Data Science Fundamentals
(CS M148)

Homework 1 Solutions

Submit to Gradescope
due: January 26, 2pm

1 Data & Bias

(a) **(6 points)** Your friend working at UCLA housing has been given the task of determining how students feel about the new dorms. Your friend wants to accomplish this by scraping Twitter and news article comment sections for tweets containing keywords and hashtags related to the new dorm and then running them through a model that does sentiment analysis. This model contains an algorithm that says whether the text exhibits positive, neutral, or negative sentiment. What kinds of selection bias might your friend's data collection method exhibit? (Refer to Week 1 Lecture 2, slide 39 for a list).

Potential biases:

- Voluntary bias: only people that want to talk about the new dorms will tweet
- Under-coverage bias: not everyone uses Twitter
- Non-response bias: see voluntary bias
- Response bias: people that tweet tend to feel strongly (either really like it or dislike it)

Other potential issues:

- There might be tweet about the dorm that are missed in the scraping process
- Sarcasm or other forms of confounding language might be used
- It can be hard to capture a spectrum of feelings (how to distinguish between really dislike and mildly dislike?)

(b) **(6 points)** In 2018, news reports came out about how Amazon tried to use an AI tool to assist in the hiring process, but this tool was scrapped due to it displaying bias against hiring women. (i) Explain why the tool was discriminating against women? (ii) The prediction of the AI program was based on the CV of the candidates, including their gender, the name of their university, previous positions, extra-curricular activities, etc. Amazon modified the AI program to make it not consider the gender of the candidate, but this was not enough to eliminate the bias. Explain why dropping the gender could not eliminate the bias.

(i) The training data for "successful" candidates was predominantly male, since males were overrepresented in Amazon's hiring before they implemented this AI tool. Thus, the model was able to pick up on features of resumes that males tend to exhibit, and so (ii) even if gendered words are removed, the AI can still infer gender from the rest of the resume. For example, the algorithm tends to discriminate against women studied in all-women universities, or participated in activities that included the word "women's," as in "women's chess club captain", etc. This article describes it well, calling the issue "self-selection bias."

2 KNN regression

Consider the following data points (x, y) = (0, 1), (1, 0), (2, 5), (3, 2), (4, 5).

Data Science Fundamentals
(CS M148)

Homework 1 Solutions

Submit to Gradescope
due: January 26, 2pm

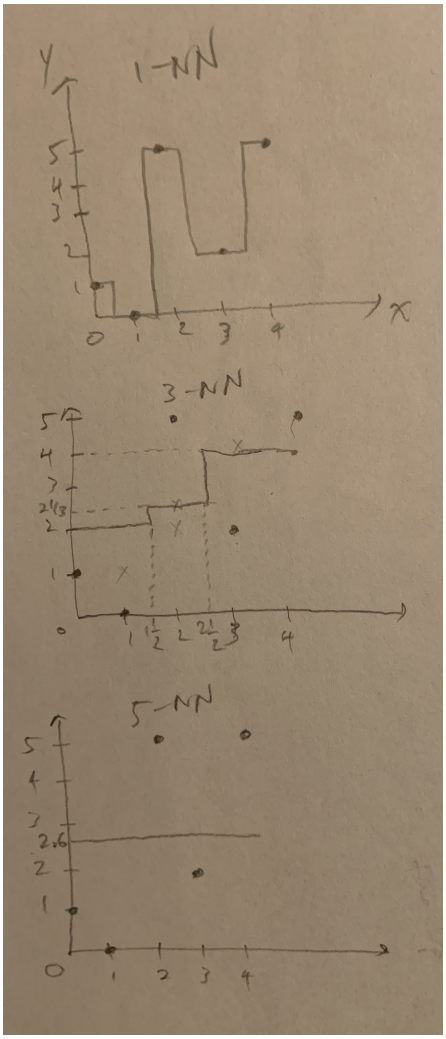


Figure 1: Answer for 2(a)

(a) **(5 points)** Draw the prediction for all x from [0, 4] using 1-NN, 3-NN, 5-NN regression based on those data points.

(b) **(3 points)** Given a test data set (0.5, 1), (1.5, 3), (2.5, 4), (3.5, 3). Which value of K (in the range 1 to 5) is the best and why?

We'll use MAE to calculate errors. MSE is also acceptable, but not shown here.

$K = 1$: We make the following predictions: (0.5, 0), (1.5, 0), (2.5, 2), (3.5, 2). This results in an MAE of

$$|1 - 0| + |3 - 0| + |4 - 2| + |3 - 2| = 7$$

.

$K = 2$: Predict (0.5, 0.5), (1.5, 2.5), (2.5, 3.5), (3.5, 3.5). MAE =

$$|1 - 0.5| + |3 - 2.5| + |4 - 3.5| + |3 - 3.5| = 2$$

$K = 3$: Predict (0.5, 2), (1.5, 2), (2.5, 2.333), (3.5, 4). MAE =

$$|1-2| + |3-2| + |4-2.333| + |3-4| = 4.667$$

$K = 4$: Predict (0.5, 2), (1.5, 2), (2.5, 3), (3.5, 3). MAE =

$$|1-2| + |3-2| + |4-3| + |3-3| = 3$$

$K = 5$: Predict (0.5, 2.6), (1.5, 2.6), (2.5, 2.6), (3.5, 2.6). MAE =

$$|1-2.6| + |3-2.6| + |4-2.6| + |3-2.6| = 3.8$$

Based on these results, $K = 2$ is the best since it has lowest MAE. You can also find the K value that gives the lowest MSE.

(c) **(3 points)** What is R^2 when K = 1 on the training data? Is this a good or bad model? Why?

$R^2 = 1$, and this is a bad model because you're overfitting to the data.

3 Linear Regression

Assume that you are fitting a linear regression of the form $Y = \beta_0 + \beta_1 X$ to a data set of n points: $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, using a MSE loss: $\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_i])^2$. The closed form solution for β_0, β_1 can be derived analytically, i.e. by taking the derivative of the loss w.r.t β_1, β_0, and set it equal to 0.

(a) **(5 points)** Derive the derivative of the above loss function and find the closed-form solution for β_0, β_1.

We'll find $\hat{\beta}_0$ first.

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)(-1)) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\beta_0 - \frac{1}{n}\sum_{i=1}^{n}\beta_1 x_i = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Now find $\hat{\beta}_1$, which is a bit trickier:

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)(-x_i)) = 0$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$\beta_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x^2 - n\bar{x}^2}$$

$$\beta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

**You can find the step-by-step derivation at the end of the homework solution.

(b) **(6 points)** Assume that you want to fit a line $Y = \beta_0$ to your training dataset. Using the argument in part (a), show that the optimal value for β_0 is equal to the mean of the y_i values if we use MSE, and is equal to the median of the y_i values if we use MAE.

MSE: Fitting a line $Y = \beta_0$ means a horizontal line, so $\beta_1 = 0$. Thus,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - 0(\bar{x}) = \bar{y}$$

MAE: We use the loss expression for MAE.

$$L = \frac{1}{n} \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - (\beta_0 + \beta_1 x_i)}{|y_i - (\beta_0 + \beta_1 x_i)|}(-1) = 0$$

Again, because it is a horizontal line, $\beta_1 = 0$.

$$\sum_{i=1}^{n} \frac{y_i - \beta_0}{|y_i - \beta_0|} = 0$$

$$\sum_{i=1}^{n} \text{sgn}(y_i - \beta_0) = 0$$

Interpreting this final equation, in order for it to be true, it means that half of the β_0 values must be greater than y_i and the other half of the β_0 values must be less than y_i. Thus, the only value of β_0 that satisfies this condition is the median of y_i, as desired. (Note: sgn indicates the sign function).

4 Linear Regression: goodness of fit & Interpretation

1- **(6 points)** US population was around 5 million in 1800, 23 million in 1850, 76 million in 1900, 161 million in 1950, and 291 million in 2000.

(a) use the analytic solution you derived in question 3(a) to fit a linear regression to the above data. What are β_0, β_1?

$\beta_0 = -2,586,800,000$ $\beta_1 = 1,420,000$

(b) What is R^2 for your model? Based on the value of R^2 can we say weather the estimated regression line fits the data well?

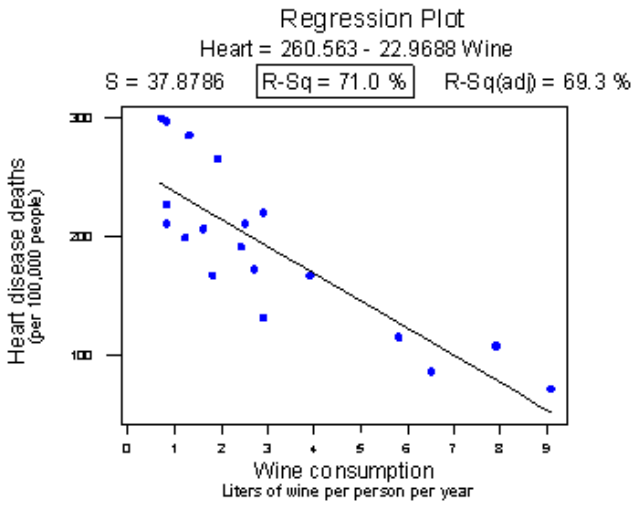
The value of R^2 is 0.9148. That is, only 8.52% of the variation in U.S. population is left to explain after taking into account the year in a linear way. However, Plotting the data suggests that a curve would describe the relationship even better. Indeed, the large value of 92.0% should not be interpreted as meaning that the estimated regression line fits the data well. (Its large value does suggest that taking into account year is better than not doing so. It just doesn't tell us that we could still do better.) Hence, from just the R^2 value alone, we cannot conclude if the regression line fits the data well or not.

(c) Plot the residuals versus year. What do you conclude?

Figure 2: Residual plot for 4-1(c)

From looking at the plot of residuals versus year, we can see that the residuals are not randomly distributed about 0, which indicates that a linear model is not expressive enough for this data. A linear model would be inappropriate to use - even though we have a high R^2 value, it doesn't mean that a linear model is a good fit for the data.

2- **(4 points)** The following plot shows how the number of deaths due to heart disease varies with wine consumption, in different countries. Based on the values of R^2 and correlation reported on the top of the plot, can we conclude that drinking more wine reduces the risk of heart disease?
From the figure, we can see that $R^2 = 0.71$ and $R = -\sqrt{R^2} = -0.843$. Thus, we can conclude that drinking more wine is strongly correlated with reduced heart disease, but we cannot necessarily say that drinking more wine *reduces* the risk of heart disease. Correlation is not equal to causation. Indeed, there may be other differences in the behavior of the people in the various countries that



really explain the differences in the heart disease death rates, such as diet, exercise level, stress level, social support structure and so on.

3- **(6 points)** *[You can use Python]* The Agriculture data contains data from 14 countries. The first column shows the amount of diseased crops (diseased) in million tons, that needs to be destroyed. The second column shows the amount of pesticide in tons (pesticide), and the third column shows the percentage of sunny days in a year (sun).

(a) Report β_0, β_1 for two *linear* classifiers that model: (i) diseased based on pesticide, and (ii) pesticide based on sun.

Diseased vs. pesticide: $\beta_1 = 1.47$, $\beta_0 = -30.83$

pesticide vs. sun: $\beta_1 = -0.681$, $\beta_0 = 113.985$

(b) Report R^2 for the above classifiers and explain the relationships between diseased, pesticide, and sun. Explain why it seems that the disease rate increases as more pesticide is used?

Diseased vs. pesticide: $R^2 = 0.658$. This shows a moderate positive association between the variables, possibly suggesting that as pesticide increases, so too does disease.

Pesticide vs. sun: $R^2 = 0.639$. This shows a moderate (negative) association between the variables, possibly suggesting that as sun increases, pesticide decreases.

Higher values of sun tend to be associated with lower values of both pesticide and disease, so low values of pesticide and disease tend to occur together. Similarly, lower values of sun tend to be associated with higher values of both pesticide and disease, so high values of pesticide and disease tend to occur together. Sun is a lurking variable here and is likely the real driver behind crops disease rates (lower amount of sunlight lead to increase in disease rate, which lead to an increase in the amount of pesticide).

4-**(15 points)** *[You can use Python].* The Experiment dataset containing a thousand (x, y) data points, from a scientific experiment.

(a) Fit a linear model to the data and compute β_0, β_1.

x vs. y: $\beta_0 = 5.006$, $\beta_1 = 0.0998$

(b) Compute and interpret the R^2 value. Does it show a strong linear relation between x and y?

$R^2 = 0.243$, indicating that 24.3% of the variation in x is explained by y. This shows a weak linear relationship between the 2 variables.l

(c) Conduct the test $H_0 : \beta_1 = 0$ (reject the null hypothesis if the p-value for β_1 is less than 0.05). What is your conclusion?

Running the dataset through statsmodels.OLS and printing out the summary, we get that β_1 has a standard error of 0.006 and a t value of 17.897. The value $P(t > 17.897)$ for a t-distribution with $n - 2 = 1000 - 2 = 998$ degrees of freedom is equal to 0, which is less than the p-value threshold of 0.05, so we reject the null hypothesis. We conclude that there is a significant relationship between x and y.

(d) Calculate a 95% confidence interval for β_1, using $\beta_1 \pm 2 \times SE(\beta_1)$, and interpret your interval. Suppose that if $\beta_1 \geq 1$, then we consider it to be meaningfully different from 0, in our research. Does the 95% confidence interval suggests that β_1 is meaningfully different from 0?

95% CI $= [0.0998 - 2 * 0.006, 0.0998 + 2 * 0.006] = [0.0878, 0.1118]$. From this, we see that the 95% CI does not suggest that β_1 is meaningfully different from 0.

(e) Summarize the contradiction you've observed in parts (c) and (d). What is causing the contradiction, and what would you recommend we should always do while analyzing data?

The large sample size results in a sample slope that is significantly different from 0, but not meaningfully different from 0. The scatter plot, which should always accompany a simple linear regression analysis, illustrates this.

5- **(10 points)** *[You can use Python]* The Volcano dataset contains 21 consecutive volcanic eruptions. Use a linear model to predict the time until the next eruption (next), given the duration of the last eruption (duration).

(a) Compute and interpret the R^2 value.

$R^2 = 0.749$, indicating that 74.9% of the variation in the time until the next eruption is explained by eruption duration.

(b) If the duration of the last eruption was 3 minutes, obtain a 95% prediction interval for the time until the next eruption occurs, and interpret your prediction interval.

Using get_prediction(input).summary_frame(alpha=0.05), where input is 3, we get a 95% confidence interval of [47.237, 73.529] minutes. This means that we are 95% confident that the true time until the next eruption will be between 47.237 and 73.529 minutes.

(c) Suppose you can only wait 60 minutes for the next eruption to occur. Can you make a decision based on the above prediction interval?

No, you cannot make a decision between the interval includes values both below and above 60 minutes. As a result, you have no idea if the volcano will erupt in the next 60 minutes.

5 Interpretation of Coefficients in Linear Regression

(a) **[15 points]** Suppose that we want to model the effect of sunlight on the growth of three different types of plants. Suppose we are expecting a linear growth-response over a given range of sunlight, and hence we can model the outcome Y (amount of growth) as a linear function of the sunlight X_1 and the plant type X_2. How do you model the effect of sunlight on the growth of different plant types? How do you interpret each coefficients in your model?

We can build a linear regression model. Notice that X_2 is a categorical feature with three levels, we can use one hot encoding to encode this feature. $X_2^1 = 1$, if we have plant type 1 otherwise $X_2^1 = 0$. $X_2^2 = 1$, if we have plant type 2 otherwise $X_2^2 = 0$. Furthermore, since we are going to study the effect of sunlight on the growth of different plant types, we may consider an interaction term between X_1 and X_2. Therefore, we have our model as

$$Y = \beta_0+\beta_1 X_1+\beta_2 X_2^1+\beta_3 X_2^2+\beta_4 X_1 X_2^1+\beta_5 X_1 X_2^2+\epsilon = \begin{cases} \beta_0 + (\beta_1 + \beta_4)X_1 + \beta_2 + \epsilon & \text{if plant type 1} \\ \beta_0 + (\beta_1 + \beta_5)X_1 + \beta_3 + \epsilon & \text{if plant type 2} \\ \beta_0 + \beta_1 X_1 + \epsilon & \text{otherwise} \end{cases}$$

β_0 represents the expected amount of growth for plant type 3 if we do not have any sunlight.

β_1 indicates the expected change in the amount of growth for plant type 3 per unit change of the amount of sunlight.

β_2 represents the expected difference of the amount of growth between plant type 1 and plant type 3 when there is no sunlight.

β_3 represents the expected difference of the amount of growth between plant type 2 and plant type 3 when there is no sunlight.

β_4 represents the expected additional difference of the amount of growth between plant type 1 and plant type 3 per unit change of the amount of sunlight when we have sunlight.

β_5 represents the expected additional difference of the amount of growth between plant type 2 and plant type 3 per unit change of the amount of sunlight when we have sunlight.

(b) **[5 points]** How do you test the null hypothesis for each independent variable X_1 and X_2 to indicate if they have a significant correlation with the dependent variable Y?

We can perform a hypothesis testing for each parameter. To achieve this, we can use bootstrap to compute the mean and standard errors for the given parameter. Then we can compute the t-statistic by the ratio of the obtained mean and the standard errors. Finally, we can compute the p-value and compare it to the predefined significant level. If the p-value is less than the significant level, then we can conclude that the given independent variable has a significant correlation with the dependent variable Y.

6 Model Evaluation

(5 points) You have a dataset where the only y values are 0 or 1 (a binary classification problem). Out of the 100 data points in this dataset, there is a minority group of 5 data points with $y = 0$

and the rest of data points have $y = 1$. What is the issue if you randomly sample 80% of the data points as your training data? What evaluation strategy you should use to address this?

If we do a naive 80/20 train-test split, the vast majority of points (or perhaps even all the points) will have $y = 1$, and so the model will most likely learn to simply predict $y = 1$. Then, when we evaluate the model on the test set, we will get a misleading accuracy because the test set will likely contain only a few $y = 0$ data points too. A better evaluation strategy would be to use K-fold cross validation so that we can get an average of scores across all splits. Alternatively we can also use a different classification metric, such as precision and F1 score, which is better for imbalanced datasets.

Complete Derivation for Question 3b

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)(-x_i)) = 0$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$-n\bar{x}\bar{y} + \sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$-n\bar{x}\bar{y} + n\bar{x}\bar{y} - n\bar{x}\bar{y} + \sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$-\sum_{i=1}^{n} \bar{x} y_i + \sum_{i=1}^{n} \bar{x}\bar{y} - \sum_{i=1}^{n} x_i \bar{y} + \sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$\sum_{i=1}^{n} (-\bar{x} y_i + \bar{x}\bar{y} - x_i \bar{y} + x_i y_i) = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x})$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \beta_1 (-n\bar{x}^2 + \sum_{i=1}^{n} x_i^2)$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \beta_1 (-2n\bar{x}^2 + n\bar{x}^2 + \sum_{i=1}^{n} x_i^2)$$

Data Science Fundamentals
(CS M148)

Homework 1 Solutions

Submit to Gradescope
due: January 26, 2pm

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \beta_1(-2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2 + \sum_{i=1}^{n}x_i^2)$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \beta_1\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$