Data Science Fundamentals (CS M148) Submit to Gradescope
released: January 27, 2023 **Homework 1** due: February 3

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 key words 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).

(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?

2 KNN regression

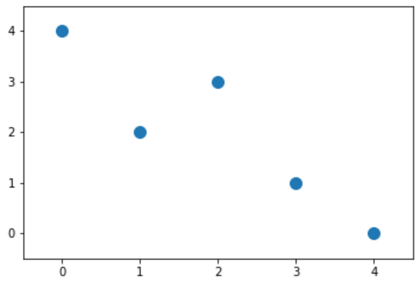


Figure 1: Figure for 2(a)

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.

(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.

(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? You can either report the MSE in fractional number or round to two decimal places. (**Hint.** For dataset of size n, we have MSE $= \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2$, where $f(x_i)$ is the model's prediction.)

(c) **(3 points)** Now let's try to understand more about KNN and R^2 score. Consider a more general case, we have a training dataset of n data points: $\big\{(x_1, y_1), \ldots, (x_n, y_n)\big\}$. 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 $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.)

3 Linear Regression: goodness of fit & Interpretation

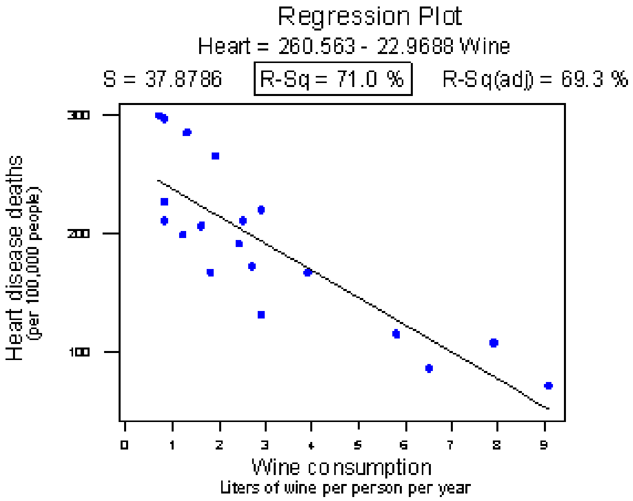
1- **(6 points)** US population was around 9 million in 1820, 40 million in 1870, 92 million in 1910, 151 million in 1950, and 281 million in 2000.

(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?

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

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

2- **(4 points)** The following plot shows how the number of deaths due to hearth 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.



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 working experiences(Experience),

(a) Report β_0, β_1 for two *linear* classifiers that model: (i) consumption based on income, and (ii) income based on working experience.

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

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.

(b) Is there a strong linear relation between x and y? Explain your reasoning.

(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). Analyze your result

(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?

(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?

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.

(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.

(c) If you need to leave in 50 minutes, can you determine if you can see the eruption based on the data? Explain your reasoning.

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.

(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 β_0, β_1, ... to denote the coefficients and write the model in the form of $Y = \beta X + \ldots + \epsilon$.

(c) **(10 points)** How do you interpret each coefficients in your model? Your answer should include interaction terms (for example, $\beta X_i X_j$). **Hint.** When doing interpretation, try to discuss by cases. For example, when the fish species is A/B/C.

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.

(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.