Data Mining and Machine Learning Homework #3: Classification methods
Due on Wednesday, Mar 2, 2011, in class.

*Format of the report: Explain clearly your methods and your results. Point out what you find interesting and try to make sense of it. Use tables and plots if they better represent your results and explain them. Please put your code and less important output in the appendix.*

1. Explore the whole Zip code data(with all 10 digits) and classify the images by using the methods we disscussed in class, namely k-NN, LDA, QDA, reduced rank LDA, and logistic regression. Compare the performance these methods. Cross-validation should be considered as you choose the dimension for reduced rank LDA.

2. Regularization for a classification problem.

   (a) Read the Section 5.2.3 (*Example: Phoneme Recognition*) from the textbook and information on the data "Phoneme" from the book website. Make sure that you understand the data, the problem of using raw logistic regression, and how the regularization is done. Conduct logistic regression with regularization on the subset of the data, as described in the textbook and confirm your result with the table on training error and test error for raw and regularized methods presented in this section. Extend the analysis to the entire dataset with training sample and test sample of equal size. Report the training and test errors for raw and regularized logistic regressions.

   (b) Incorporate the idea of regularization into quadratic discriminant analysis. Decide beforehand on the series of five different choices for the number and position of the knots, and use tenfold cross-validation to make the final selection. Compare the raw and regularized (with the selected series of knots) QDA's on the training and test samples in (a).