STA561/CS571 — Fall 2013 **Homework 3** Due: September 23, 2013

Instructions: Please put all answers in a single PDF with your name and NetID and upload to SAKAI before class on the due date (there is a LaTeX template on the course web site for you to use). Definitely consider working in a group; please include the names of the people in your group and write up your solutions separately. If you look at any references (even wikipedia), cite them. If you happen to track the number of hours you spent on the homework, it would be great if you could put that at the top of your homework to give us an indication of how difficult it was.

Problem 1

Linear Regression. Some researchers, desperately in need of a machine learning expert, bring you a dataset with information on $n = 1100$ people. Their study has two explanatory predictors: X_1 = a binary indicator of gender (female = 1), and X_2 = weight. They want to use this information to help predict blood pressure Y which they believe is linearly related to X_1 and X_2.

Suppose that $\sigma^2 = 1$, and, for part (c), $\tau^2 = 1$. Use the first 1000 records for your training set, and the last 100 records for your test set. For this answer, include your R code in your solution, and do not use built in functions for linear regression.

(a) Write a program in R to estimate β using the normal equations. Estimate β from the training set.

(b) Write a program in R to estimate β using online stochastic gradient descent. Estimate β from the training set.

(c) Write a program in R to estimate β using the ridge regression normal equations. Estimate β from the training set.

For **all** of the above estimation procedures:

(d) Calculate $\mathsf{RSS}(\hat{\beta})$ in the training dataset.

(e) Calculate $\mathsf{RSS}(\hat{\beta})$ in the test dataset.

(f) For each of the estimated values of $\hat{\beta}$, what is $\mathbb{E}[Y \mid X = [1, 135]^T, \hat{\beta}]$?

(g) Separately for both features and all three regression estimations, generate a scatter plot of $Y - beta_i Xi - beta_0$ vs. the feature Xj (ALL OF THIS BEING FROM THE TRAINING DATASET). You can think of the scatter plot as (Y minus the effects of the estimated intercept and effects of feature Xi) vs (the feature Xj). Then, add the estimated regression line to the scatter plot (this will result in six figures with one regression line and a scatter plot in each figure).

So for the first figure, we are concerned with the beta estimations from the normal equations and feature $X1$. Therefore, the scatter plot should be $(Y - beta_{ols}(2) *$

$X2 - beta_{ols}(0)$) vs $X1$. Then plot the regression line ($beta_{ols}(1) * X1$) vs $X1$. Where $beta_{ols}(0)$ is the intercept term from the OLS estimation.

Similarly for the second figure, the scatter plot will be ($Y - beta_{ols}(1) * X1 - beta_{ols}(0)$) vs $X2$. And then the line on top of this scatter plot will be ($beta_{ols}(2) * X2$) vs $X2$.

Then the third figure will consist of the SGD beta estimations and the feature $X1$. So the scatter plot will be ($Y - beta_{sgd}(2) * X2 - beta_{sgd}(0)$) vs $X1$. Then plot the regression line ($beta_{sgd}(1) * X1$) vs $X1$. Where $beta_{sgd}(0)$ is the intercept term from the SGD estimation.

And so on...

Summarize your findings:

(h) Comment on the propensity of these estimation procedures to 'overfit' to the training data.

(i) The researchers need an answer! Suggest the best estimation procedure for the researchers' question and justify your choice.

Problem 2

*L*ogistic Regression. The researchers are back again! This time they are interested in doing prediction for a binary outcome Z (an indicator of adverse reaction to a drug they are testing), which they again believe is linearly related to X_1 and X_2.

Again, use the first 1000 records for your training set, and the last 100 records for your test set.

(a) Write a program in R to estimate β using Iteratively Reweighted Least Squares (book section 8.3.4). Estimate β using the training data.

(b) Calculate $\mathsf{RSS}(\hat{\beta})$ in the training dataset.

(c) Calculate $\mathsf{RSS}(\hat{\beta})$ in the test dataset.

(d) What is $\mathbb{E}[Z \mid X = [1, 135]^T, \hat{\beta}]$?

(e) Separately for both features and all three regression estimations, generate a scatter plot of $Z - beta_i Xi - beta_0$ vs. the feature Xj (ALL OF THIS BEING FROM THE TRAINING DATASET). You can think of the scatter plot as (Z minus the effects of the estimated intercept and effects of feature Xi) vs (the feature Xj). Then, add the estimated regression line to the scatter plot (this will result in two figures with one regression line and a scatter plot in each figure).

So for the first figure, we are concerned with the beta estimations from the IRLS algorithm and feature $X1$. Therefore, the scatter plot should be $(Z - beta(2) * X2 - beta(0))$ vs $X1$. Then plot the regression line $(beta(1) * X1)$ vs $X1$. Where $beta(0)$ is the intercept term from the IRLS estimation.

Similarly for the second figure, the scatter plot will be $(Z - beta(1) * X1 - beta(0))$ vs $X2$. And then the line on top of this scatter plot will be $(beta(2) * X2)$ vs $X2$.

(f) Comment on the correlation between the features (is there correlation?) and how this may or may not impact your estimates of $\hat{\beta}$.

(g) The researchers need an answer! What decision rule will you use to predict Z^* given a new X^*?