Project 2 - Classification with Dimensionality Reduction and Performance Evaluation - Due 02/27/18

Basic requirement (80)

- Task 1 (10): Preprocess the data set. Denote the original dataset as X and the preprocessed data set as nX
  - Dataset used in this project: pima.tr (200 samples) and pima.te (332 samples). They can be downloaded from the Testing Dataset link on the course website, or http://www.stats.ox.ac.uk/pub/PRNN/.
  - Each sample in the data set has 7 dimensions. The classification result is 'no': (no diabetes) or 'yes' (diabetes)
  - Refer to the README file for details on the features used
  - Before you can use the dataset, you need to do some preprocessing
    - 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. For each sample in the test set, use the same m_i and sigma_i you derived from the training set
- Task 2 (15): Transform the preprocessed 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 (15): 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 (40): Classification
  - Task 4.1: Use nX. Classify the test set using discriminant functions (Cases I, II, and III) ~~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 three ~~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 three classifiers.
    - Vary the prior probability and plot precision and recall with respect to prior probability
    - (+15 Bonus for UG) Draw a ROC curve for each of the four classifiers by varying the prior probability. Use the k that generates the best accuracy in the previous drawing.
  - Task 4.2: Repeat Task 4.1 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 4.3: Repeat Task 4.1 on fX.

Report (20)