Lab 4: Classification

Assigned: November 24, 2023 Due: December 8, 2023

Lab objectives

Using the features generated in lab 3:

(1) Practice feature screening/selection using Fisher’s ratio

(2) Selecting suitable features and classifiers for the classification problem

Description of the data set

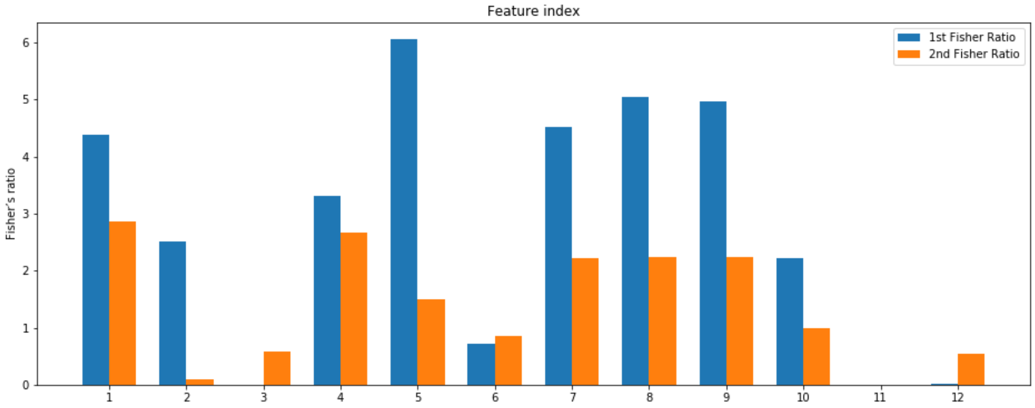
This lab is a continuation of Lab 3 and we will use the same data set for classifier design. In Lab 3, we performed data preprocessing and feature generation. In this lab, we will investigate the classification performance when both features and classification algorithms are varied. For your convenience, we have extracted 10 features, which are available in the attached csv file. To avoid any bias inherited from your results from previous labs and facilitate the data-driven study, we have normalized and anonymized all features. Note that this processing will not influence the classification performance. The dataset has been randomly partitioned into a training set and a test set.

Analysis procedure

1. Fisher's ratio.

    a) Calculate Fisher's ratio for Clean-L1 (Fisher's ratio 1) and Clean-L2 (Fisher's ratio 2) for all 10 features. Report the results in one bar plot. The vertical axis is Fisher's ratio, and the horizontal axis is the feature name. Use two bars to represent two Fisher's ratios for each feature. (15 points)

Example:



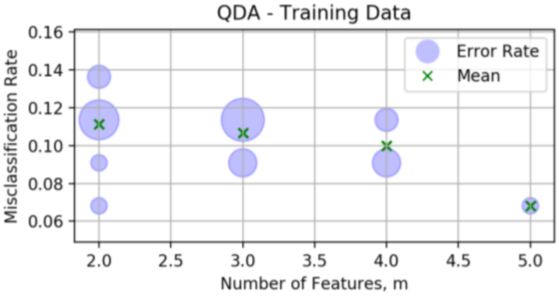
   b) Intuitively, is Clean-L1 or Clean-L2 relatively easy to be distinguished? (5 points)

2. Feature screening. Select five features from the feature pool based on the following criterion: max (Fisher's ratio 1 + Fisher's ratio 2). Report the selected features. (20 points)

3. Feature selection & classification. Repeat the following procedure for both classifiers: LDA and QDA. (Some useful links for LDA and QDA.)

   a) Train and test the classifier for any possible combination of m features from the selected four features from step 2, where $m \in \{2,3,4,5\}$. Organize the result as $[m, e]$, where e is the test misclassification rate. Plot $[m, e]$ in a scatter plot. Note that you will have different numbers of points for different values of m (use suitable python package to generate all combinations). See the sample code for how to generate this plot. (25 points)

   Example: Larger circles indicate multiple combinations of m that have the same misclassification rate. You can use any method to present your result.



   b) In general, which m (number of selected features) performs best? (5 points)

   c) From all combinations studied in (a), find the best classifier and report its confusion matrices for both training and test data sets. Are the training and test errors comparable? (20 points)

4. Does LDA or QDA perform better based on your results above? (10 points)