Project 1: Supervised Learning Using Baysian Decision Rule - Two Category Classification (Due 09/19)

Objective:

The objective of this project is, first of all, to learn how to implement supervised learning algorithms based on Baysian decision theory. The second objective is to get you familiar with the design flow when applying machine learning algorithms to solve real-world problems. Some practical considerations include, for example, 1) the selection of the right pdf model to characterize the data distribution in the training set, 2) the selection of the right ratio of prior probability, 3) the different ways to evaluate the performance of the learning algorithm, and 4) how differently the same ML algorithm performs when applied to different datasets.

Data Sets:

The synthetic dataset: synth.tr (the training set) and synth.te (the test set) from Ripley's Pattern Recognition and Neural Networks.

Algorithm:

You need to implement the three cases of the discriminant function (parametric learning) and kNN (non-parametric learning) based on Baysian decision theory, 1) minimum Euclidean distance classifier (linear machine), 2) minimum Mahalanobis distance classifier (linear machine), 3) the generic form of Baysian decision rule (quadratic machine), where Gaussian pdf is assumed, and 4) kNN.

Performance Metrics:

Three metrics are used to evaluate the performance of the ML algorithms, including 1) overall classification accuracy, 2) classwise classification accuracy, and 3) run time.

Tasks:

- (5 pts) Show a scatter plot of the training set of the two classes. From visual inspection, do you think single-modal Gaussian is a good/reasonable model for the pdf?
- (15 pts) Plot a figure with the x-axis showing the different "k" values in kNN and the y-axis showing the overall classification accuracy.
- (50 pts) Assuming equal prior probability, generate a table summarizing the overall classification accuracy, classwise accuracy, and run time of the four supervised learning algorithms, with each row indicating a learning algorithm and each column indicating a performance metric. For kNN, choose the best "k" you obtained from Task 2.
- (10 pts) Provide a comprehensive discussion (0.5 ~ 1 page) on the results shown in the table, including the effect of using different assumptions of the covariance matrices.
- (15 pts) Using the synthetic dataset, illustrate the four decision boundaries from the three cases of parametric learning algorithms on the same figure as the scatter plot of the testing dataset. The fourth one should be that from kNN (with the best "k"). Note that for kNN, there won't be an analytical equation to describe the boundary. Comment on the differences.
- (5 pts) Final discussion.