Project 2 - Non-parametric Classification with Dimensionality Reduction - Due 10/10/19

Objectives

Design non-parametric classifier and compare that with parametric classifiers. Implement both supervised and unsupervised dimensionality reduction approaches as preprocessing steps to classification. Also practice nomalization as another preprocessing step.

Data set used:

Download *pima.tr* and from Ripley's Pattern Recognition and Neural Networks. This is a 2-category 7-dimensional classification problem.

Basic requirement (80)

- Task 1 (10): Preprocess the data set by normalizing it first. Refer to the README file for details on the features/attributes used to differentiating between diabetes patients vs. normal patients. Denote the original dataset as X and the normalized data set as nX
  - Change 'yes' and 'no' to 1 and 0 indicating 'with disease' and 'without disease'.
  - Delete the first row in the data set
  - Normalize the data set to make the features comparable (or with the same scale). Suppose x is a sample vector, m_i is the mean of each feature i, sigma_i is the standard deviation of each feature i, then normalization is conducted by (x-m_i)/sigma_i. Keep in mind that you also need to normalize the samples in the test set. Be careful which mean and standard deviation you should use. (For each sample in the test set, use the same m_i and sigma_i you derived from the training set.)
- Task 2 (10): Transform the normalized dataset using principal component analysis (PCA). Denote the transformed data set as pX.
  - Use PCA to derive a new set of basis and choose the major axes with an error rate not greater than 0.10.
  - Represent the data using this new set of basis for a reduced dimension
- Task 3 (10): Using Fisher's linear discriminant (FLD) method to derive the projection direction that best separates the projected data, and generate the projected data. Denote it as fX.
- Task 4 (25): Implement kNN
  - The implementation should be able to flexibly change the value of "k"
  - The implementation should be able to measure the run time
- Task 5 (25): Performance Evaluation
  - Task 5.1: Use nX. Classify the test set using discriminant functions (Cases I, II, and III from project 1) as well as kNN.
    - Draw a performance curve with accuracy vs. k values where prior probability is calculated based on the training set.
    - Compare the performance of all four classifiers using prior probability determined by the training set, for fair comparison. Provide TP, TN, FP, FN values.
    - Vary the prior probability and plot sensitivity and specificity with respect to prior probability for the four classifiers.
  - Task 5.2: Repeat Task 5.1 on pX.
    In addition to the above, plot sensitivity and specificity curves against different error rate (or different numbers of eigenvectors, from 1 to 7)
  - Task 5.3: Repeat Task 5.1 on fX.

Report (20)