EE 657 - Pattern Recognition and Machine Learning

Jan - May 2018

Assignment - 2

☞ Submission Deadline: **22 Apr 2018 (Sunday), 9:00 PM**

☞ Submission should include:

1. Well documented code written in Python.
2. Report (pdf) explaining the summary of the experiments performed, the observations made and the conclusions. Include informative plots, figures and/or tables that justifies your findings.

OR

3. Jupyter notebook (a.k.a IPython notebook) with all the details mentioned in 2.

☞ Assignment submission link will be made online few days prior to the submission deadline.

1. Upload *.zip* file named as $< name > \_ < rollnumber >$.

2. Make sure that every details are correct in the first upload itself to avoid multiple submissions. Make the submission sufficiently in advance of the deadline to avoid last minute rush.

☞ The implimentation should be carried out using **Scikit-Learn** (python machine learning library). Use sklearn.tree.DecisionTreeClassifier and sklearn.tree.DecisionTreeRegressor for classification and regression respectively. The modules sklearn.model_selection, sklearn.metrics may be used for model selection and sklearn.tree.export_graphviz may be used for plotting the tree model.

☞ Useful References:

1. https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/.
2. http://scikit-learn.org/stable/modules/classes.html.
3. http://scikit-learn.org/stable/modules/model_evaluation.html.
4. http://scikit-learn.org/stable/auto_examples/index.html.

---

1. **Dataset Description:** *Wisconsin Diagnostic Breast Cancer(WDBC) dataset* from the *UCI repository*. Each row in the dataset represents a sample of biopsied tissue. The tissue for each sample is imaged and 10 characteristics of the nuclei of cells present in each image are characterized. These characteristics are: *Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Number of concave portions of contour, Symmetry, Fractal dimension.* Each sample used in the dataset is a feature vector of length 30. The first 10 entries in this feature vector are the mean of the characteristics listed above for each image. The second 10 are the standard deviation and last 10 are the largest value of each of these characteristics present in each image.

   - **Training data:** ‘ `trainX.csv` ’ consisting of 455 samples, 30 attributes. The label associated with each sample is provided in ‘ `trainY.csv` ’ . A label of value 1 indicates the sample was for malignant (cancerous) tissue, 0 indicates the sample was for benign tissue. .

- **Test data:** ‘ `testX.csv` ′ consisting of 57 samples, 30 attributes. The label associated with each sample is provided in ‘ `testY.csv` ′.

Problem: Use decision trees to classify the test data. Estimate the misclassification rates of both classes and populate the 2x2 confusion matrix.
(i) Report the following: (a) Plot of decision tree model. (b) the total number of nodes in the tree. (c) the total number of leaf nodes in the tree.
(ii) Train your binary decision tree with increasing sizes of training set, say 10%, 20%, ..., 100%. and test the trees with the test set. Make a plot to show how training and test accuracies vary with number of training samples.

2. **Dataset Description:** The data set is derived from a two-year usage log of a Washington, D.C. bike-sharing system called Captial Bike Sharing (CBS). Bike sharing systems are variants of traditional bicycle rentals, where the process of renting and returning is heavily automated; typically, bikes can be rented at one location and returned at another without ever having to deal with a human being. The goal is to predict the daily level of bicycle rentals from environmental and seasonal variables using decision trees.

   - **Data:** ‘ `bikes.csv` ′ has the following attributes:

     - `date:` The full date, in year-month-day format.
     - `season:` Season of the year, 1 to 4
     - `year:` Year, 0=2011, 1=2012
     - `month:` Month (1 to 12)
     - `holiday:` Whether the day is holiday or not
     - `weekday:` Day of the week (coded by 0-6)
     - `workingday:` 1 for working days, 0 for weekends and holidays
     - `weather:` Weather, coded as follows:
       1. Clear to partly cloudy
       2. Mist but no heavier precipitation
       3. Light rain or snow, possibly with thunder
       4. Heavy rain or snow
     - `temp:` Normalized temperature in Celsius. The values are derived via $(t - t_{min})/(t_{max} - t_{min})$; $t_{min} = -8$, $t_{max} = +39$.
     - `humidity:` Normalized humidity (actual humidity divided by 100).
     - `windspeed:` Normalized wind speed (actual wind speed in miles per hour divided by 67).
     - `count:` Count of total bike rentals that day, including both casual and registered users.

   ⋆ The response variable of interest is `count`, the total number of rentals each day.

   Problem: Build a regression tree predicting daily bike rentals from all available variables.
   (i) Report the following: (a) Plot of regression tree (b) The total number of leaf nodes in the tree, (c) Into how many different groups of days does the tree divide the data, (d) Which variables appear in the tree, (e) Which variables are important, (f) The MSE.

   (ii) Now re-code the months so that January and February share one code, May through October shares another, and March, April, November and December share a third. Re-estimate the regression tree. How does the tree change (if at all)? What is the MSE? Did it improve the fit?

****