STATS 202: DATA MINING AND ANALYSIS HOMEWORK #3

INSTRUCTOR: LINH TRAN, HOMEWORK #3, DUE DATE: AUGUST 2, 2023, STANFORD UNIVERSITY, AND STUDENT: ADAM KAINIKARA

Problem 1 (7 points)

Chapter 6, Exercise 3 (p. 283).

$\Sigma_{i=1}^{n}(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_{ij})^2$ subject to $\Sigma_{j=1}^{p}|\beta| \leq s$

The left square part is RSS

a) (iv) As we increase s from 0 the training RSS will steadily increase. Increasing s makes us restrict the β (coefficients) less and less. By restricting coefficients less, the model will become more flexible. When the model becomes more flexible the training RSS will decrease.

b) (ii) As we increase s from 0 the test RSS would at first decrease but then slowly start increasing and would form a U shape. Increasing s makes us restrict the β (coefficients) less and less. By restricting coefficients less, the model will become more flexible. However at some point on test data, the RSS will once again increase.

c) (iii) As we increase s from 0, variance would steadily increase. Increasing s makes us restrict the β (coefficients) less and less. By restricting coefficients less, the model becomes more flexible. As flexibility increases, variance increases. This effect is like the bias variance trade off graph.

d) (iv) As we increase s from 0, squared bias would steadily decrease. Increasing s makes us restrict the β (coefficients) less and less. By restricting coefficients less, the model becomes more flexible. As flexibility increases, squared bias decreases. This effect is like the bias variance trade off graph.

e) (v) As we increase s from 0, the irreducible error would stay the same. Irreducible error is always there and does not come from the fitted model. So changing the flexibility of the model will have no impact.

Problem 2 (7 points)

Chapter 6, Exercise 4 (p. 284).

$\Sigma_{i=1}^{n}(y_i - \beta_0 - \Sigma_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\Sigma_{j=1}^{p}\beta_j^2$

Ridge?

a) As we increase λ from 0 the training RSS will increase. Increasing λ increases the penalty term. With the penalty increasing and becoming more significant, the coefficients decrease. This leads to more error.

b) Initial, increasing l shrinks the coefficients. So as it initially increases RSS will decrease since over fitting is reduced. However, if we continue to increase l, the model may become too simple (due to smaller coefficients) and start to under fit, leading to an increase in the test RSS.

c) As we increase λ from 0 the variance will increase. With the coefficients changing as λ increases, the variability between how well the models can fit increases which increases the variance.

d) As we increases λ the bias will decrease. As λ increases, and the coefficients decrease, the model may become more simple. This leads to a decrease in bias.

e) Remain constant. The irreducible error represents the inherent noise in the data that cannot be reduced through modeling. As λ increases, irreducible error i always there.

Problem 3 (7 points)

Chapter 6, Exercise 9 (p. 286). Don't do parts (e), (f), and (g).

a) Split the data into a 75% training and 25% test.

b) Fit a linear model. Got a MSE for linear model of 1503017.4360986822 and a r squared value of 0.9134874545115684

c) Fit a ridge model. Got a MSE of 1502973.7121541149 and a r squared value of 0.9134921934779143.

d) Fit a lasso model. Got a MSE of 35152969.33526452 and 9 non zero coefficients

Coding question

Problem 4 (7 points)

Chapter 7, Exercise 1 (p. 321).

a) We are given that $f(x) = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta^4(x-\xi)^3+$ where $(x-\xi)^3$ is its normal polynomial self if $x \geq \xi$ and is 0 otherwise

We are also given that $f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$

We want to find the polynomial and coefficients such that $f(x) = f_1(x)$

In part 'a' it is given that $x \leq \xi$ and because of this $f(x) = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

So in order for $f(x) = f_1(x)$ to be true the coefficients are:

$a_1 = \beta_0, b_1 = \beta_1, c_1 = \beta_2, d_1 = \beta_3$

b) Now given $f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$

In part 'b' it is given that $x \geq \xi$ so $f(x) = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x-\xi)^3$

$= \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x^3 - 3x^2\xi + 3x\xi^2 - \xi^3)$

Multiply out β_4 and rearrange the equation to follow the form of $f_2(x)\, ie: x^0, x^1, x^2 \ldots$

$(\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3$

In order for $f(x) = f_2(x)$

$a_2 = \beta_0 - \beta_4\xi^3, b_2 = \beta_1 + 3\beta_4\xi^2, c_2 = \beta_2 - 3\beta_4\xi, d_2 = \beta_3 + \beta_4$

c) Show$f_1(\xi) = f_2(\xi)$

$f_1(\xi) = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3$

$f_2(\xi) = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)\xi + (\beta_2 - 3\beta_4\xi)\xi^2 + (\beta_3 + \beta_4)\xi^3$

$\beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3 = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)\xi + (\beta_2 - 3\beta_4\xi)\xi^2 + (\beta_3 + \beta_4)\xi^3$

$= \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3 = \beta_0 - \beta_4\xi^3 + \beta_1\xi + 3\beta_4\xi^3 + \beta_2\xi^2 - 3\beta_4\xi^3 + \beta_3\xi^3 + \beta_4\xi^3$

$\beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3 = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3$

d)$Show: f_1'(\xi) = f_2'(\xi)$

$f_1'(\xi) = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2$

$f_2'(\xi) = -3\beta_4\xi^3 + \beta_1 + 9\beta_4\xi^2 + 2\beta_2\xi - 9\beta_4\xi^2 + 3\beta_3\xi^2 + 3\beta_4\xi^2$

$f_2'(\xi) = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2$

Therefore $f_1'(\xi) = f_2'(\xi)$

e)$Show: f_1''(\xi) = f_2''(\xi)$

$f_1''(\xi) = 2\beta_2 + 6\beta_3\xi$

$f_2''(\xi) = 2\beta_2 + 6\beta_3\xi$

Therefore $f_1''(\xi) = f_2''(\xi)$

Problem 5 (7 points)

Chapter 7, Exercise 8 (p. 324). Find at least one non-linear estimate which does better than linear regression, and justify this using a t-test or by showing an improvement in the cross-validation error with respect to a linear model. You must also produce a plot of the predictor X vs. the non-linear estimate f^(X).

Warning: Ignoring XDG_SESSION_TYPE=wayland on Gnome. Use QT_QPA_PLATFORM=waylan to run on Wayland anyway. T-test statistic: 0.16492863090989973 P-value: 0.8690858275585926 Linear CV error: 6.502276260524977 Polynomial CV error: 6.502276260524972

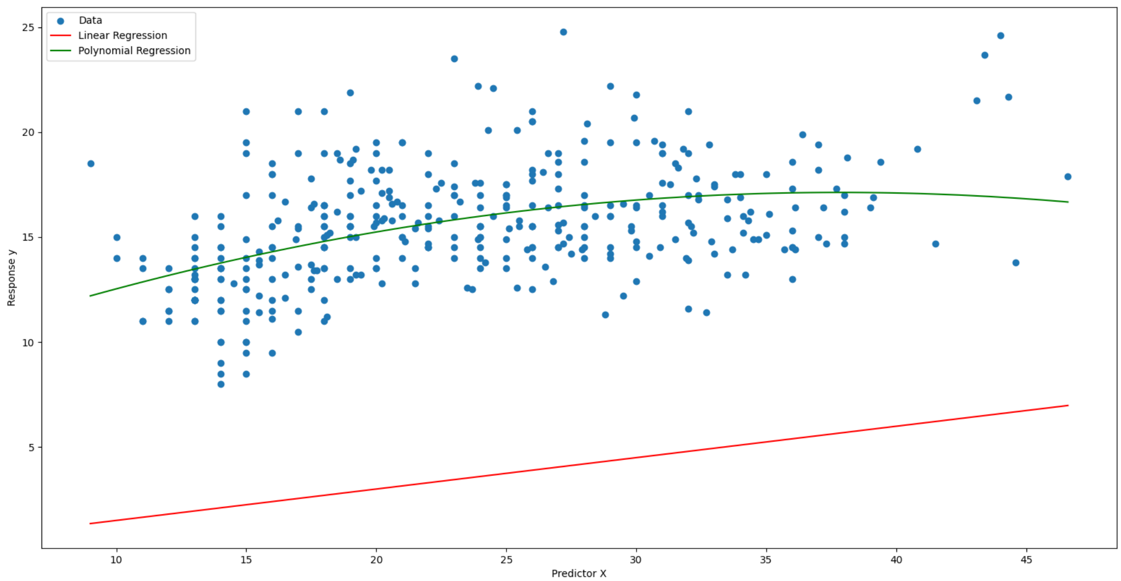

FIGURE 0.1. Poly and line

Problem 6 (7 points)
Chapter 9, Exercise 1 (p. 398).
Drawing a hyper plane

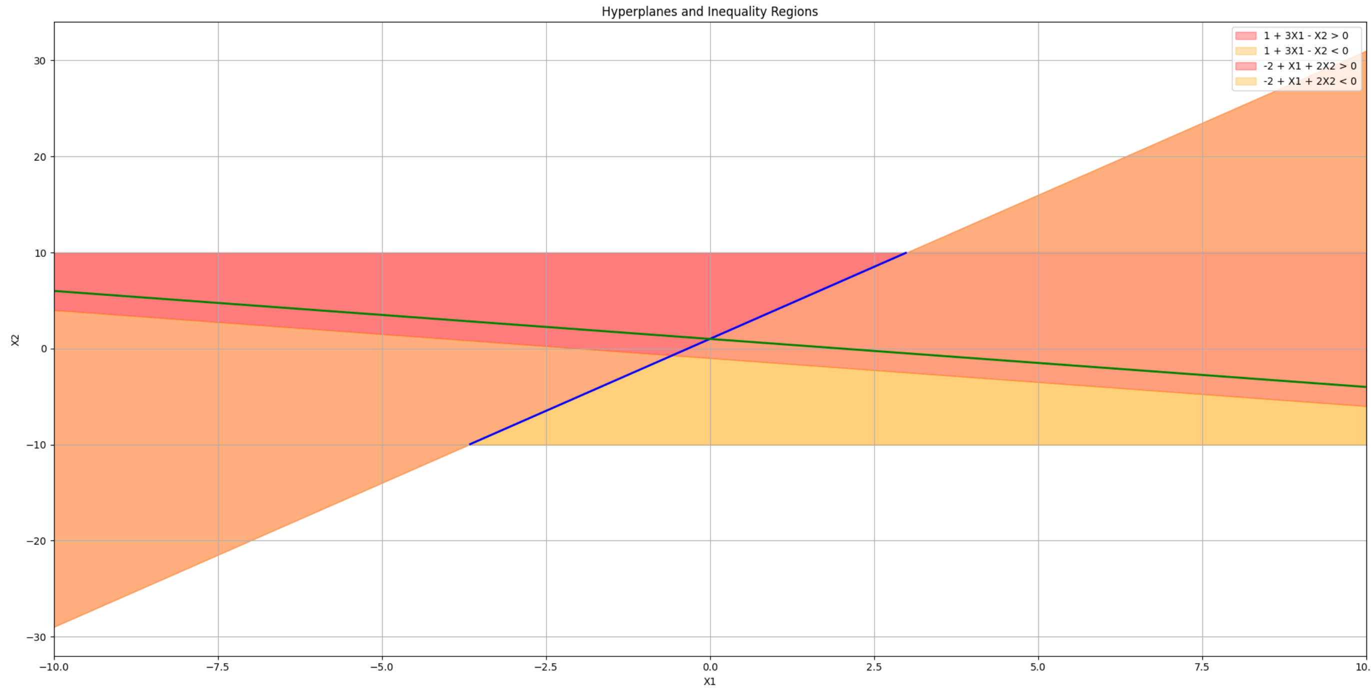

FIGURE 0.2. Hyper planes

Problem 7 (8 points)

Chapter 9, Exercise 8 (p. 401).

Note: Likely got some part of it wrong as I got the same accuracy scores for all and the best c value was 0.01.

Fitted a support vector classifier to the training data using C = 0.01, with Purchase as the response and the other variables as predictors. There were 612 support points. Training accuracy score of 0.38249999999999995 and test accuracy score of 0.4111111111111111 0.1 Fitted a support vector classifier to the training data using the best C = 0.1, and got Training accuracy score of 0.38249999999999995 and test accuracy score of 0.4111111111111111 Fitted a support vector classifier to the training data using C = 0.01, with Purchase as the response and the other variables as predictors. Used radial. There were 612 support points. Training accuracy score using radial is 0.38249999999999995 and test accuracy score of 0.4111111111111111 Fitted a support vector classifier to the training data using the best C = 0.01 with radial, and got Training accuracy score of 0.38249999999999995 and test accuracy score of 0.4111111111111111 Fitted a support vector classifier to the training data using C = 0.01, with Purchase as the response and the other variables as predictors. Used poly. There were 612 support points. Training accuracy score using radial is 0.38249999999999995 and test accuracy score of 0.4111111111111111 Fitted a support vector classifier to the training data using the best C = 0.01 with poly, and got Training accuracy score of 0.38249999999999995 and test accuracy score of 0.4111111111111111