Lab Assignment 2

Vasant Honavar
DS 310 - Machine Learning

October 30, 2018

In all of the following exercises, if there is a need for a random seed, set it to 1234. Using different sklearn libraries are permitted as long as usage is well-understood and explained in the code. In case you would need to interpret your results, do so in your Ipython Notebook by changing the cell type and writing your interpretation immediately below the code and its result so that the interpretation can be matched with the result and the code. Submit a single Ipython Notebook in which all of the answers are organized in a way that can be run and evaluated.

1. **Random forest (RF)**. From sklearn import the Breast Cancer dataset. Randomly split it into train/test with 70/30 ratio. Using these data, train a Random Forest Classifier from the ensemble library of sklearn using 100 trees. Then, report the following:

   (a) The prediction values of the RF classifier on the test data (i.e., report 0's and 1's as class labels).

   (b) Report the feature importance obtained from the trained RF and plot a histogram of them.

2. **Regularized softmax linear classifiers**. Logistic regression is a well-known method for binary classification. In order to extend logistic regression to the case of multi-class classification, one could use the softmax function defined in slide #49 of the lecture slides on "probabilistic models." Additionally, one could add regularizers (e.g., an L2 norm to obtain sparse solutions) explained in slides #32-38 of the same lecture slides. In this excercise, we are interested in implementing an L2 regularized softmax classifier. After having your classifier implemented, train/test it with 70/30 ratio on the Iris dataset from sklearn (which has 3 class labels) and report the average accuracy obtained on the test data.