IST 597: Homework 2

Justin Silverman

Problem 1 (50 points)

a. (1 point) Do a brief exploratory data analysis of the training data in `data_train_hw2_problem1.csv`. You will be building a model of `y` vs. `x1`, `x2` and `x3`. Just report if you find anything unusual or noteworthy (no is a perfectly acceptable answer).

b. (5 points) Fit the following model to the training data in `data_train_hw2_problem1.csv` using maximum likelihood. Report the resulting estimate for $\beta = (\beta_1, \beta_2, \beta_3)$ and σ^2 (that is, report $\hat{\beta}$ and $\hat{\sigma}^2$).

$$y_i \sim N(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \sigma^2)$$

If you want to be eligible for partial credit it may help if you show your work / explain how you got your answer.

c. (5 points) Using Mean Squared Error (MSE) as a performance metric. How well does your model perform on the training data?

d. (5 points) Now evaluate the MSE on the held-out testing data. How well does your model perform?

e. (10 points) Let $\hat{y}$ denote the model predictions at the training data, that is

$$\hat{y}_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}.$$

Plot the residuals, $\epsilon = (y - \hat{y})$, as a function of y. Explain what you see and explain other potential uses for this plot.

f. (9 points) Pretend this is real data and you want to get the best predictive / inferential model. Given what you found in part (c), what do you suggest doing now?

g. (5 points) Implement your solution and retrain your model on the training data. Give the resulting improved estimates for $\hat{\beta}$.

h. (5 points) Now evaluate the MSE of the new model on the training and testing sets. (Report the results)

i. (5 points) What we have done here is not great practice. We looked at our testing test to improve our original model. Explain why this could be a problem: why might your new fitted model not generalize well to new data? Explain how you could have recognized the problem before part (d) and corrected the issue before ever looking at the testing data.

Problem 2 (20 points)

a. (1 point) Do a brief exploratory analysis of the training data in `data_train_hw2_problem2.csv`. You will be building a model of `y` vs. `x1`, `x2` and `x3`. Just report if you find anything unusual or noteworthy (no is a perfectly acceptable answer).

b. (9 points) Let

$$X = \begin{bmatrix} | & | & | \\ x_1 & x_2 & x_3 \\ | & | & | \end{bmatrix}$$

What is the value of $X^T X$? What are the Eigenvalues of this matrix? What is this telling you? *Please answer this latter question truthfully and don't come back after doing the rest of the problem where you will likely learn the answer (getting this later question right or wrong won't change the scoring on this part of the question).*

c. (10 points) Using OLS, fit the model

$$y_i \sim N(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \sigma^2).$$

Whats wrong here? If your code threw an error, explain why. If it didn't, look into the documentation and report what assumption the code is making? If that later question is too difficult, you can get full credit by reporting what mathematical step is being used to not throw an error.

Problem 3 (30 points)

In this problem I have given you two dataset. One is labeled `data_train_hw2_problem3.csv` that contains both the predictive target `target` and predictors (columns starting with `V`). The other dataset is unlabeled `data_test_hw2_problem3.csv` (it contains only the predictors but no column `target`). Your goal is to build the best predictive model for `target` in the unlabeled dataset that you can.

Submit your predictions for the missing column `target` in the unlabeled dataset as a text file with the name `predictions_hw2p3.txt` (e.g., one value per row in the same order as the unlabeled dataset, don't include row or column labels).

Your answers will be scored based on the Mean Squared Error of your predictions. If you know how to produce probabilistic predictions, know that only the mean prediction will be scored.