CS 498 AML: Homework 5

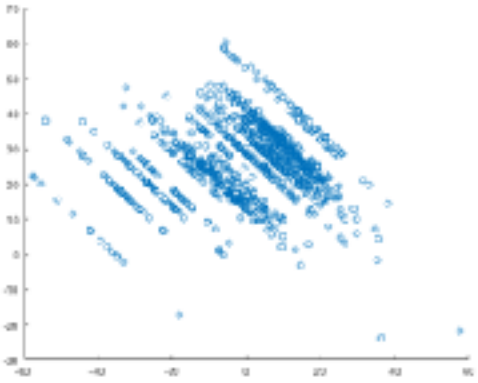
Rauhul Varma (rvarma2), Tadas Aleksonis (alekson2)

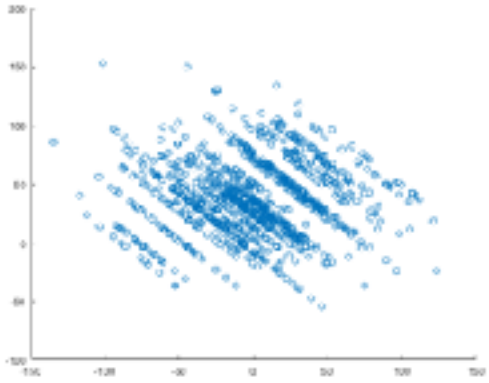
Problem 1

ALL GRAPHS AND VALUES ARE ORDERED: (LATITUDE, LONGITUDE)

a. First, build a straightforward linear regression of latitude (resp. longitude) against features. What is the R-squared? Plot a graph evaluating each regression.

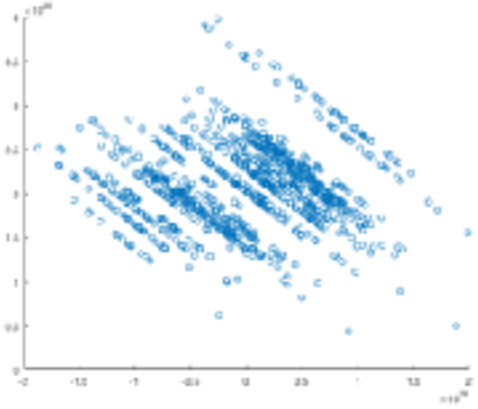
R^2 Values: (0.2928, 0.3646).

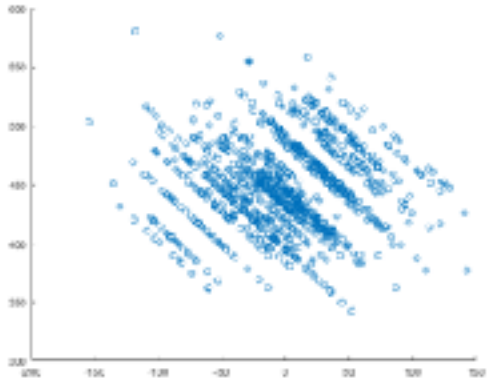




b. Does a Box-Cox transformation improve the regressions? Why do you say so? For the rest of the exercise, use the transformation if it does improve things, otherwise, use the raw data.

Yes, a Box-Cox transformation improves the regressions slightly, we know this because the R2 values increased after transforming the data, new values: (0.3264, 0.3647).





c. Use glmnet to produce:
    a. A regression regularized by L2. Is the regularized regression better than the unregularized regression?

        No the regularized regression produced a model that represented the data slightly less well with R^2 Values: (0.2913, 0.3629).

    b. A regression regularized by L. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?

        This regression used 100 of the 117 variables and also did not produce a better model than the unregularized regression, with R^2 values: (0.2585, 0.3327).

Problem 2

Use logistic regression to predict whether the user defaults. You should ignore outliers, but you should try the various regularization schemes we have discussed.

Using an alpha value of 0.5 we created a model that used 18 of the 23 explanatory variables and could predict whether the user defaults with a 80.43% accuracy. When the alpha was increased to 1 (lasso regression) we achieved a slightly better model that used 17 explanatory variables to predict whether the user defaults with a 80.65% accuracy.

Problem 3

Use a binomial regression model (i.e. logistic regression) with the lasso to predict tumorous/ normal. Use cross-validation to assess how accurate your model is. Report both AUC and deviance. How many genes does the best model use?

Using a logistic regression with a the lasso to predict tumorous vs normal produced a model that used 10 of the 2000 genes (explanatory variables). This model could predict tumorous vs normal with 88.71% accuracy with a deviance of 53.446 and AUC of 0.1566.