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 dining hall has been given the task of determining how students feel about this year's menus. Your friend wants to complete the task by scraping Reddit for keywords related to UCLA food and then run them through a model that can do sentiment analysis. The given model can determine if the text contains positive, negative, or neutral sentiments. Does your friend's data collection method exhibit any selection bias? Explain each kind of bias you give in the context of this situation. (Refer to Week 1 Lecture 2, slide 39 for a list).

Potential biases:

- Voluntary bias: only people that want to talk about the menu will comment in Reddit
- Under-coverage bias: not everyone uses Reddit
- Non-response bias: see voluntary bias
- Response bias: people that comment on Reddit tend to feel strongly (either really like it or dislike it)

Other potential issues:

- There might be Reddit comments about the menu 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)** Long since 2018, companies have started to explore the potential of AI as the recruiter for their hiring process, but AI recruiters were faced with many problems at that time and the idea was eventually scrapped due to bias issues, especially against women.

(1) Explain why the tool was discriminating against women? (2) The developer decides to drop the gender in their data. Would this eliminate the bias? Why?

(i) The training data for ”successful” candidates were predominantly male since males were overrepresented in the company'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 who 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 training data set $\{(x_i, y_i)\} = \{(0, 4), (1, 2), (2, 3), (3, 1), (4, 0)\}$, as shown in Figure 1.

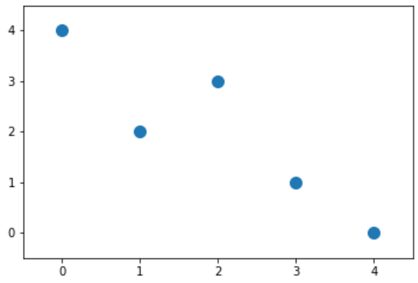


Figure 1: Figure for 2(a)

(a) **(5 points)** Based on the given training data points, draw the KNN predictive regression for $x \in [0, 4]$ using 1-NN, 3-NN, 5-NN regression respectively. You can simply draw with pen and paper. The line does not have to be precise for fractional numbers (for example, $y = 1/3$), roughly draw it and denote on plot what value it should take.

The cutoff value for $k = 1$ is at $1/2$, $3/2$, $5/2$ and $7/2$. The cutoff value for $k = 3$ is at $3/2$ and $5/2$.

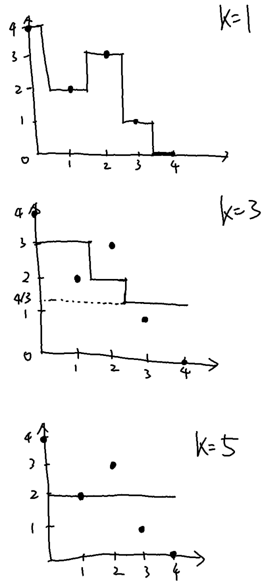


Figure 2: Answer for 2(a)

(b) **(3 points)** Given a test data set $\{(0.3, 3), (1.8, 2), (3.8, 1)\}$, use MSE (Mean Squared Error) to evaluate the performance of 1-NN, 3-NN and 5-NN. Which value of K (choosing from $\{1, 3, 5\}$) is the best? (**Hint.** For dataset of size $n$, we have MSE $= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$, where $f(x_i)$ is the model's prediction.)

$\text{MSE}_{K=1} = 1$, $\text{MSE}_{K=3} = \frac{1}{27}$ and $\text{MSE}_{K=5} = \frac{2}{3}$. $K = 3$ is the best.

(c) **(3 points)** Now let's consider a more general case. Given a training dataset of $n$ data points: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. What will be the $R^2$ score for KNN regression using $K = 1$ on the training data? What about $K = n$? What are the problems with each of these KNN models ($K = 1$ and $K = n$)? (**Hint.** $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.)

For $K = 1$ we have $R^2 = 1$ and for $K = n$ we have $R^2 = 0$. 1-NN overfits the training data and $n$-NN underfits the training data.

## 3 Linear Regression: goodness of fit & Interpretation

1- **(6 points)** US population was around 9 million in 1820, 40 million in 1870, 92 million in 1910, 151 million in 1950, and 281 million in 2000.

(a) The closed-form solution of linear regression with an MSE loss is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Use the formula to fit the above data. What will the population be like in 2010 under this model?

$\beta_0 = -2,732,678,350$ $\beta_1 = 1,490,722$

(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 0.9323. That is, only 6.77% of the variation in the U.S. population is left to explain after taking into account the year in a linear way. However, from the $R^2$ value alone, we cannot conclude if the regression line fits the data well or not.

(c) Plot the residuals versus year. Do you think this is a good model? Why

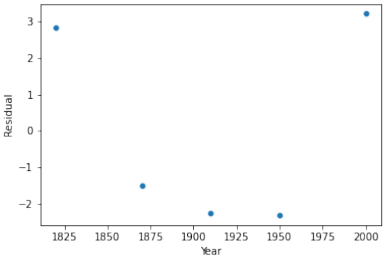
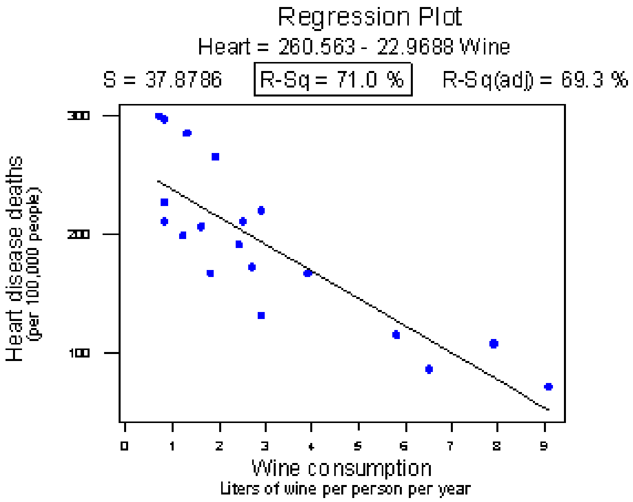


Figure 3: Residual plot for 3-1(c)

This is not a good model. 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. Is there a strong correlation between heart disease and wine consumption? Can we conclude that drinking more wine will reduce the risk of heart disease? Explain your reasoning.



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 Income Data contains data from 14 individuals. The first column shows the average income per year (Income). the second column shows the average spending per year (Consumption), and the third column shows the number of years of working experience (Experience).

(a) Report $\beta_0, \beta_1$ for two *linear* classifiers that model: (i) consumption based on income, and (ii) income based on working experience.

consumption vs. income: $\beta_1 = 4.27$, $\beta_0 = 0.62$

pesticide vs. sun: $\beta_1 = 7.58$, $\beta_0 = 35.4$

(b) Report $R^2$ for the above classifiers and explain the relationships between consumption, working experience, and income. Analyze the potential reason behind this.

consumption vs. income: $R^2 = 0.582$. This shows a moderate positive association between the variables, possibly suggesting that as income increases, so too does consumption.

income vs. experience: $R^2 = 0.603$. This shows a moderate positive association between the variables, possibly suggesting that as experience increases, income decreases.

Higher values of income tend to be associated with higher values of both experience and consumption, so high values of experience and consumption tends to occur together. The reason could be that more experienced people are more likely to get high-paid job, which in turn makes them able to affor more expensive consumptions. Reasonable answers would suffice.

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 $\beta_0, \beta_1$.

x vs. y: $\beta_0 = 0.0992$, $\beta_1 = 4.528$

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

$R^2 = 0.2107$, indicating that 21.07% of the variation in $x$ is explained by $y$. This shows a weak linear relationship between the 2 variables.

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

Running the dataset through statsmodels.OLS and printing out the summary, we get that $\beta_1$ has a standard error of 0.006 and a t value of 16.323. The value $P(t > 16.323)$ 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 $\beta_1$, 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 suggest that $\beta_1$ is meaningfully different from 0?

95% CI $= [0.0992 - 2 * 0.006, 0.0992 + 2 * 0.006] = [0.0872, 0.1112]$. From this, we see that the 95% CI does not suggest that $\beta_1$ 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?

in (c), we concluded that there is a meaningful relationship between the response variable and the explanatory variable, but in (d) we observed no such meaningful relationship. The contradiction is caused by different evaluation criteria. In (c), we set a p-value of 0.05 to determine significance, while in (d), we picked a value to determine the cutoff for if $x$ is significant. Thus, when analyzing data, we should always establish clear evaluation criteria first.

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) Is the linear model a good model? Analyze your result using $R^2$.

Based on $R^2$ alone, the linear model is a good fit here. $R^2 = 0.749$, indicating that 74.9% of the variation in the time until the next eruption is explained by eruption duration.

Students may make comments about judging from $R^2$ itself is not sufficient to judge if a model is good or not, which is a fair point. As long as the $R^2$ calculation is correct and there is some explanation on its meaning, the student can get the full score.

(b) If the duration of the last eruption was 5 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 5, we get a 95% confidence interval of [65.986, 93.941] minutes. This means that we are 95% confident that the true time until the next eruption will be between 65.986 and 93.941 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?

Yes, you can make a decision since both ends of the intervals are above 60 minutes. You cannot see the eruption in the next 60 minutes and you should just leave now.

## 4 Interpretation of Coefficients in Linear Regression

Suppose that we want to model the market sales of fish in a fish market on the weight of three different species of fish. Moreover, we are expecting a linear growth-response over a given range of weight with the sales. Hence, we want to model the outcome $Y$ (sales) as a linear function of the weight $X_1$ and the fish specie $X_2$. There is no ordinal relationship between the fish species.

(a) **(5 points)** As $X_2$ is a categorical feature, we need to first convert it through encoding. Which of the following encoding will be more preferable? Explain your reasoning.

(1) Create one variable $X_2 = \{1, 2, 3\}$. Specifically, let $X_2 = 1$ if fish species is $A$, $X_2 = 2$ if fish species is $B$, and $X_2 = 3$ if fish species is $C$.

(2) Create three indicator variables $X_2^A$, $X_2^B$ and $X_2^C$. Specifically, let $X_2^A = 1$ if fish species is $A$ and 0 otherwise. $X_2^B$ and $X_2^C$ are encoded similarly.

(2) one-hot encoding will be more preferable. Since the categorical feature "fish species" does not have ordinal relationship, it will be better to use one-hot encoding and will be easier to interpret.

(b) **(5 points)** Based on the encoding you chose, how do you model the weight of the fish on the sales of different fish species? **Hint.** Use $\beta_0$, $\beta_1$, $\beta_2$, ... to denote the coefficients and write the model in the form of $Y = \beta X + \ldots$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^A + \beta_3 X_2^B + \beta_4 X_2^C + \beta_5 X_1 X_2^A + \beta_5 X_1 X_2^B + \beta_6 X_1 X_2^C + \epsilon$$

Also correct:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^A + \beta_3 X_2^B + \beta_5 X_1 X_2^A + \beta_5 X_1 X_2^B + \epsilon$$

Coefficients can take any notations, no need to strictly be as written. For example, $\beta_2^A$ is also correct.

(c) **(10 points)** How do you interpret each coefficients in your model? Your answer should include interaction terms.

When the fish species is A,

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5)X_1 + \epsilon$$

- $\beta_0$ represents the base sale for all fish (when weight is 0).
- $\beta_1$ represents the expected change of growth of sale for all fish per unit change of the amount of weight.
- $\beta_2$ represents the base sale for fish type A.
- $\beta_5$ represents the expected additional change of the amount of growth of sale per unit change of the amount of weight for fish type A.

Others are similar.

## 5 Model Evaluation

You have a dataset where the label $y$ takes value either 0 or 1 (a binary classification problem). Suppose the dataset consists of $1,000$ data points, with 10 being negative (i.e. $y = 0$). The rest of the 990 data points have $y = 1$.

(a) **(3 points)** If we consider a baseline model that predicts $y = 1$ for all data, what will be its accuracy on the dataset? Do you think accuracy will be a good evaluation for this dataset? If not, what will be a better evaluation metric? Briefly explain your reasoning.

Accuracy is 0.99. It is not a good evaluation metric when the dataset is imbalanced.

For the better evaluation metric, this question in particular can be a bit tricky. False positive rate will be a good evaluation metric. You can also check the confusion matrix or ROC. Alternatively, you want to switch the labels (0 to 1 and 1 to 0) and use f1 score/recall. Otherwise, the three evaluation metircs (precision, recall, f1) will all give results similar to accuracy. This question can be viewed as a special case where the four evaluation metrics all have problems. For this is not covered, any answer on evaluation metric other than accuracy will receive full credit.

(b) **(2 points)** What is the problem with the given dataset? Other than choosing a different evaluation metric, propose one method that can address the problem.

Dataset is imbalanced. Any method that works will receive full credit. For example, downsampling the majority, generating more negative data through augmentation, consider weight balancing in the loss function, ect.