Statistical Learning I - Homework # 4, Due November 2

1. We will be looking at the **trees** data which is already in the standard R package. First build a linear model for predicting the volume of the trees from the girth, height, and the product girth*height and look at a summary of this model. Now bootstrap the tree data 1000 times and, for each bootstrap sample, fit the same model and save the coefficients. Now look at the standard deviations for the coefficients across the bootstrap samples. How to they compare with the standard errors given in the summary of your first model?

2. Load the $MASS$ library into R. We will be working with the **Pima.tr** data set which was collected on female Pima indians. The objective is to build a model to classify female Pima indians as either diabetic or non-diabetic based on the other variables. If you type "**help**(Pima.tr)" you will see what the variables represent.

- **a.** Build a logistic regression model to classify the woman. Feel free to use the **step** function to choose a suitable model (for now we will not use cross-validation). Construct a confusion matrix to show how well the model classified the training data.
- **b.** Build an LDA model to classify the women and construct a confusion matrix for this model as well.
- **c.** Build a QDA model to classify the women and construct a confusion matrix for this model as well.
- **d.** Which models did the best job classifying training data?
- **e.** Now build confusion matrices for each of the three models with predictions for the **Pima.te** data set. Which models generalize best?

3. Suppose that $P(X = x|G = j) = f_j(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{(x-\mu_j)^2}{2\sigma_j^2})$ for $j = 1, 2$ and where $\sigma_1^2 \neq \sigma_2^2$. Let π_1 and π_2 be the prior probabilities that $G = 1$ and $G = 2$ respectively.

- **a.** The decision boundary for classifying an observation as either group 1 or 2 using a QDA is given by the zeroes of a quadratic function of x. Write out this quadratic equation as a function of the parameters in the model.
- **b.** Let $\sigma_1 = 1/3$, $\sigma_2 = 1/2$, $\mu_1 = 1$, $\mu_2 = 2$, and $\pi_1 = \pi_2 = 1/2$. What values of x would predict $G = 1$?

4. Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed Poisson random variables with mean θ. What is the log-likelihood function? Maximize it as a function of θ and give the maximum likelihood estimator $\hat{\theta}$ for the mean of this distribution.