and test classifiers.

In this assignment you will explore linear models and compare different methods of learning linear classifiers and regression functions.

In two problems the data is synthetic while in the other two you have real data. For the synthetic data, the details of how the data is generated is given. For the synthetic data sets in one problem, the training and test sets are separately given. For the real data sets and one synthetic data set, you have to decide what you would use as training set and what you would use as test set. You need to clearly explain this in your report.

Each of the problems are described in detail below. For each problem, you are asked to try some algorithms. This is the minimum exploration you are required to do. You are welcome to explore further if you can think of some other interesting things to do.

You can implement the learning algorithms on any platform you want (C, C++, MATLAB, Python etc.). You are welcome to use codes that are freely available from any source. You are not required to submit any codes/implementation.

What you need to submit is a report summarizing your exploration of the data sets. For each problem, briefly describe what is done and then present all the results obtained. Discuss all points from your results that you consider interesting or worth discussing. The final submission should be in the form of a short PDF file.

The grading depends on whether or not you have done all explorations that are asked for, how you presented the results, your discussion of results and whether you have done some exploration on your own.

The problems are described below.

1. In this problem you are given three synthetic data sets. All are 2-class classification problems. In each case you are required to compare linear least squares and logistic regression for learning a linear classifier.

   In each case you are given 2000 training samples and 1000 test samples. Use different training set sizes (by randomly sampling from the given training set). Vary training set sizes as 10, 50, 100, 500, 1000. In each case assess the learnt classifier using the full test set.

   The three data sets are the following.

   a. 2D Data, features independent Gamma distributed. Class-I: Gamma(shape=0.5, scale=1); Class-II: Gamma(shape=2.0, scale=2).
   File names: *Gamma_train.txt, Gamma_test.txt*

   b. 2D data. Class-I: Uniform over $[0.5,\ 6.0] \times [0.5,\ 6.0]$; Class-II: Uniform over $[0,\ 1] \times [0,\ 1]$.
   File names: *Uniform_train.txt, Uniform_test.txt*

   c. 10D Guassian data. Both class conditional densities have I as covariance matrix. The mean vector for class-I is all-zeros and that for class-II is all-ones.
   File names: *Normal_train_10D.txt, Normal_test_10D.txt*

2. The data set for this problem is the celebrated Iris data. It is about classifying a plant species based on four measurements. The feature vector is four dimensional. There are three classes. The data file is *iris_dataset.txt*. The last entry in each row is the class label.

   This is a multiclass classification problem. You are required to compare two methods: (i). learn three linear 2-class classifiers using 'one vs rest' strategy through linear least squares; (ii). learn a 3-class linear classifier (by taking the target or prediction variable as a 3-dimensional one-hot vector) through linear least squares. In your report explain clearly and concisely what the final classifier is in each case and how the three weight vectors learned are used in the final classification in each case.

   This is a well-known standard dataset. More details of the dataset can be obtained from
   https://archive.ics.uci.edu/ml/datasets/iris

3. This is a 2-class classification problem. Each feature vector is 24 dimensional. The features represent various financial attributes of a person

and the class labels denote whether or not a person is 'good' for extending credit. In the given data all features are numeric. The data file name is *german.data-numeric*. The last entry in each line is the class label.

On this data you are required to compare the performance of linear least squares and logistic regression.

This is also a standard data set used for testing ML algorithms. You can get more details about the data set at
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

4. The last problem is a 1D regression problem. The data file is *1D_regression_data.txt*. The file contains (x_i, y_i) pairs. The x_i are essentially sampled uniformly between -6 and 6. Here, y is a cubic polynomial function of x. (The function is: $y = 0.25x^3 + 1.25x^2 - 3x - 3$). In the data, we computed y for the given x and then added zero-mean Gaussian noise. Use linear least squares to fit polynomial functions of different degrees to the data. Discuss how you can decide on the degree of the polynomial function to fit to this data.