Final Project

The objective of the final project is to integrate various machine learning techniques to achieve the best performance. Final project is a group effort. Each group can have 2-3 members. You are required to apply ALL techniques learned in this semester.

Schedule

- (5) Milestone 1: Group formation and topic selection (due 11/05). Submit through Canvas. Approval and comments will be returned in one day. The same topic cannot be chosen by more than 1 group. The topic follows the first-come first-served rule. So decide your group formation and pick a topic as soon as possible.
- (5) Milestone 2: Background study (2-page report due 04/25). Submit through Canvas.
- (5) Milestone 3: Preliminary results. Applying at least one learned technique successfully on the chosen dataset and submit a 1-page report. (Due 11/19)
- (100) Final presentation (12/03 and 12/12) (Presentation slides due the midnight before the presentation. Submit through Canvas)
- (85) Final report (due 12/13). Submit through Canvas.

Potential Topics

Each group can choose one topic from the following sources

- KDD-Cup 1997-2006
- Kaggle Competitions
- Other topics: You can select a topic yourself from other resources with the approval from the instructor.

Requirement

General steps involved in a pattern recognition problem include

- Data collection (raw data)
- Feature derivation (how to derive features from the raw data)
- Feature selection (dimensionality reduction, Fisher's linear discriminant or PCA)
- Classification
    - classification based on supervised learning or unsupervised learning
    - classification based on parametric or non-parametric density estimation in supervised learning
    - classifier fusion
- Performance evaluation
- Feedback system

You are required to evaluate the effect of various aspects of the classification process, including but not limited to

- the effect of assuming the data is Gaussian-distributed
- the effect of assuming parametric pdf vs. non-parametric pdf
- the effect of using different prior probability ratio
- the effect of using different orders of Minkowski distance
- the effect of knowing the class label
- the effect of dimension of the feature space (e.g., changed through dimensionality reduction)
- the effect of classifier fusion

To be more specific, you need to at least go through the following steps:

- Data normalization
- Dimensionality reduction
- Classification with the following
  - MPP (case 1, 2, and 3)
  - kNN with different k's
  - BPNN (can use open-source software package or the toolbox comes with MATLAB)
  - Decision tree (can use open-source software package or the toolbox comes with MATLAB)
  - SVM (can use open-source software package or the toolbox comes with MATLAB)
  - clustering (kmeans, wta, kohonen, or mean-shift)
- Classifier fusion
- Evaluation (use n-fold cross validation to generate confusion matrix and ROC curve if applicable).