Principles of Machine Learning(CS4011)
Programming Assignment#1

S Ram Ananth
ME15B153

August 30, 2017

1 Synthetic Data-set Generation

1.1 Goal

Generate 2-classes data with 20 features each. Each class is given by a multivariate Gaussian distribution whose covariance matrix should not be spherical in nature. Generate 2000 samples for each class and make sure there exists some overlap within the classes by making adjusting the centroids. Partition them into train set and test set by randomly picking 70 % data of each class (1400 points each) into train set and the remaining 30 % into the test set. This dataset is referred to as DS1.

1.2 Approach

We need to generate 2 different Gaussian distributions with a random centroid close to each other such that they sufficiently overlap. Also, the covariance matrix needs to be positive semi-denite and symmetric. Let A be a randomly generated matrix of order pxp. Now the matrix B generated from A as follows will be positive semi-denite and symmetric.

$$\mathrm{B} = AA^T$$

Thus, generating the covariance matrix B like this, it can be used to generate a multivariate Gaussian distribution.

1.3 Result

Thus the two distributions are generated and classes are labelled as 0 and 1 respectively. Performance reported on DS1 are on the test-set which was created randomly using 30 % of the generated data.

2 Linear Classication

2.1 Goal

We need to learn a linear classifier by using regression on indicator variable of dataset DS1 and report the best fit accuracy, precision, recall and F-measure achieved by the classifier, along with the coefficients learnt.

2.2 Approach

A linear classifier performs a regression on the data points (X) and labels (Y) . Regression tries to learn weights w to predict a label for data point x such that Mean Squared Error is minimised .

$$y_{pred} = w_0 + w^T x$$

subject to minimising,

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - y_{pred})^2$$

where m is number of training examples. Classification is done by providing a threshold(0.5 in this case) such that all predicted values more than the threshold are labelled as 1 and rest as 0.

2.3 Result

The following metrics were obtained.

Accuracy	Precision	Recall	F-Score
0.7541	0.7634	0.7366	0.7497

Coefficients are saved as .csv file for easier readability.

3 k-NN Classifier

3.1 Goal

To use k-NN (k Nearest Neighbours) to build a classier to classify the points in DS1.

3.2 Approach

k-NN is a classifier which models a function which is region-wise constant. Given a point, k-NN predicts the majority class in the neighbourhood formed by the k nearest points from the trainset. As a result, it predicts the same class for a given region as the majority class doesn't change in that neighbourhood.

One important point to be noted is choosing a suitable value for k as higher k will result in smoother decision boundaries. Here we used different values of k ranging from 3 to 10 and performance metrics were reported.

3.3 Result

Performance metrics for different k are tabulated below.

k	Accuracy	Precision	Recall	F1-Score
3	0.5575	0.5602	0.5350	0.5473
4	0.5625	0.6027	0.3666	0.4560
5	0.5808	0.5820	0.5733	0.5776
6	0.5908	0.6306	0.4383	0.5172
7	0.5991	0.6031	0.5800	0.5913
8	0.5866	0.6176	0.4550	0.5240
9	0.5825	0.5849	0.5683	0.5765
10	0.5908	0.6187	0.4733	0.5363

As we observe linear regression using indicator variables performs better than k-NN on the whole. Among these values of k best fit is for $k = 7$.

4 Data imputation

4.1 Goal

We need to perform regression on the Communities and Crime (CandC) Data Set from the UCI repository . However as it is a real life dataset it contains missing values. Goal is to perform missing value imputation before proceeding to do the regression.

4.2 Approach

The first five features are non predictive features and hence aren't required to perform regression and are removed. The missing value imputation is done in general by estimating the distribution and extrapolating it. Hence mean,mode and median can be used as they are statistical estimate of the distribution. Mode imputation is highly benefitial only for categorical and text data and since all data is numeric it won't be much use here. Among median imputattiion and mean imputation, median is more preferable as mean is very good estimate only for uniform distributions. Therefore missing values were filled using median of all other the other values in that feature.

4.3 Result

The missing values were imputed using median value of the rest of values of that feature and completed data is produced.

5 Linear Regression

5.1 Goal

The goal is to fit the CandC data using linear regression and report the residual error of the best fit achieved on test data, averaged over 5 different 80-20 splits, along with the coefficients learnt.

5.2 Approach

As seen in Section 1.1, linear regression learns weights w by minimising the residual sum of squares error. Linear models were fitted to 5 different splits and average of residual error is reported.

5.3 Result

Performance of the best fit is reported below.

Residual Error (Averaged over the 5 different splits)	7.63606

Best fit coefficients are saved as .csv file for better readability.

6 Regularized Linear Regression

6.1 Goal

The goal is to use Ridge regression on the CandC data for various values of l and report the residual error on test data, averaged over 5 different 80-20 splits, along with the coefficients learnt and perform the same with a reduced set of features.

6.2 Approach

Sometimes the input data tends to have correlated features leading to large weights for some features and also in some cases model may tend to overfit the data. Hence we add another term (regularisation term) in the optimisation problem to help keep weights small. In this case the regularisation term is the L2 norm of the weights and the regularisation parameter (λ) is varied in linear space from 0.01 to 50.

6.3 Result

For the best fit case, performance was reported as:

Lambda	Residual Error (Averaged over 5 different splits)
11.42857	7.25349

Features with absolute values of weights less than 0.02 were removed and the entire task was repeated and performance was reported.

Lambda	Residual Error (Averaged over 5 different splits)
11.42857	7.15070

Coefficients for the best fit data are saved as .csv file for better readability.