BME777: Emerging Topics in Biomedical Engineering (Fall 2017)
(Machine Learning for Health Analytics)

Lab 4: Unsupervised and Algorithm Independent Machine Learning

Objectives

- Implement k-means clustering algorithm to identify the clusters within the given unlabeled data.
- Perform estimation of classification accuracies using leave-one-out and bootstrap methods.

Background

In unsupervised learning, natural clusters within unlabeled data samples (i.e. with no categorical information) are identified using iterative learning process. The functional form of the underlying probability densities of the data are assumed to be known and that the only thing that must be learned is the value of an unknown parameter vector. One elementary but popular approximate method that performs the above is the k-means clustering algorithm. The goal of the k-means clustering algorithm is to identify k mean vectors or cluster centers within the given unlabeled data. In the k-means clustering algorithm, we begin with randomly initializing the mean vectors (k cluster centers) and then assigning the data points to the nearest cluster by computing the euclidean distance. Once all the data points are assigned to one of the k clusters, the mean vectors of the k clusters are recomputed. The process is repeated until there is no change observed in the recomputed mean vectors of the k clusters. The following pseudo code describes the k means clustering algorithm:

1 **begin** **initialize** n, c, $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\ldots$, $\boldsymbol{\mu_c}$

2 **do** classify n samples according to nearest $\boldsymbol{\mu_i}$

3 recompute $\boldsymbol{\mu_i}$

4 **until** no change in $\boldsymbol{\mu_i}$

5 return $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$,$\ldots$, $\boldsymbol{\mu_c}$

6 **end**

Laboratory Exercises

Part I

Load the data file "DataLab4.mat". This loads a matrix variable "Breast_Tissue "to the MATLAB workspace. The size of the data matrix is 42 X 3 (i.e. 42 rows

and 3 columns) containing 42 unlabeled data samples with a feature dimension of 3. The dataset comprises of electrical impedance measurements on breast tissues from carcinoma, fibro-adenoma, and mastopathy classes. The specific features being: (i) Impedivity (ohm) at zero frequency, (ii) High-frequency slope of phase angle, and (iii) Area under spectrum. Perform the k-means clustering of the given data for the following conditions:

a. With c=3 and initial cluster centers be $\boldsymbol{\mu_1}(0)$=(400, 0.7, 5800)^t, $\boldsymbol{\mu_2}(0)$=(250, 0.3, 400)^t, and $\boldsymbol{\mu_3}(0)$=(300, 1.1, 1082)^t.

b. Repeat the above step with different initial cluster centers than the above.

c. Provide necessary figures (at least showing two stages of clustering progress).

Part II

Using your Lab2 solution and DataLab2_1.mat, estimate the classification accuracy of the linear classifier using Leave-One-Out and Bootstrap methods.

a. Compute the Leave-One-Out classification accuracies by iteratively calculating the classification accuracy of the linear classifier by training with $n-1$ samples and testing with the left out sample.

b. Compute the Jack-knife estimate of the final classification accuracy by calculating the mean of the Leave-One-Out classification accuracies.

c. Repeat the above steps, however using the Bootstrap method instead of the Leave-One-Out method. In the Bootstrap method at each iteration you generate a new bootstrap dataset with"n" samples by randomly selecting "n" samples from the original data with replacement. Then use a 2-fold cross validation in computing the classification accuracy for each of the new bootstrap dataset and compute the final classification accuracy by taking the mean of the classification accuracies obtained from 20 bootstrap datasets.

Lab Evaluation Information

- Submit via email the documented MATLAB code for Part I & II of the lab exercise before the demo. In the lab, the programs should be demonstrated to the TA/Instructor and each student will be evaluated on the components covered in the lab material. Marks will be provided based on the (i) demo and (ii) student′s response to the questions [3 + 4 marks].

- Submit via email a report addressing all the outcomes of the lab exercise, and your observations and conclusions. (3 marks)

Due date

- Due date for the report will be before the start of the next lab. Reports can be submitted via email till 5 pm on the due date, however demo/ evaluation in lab must be completed during scheduled lab sessions for this lab (i.e., Week 9-10).

Acknowledgment

We thankfully acknowledge UCI Machine Learning Repository for the dataset used in this lab exercise.
Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.