Data Scienve Fundomentals (CS M148)

HW1

a) Voluntary bias

Under-coverage bias

non-response bias

convenience bias

response bias

over-coverage bias

Voluntary bias is exhibiting in the data collection method. People who like to dine in UCLA dining hall are more likely to comment to Reddit platform. There can be also some negative comments, but I donot think there will be enough neutral samples.

Undercoverage bias is also exhibiting since most of the students using the dining hall is excluded. Most of the students are not commenting about UCLA food and the number drastically decreases when you try to reach people who comment UCLA food to reddit. Note that, we can only sample students, therefore study won't be covering all population.

Overcoverage bias can also happen in our context. nicknames in reddit are not unique for a person, so same person can easily comment multiple times with different nicknames.

b) 1) The tool is discriminating against woman due to the number of woman samples. Those models are trained with resumes, therefore since number of man applicants are higher, model discriminated against woman.

2) It can eliminate the bias to some extent. Firstly, it can only eliminate bias only caused by the gender section in the resumes. However, people also write member of woman club, and much more which indicate they are woman, therefore it is not easy to fully eliminate gender and race bias since they are closely related with the resume of human.

2) KNN Regression

[0,4)
[1,2)
[2,3)
[3,1)
[4,0)

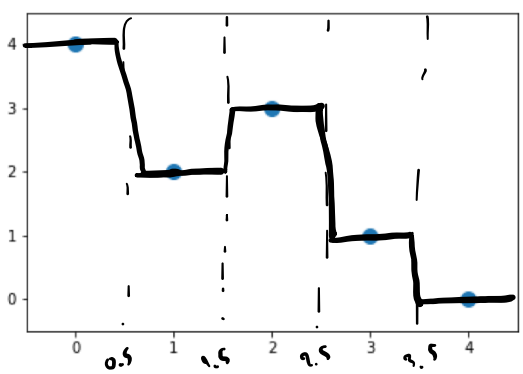


Figure 1: Figure for 2(a)

1-NN Result

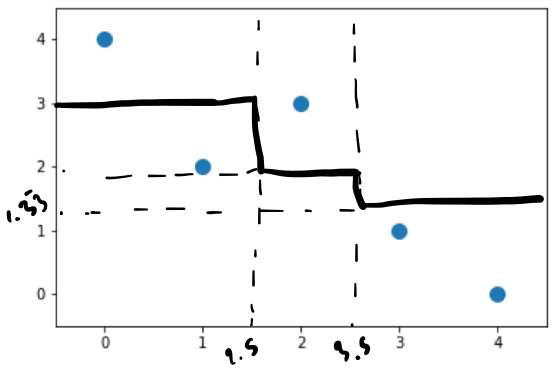


Figure 1: Figure for 2(a)

3-NN Result

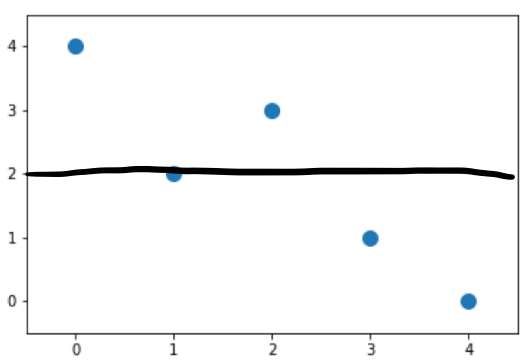


Figure 1: Figure for 2(a)

5-NN Result

b) Test = $\{(0.3,3), (1.8,2), (3.8,1)\}$

Prediction for 1-KNN:

prediction:	actual:
$(0.3,4)$	$(0.3,3)$
$(1.8,3)$	$(1.9,2)$
$(3.9,0)$	$(3.8,1)$

$$MSE = \frac{1}{3}(1^2 + 1^2 + 1^2) = \boxed{1}$$

Prediction for 3-KNN:

prediction	actual
$(0.3,3)$	$(0.3,3)$
$(1.8,2)$	$(1.9,2)$
$(3.8,\frac{4}{3})$	$(3.8,1)$

$$MSE = \frac{1}{3}\left(0^2 + 0^2 + \left(\frac{1}{3}\right)^2\right) = \frac{1}{27}$$

Prediction for 5-KNN:

prediction	actual
$(0.3,2)$	$(0.3,3)$
$(1.8,2)$	$(1.9,2)$
$(3.8,2)$	$(3.8,1)$

$$MSE = \frac{1}{3}(1^2 + 0^2 + 1^2) = \boxed{\frac{2}{3}}$$

3-KNN model yields to the least MSE error which is $\boxed{\frac{1}{27}}$. $\boxed{K=3}$

c) when $K=1$, algorithm gives the value of the nearest neighbor, thus overfitting to the data.

R^2 for training data will be $\boxed{1}$, whereas R^2 for testing data will be low. Therefore, model overfits to the data. $\sum_{i=1}^{\hat{}} (y_i - f(x_i))^2$ for training set will be zero. Thus, $R^2 = 1$

when $K=N$, algorithm gives the average of all points to every prediction, thus underfitting to the data.

R^2 for training data will be $\boxed{\text{zero}}$, whereas R^2 for testing data will be also low. Therefore, model underfits to the data.

$$\sum_{i=1}^{\hat{}} (y_i - f(x_i))^2 = \sum_{i=1}^{\hat{}} (y_i - \bar{y})^2 \text{ for training dataset,}$$

Thus $R^2 = 1 - 1 = \boxed{0}$

3) Linear Regression:

1) (1820, 9)
(1870, 40)
(1910, 92)
(1950, 151)
(2000, 281)

$$\bar{x} = \frac{1820 + 1870 + 1910 + 1950 + 2000}{5} = 1910$$

$$\bar{y} = \frac{9 + 40 + 92 + 151 + 281}{5} = 114.6$$

$$\beta_1 = \frac{(1820 - 1910)(9 - 114.6) + (1870 - 1910)(40 - 114.6) + (1910 - 1910)(\ldots) + (1950 - 1910)(151 - 114.6) + (2000 - 1910)(281 - 114.6)}{(-90)^2 + (-40)^2 + 0^2 + 40^2 + 90^2}$$

$$\beta_1 = \frac{9504 + 2984 + 1456 + 14976}{19600} = \frac{28920}{19600} = \boxed{1.49072}$$

$$\beta_0 = 114.6 - (1.49072)(1910) = \boxed{-2732.67835}$$

$$f(2010) = (1.49072)(2010) - 2732.67835 = \boxed{263.67 \text{ million}}$$

Population will be 263.67 million people by 2010

b)

$f(1820) = -19.56$
$f(1870) = 54.97$
$f(1910) = 114.5999$
$f(1960) = 174.22$
$f(1000) = 248.76$

$$R^2 = 1 - \frac{(9+19.56)^2 + (40-54.97)^2 + (92-114.59)^2 + (151-174.22)^2 + (281-248.76)^2}{(9-114.6)^2 + (40-114.6)^2 + (92-114.6)^2 + (151-114.6)^2 + (281-114.6)^2}$$

$$R^2 = 1 - \frac{(815.95 + 224.13 + 510.75 + 539.58 + 1039.09)}{11151.36 + 5565.15 + 510.76 + 1374.96 + 27666.8}$$

$$= 1 - \frac{3179.52}{46241.2} = \boxed{0.9323}$$

$R^2 = 0.9323$ is a good score which indicates the fit is a good model. However, testing data should be used to test the performance of the model. Data may have overfitted.

c)

$e(1820) = 28.565$

$e(1870) = -14.971$

$e(1910) = -22.6$

$e(1950) = -23.129$

$e(2000) = 32.235$

Residuals are scattered around 0, therefore our fit is a good model for linear regression. However, better fits can be acquired using multiple samples, and multiple features.

2) There is a negative correlation between wine consumption and heart disease deaths, however R^2 value is 0.71 which is not too high, therefore we should not conclude there is negative high correlation, however one can not deny the correlation seen from the plot. Note that, samples are very inadequate when wine consumption is bigger than 4. We should not conclude drinking more wine will reduce the risk of death from heart disease since

there are other features that are more strongly correlated with heart disease such as genetics.

3)

i) consumption based on income

a) $\beta_1 = 0.62$ $\quad y = 4.270 + 0.62x$

$\beta_0 = 4.270$

b) $R^2 = 0.58$, although consumption can not be fully explained with income. As income increases, people tend to consume more and easily. This is why slope is positive.

ii) income based on working experience.

a) $\beta_1 = 7.58$ $\quad y = 35.4 + 7.58x$

$\beta_0 = 35.4$

b) $R^2 = 0.60$, although income can not be fully explained with working experience. As working experience increases, people earn more money. This is why slope is positive.

4)

a) $\beta_0 = 4.5285$

$\beta_1 = 0.09923$

$R^2 = 0.21$

b) There is no strong relationship between x and y since R^2 is low.

c) p_value $= 0 < 0.05$, therefore we reject the null hypothesis

d) CF for $\beta_1 = [0.087309, 0.111170]$

CF for $\beta_0 = [4.395932, 4.661093]$

I have used statsmodel.api for p value and CF.

If we say β_1 is different from zero, when $\beta_1 \geq 1$, then CF for β_1 indicates that β_1 is not meaningfully different than zero.

e) c is right, β_1 is a scalable value, therefore, when data is not standardized β_1 can take any value such as $10^{-15}, 10^{-100}$. This does not mean that there is no correlation. The scale of input and output should always be considered while doing analyses or before linear fit data should be transformed using standard scaler. R^2 and p value are not effected from the scale since they are unitless. (c) suggests that β_1 is different from zero and (d) suggests that β_1 is not different from zero. Since β_1 is scalable d is wrong, c is right.

5) a) $\beta_0 = 31.0181$

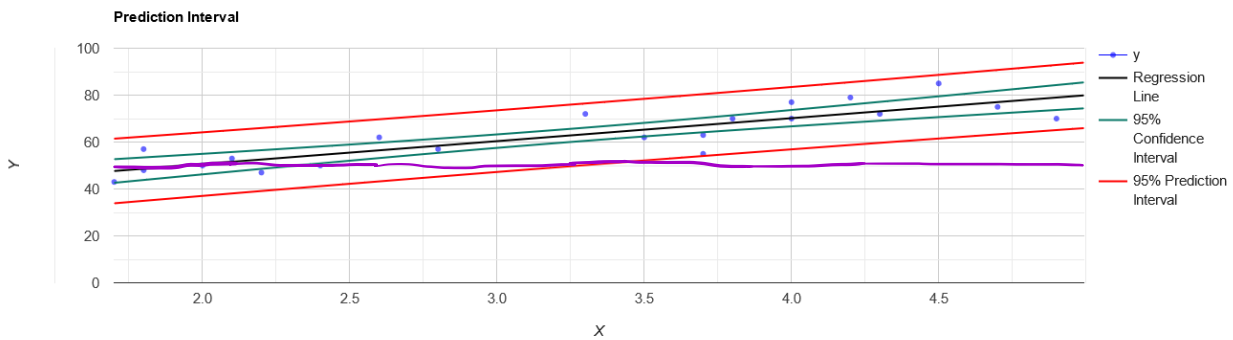
$\beta_1 = 9.7901$

$R^2 = 0.749$

Since R^2 is high, and not exactly one, our fit is a good model. Model did not over-fit nor under fit.

b) %95 prediction interval for 5 min duration is [66.754, 93.173].
Predicted value is 79.96
If the duration of last eruption is 5 mintes, next eruption will happen between 66.754 and 93.173 minutes later with significance. Therefore, it is very likely that an eruption will happen around 79 minutes.

c)



Regression line, confidence intervals and prediction intervals for this question is shown above.

Purple line is a horizontal for 50 minutes. You are very likely to observe an eruption in 50 minutes if the last eruption did not last more than 3.5 minutes. The likeliness also increases when the last eruption lasted less amount of time.

4) Interpretation of Coefficients in Linear Regression.

a) Both methods have drawbacks but

One Hot encoding (2) is generally prefered over label encoding (1) when variable is not ordinal.

The main advantage of one hot encoding is that it does not force features to be ordinal, whereas label encoding forces features to be ordinal. therefore, if the categorical variable like in the fish case is not ordinal, we should prefer one-hot encoding (2)

However, one-hot encoding has disadvantages, firstly, size of feature spare increases exponentionally. secondly, dummy variables can introduce multicollinearity which may reduce the performance of the model, due to overfitting.

b)

Without interaction:

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

With interaction:

ii) $Y = \beta_1 X_1 + \beta_2 X_2^A + \beta_3 X_2^B + \beta_4 X_2^C + \beta_5 X_1 X_2^A + \beta_6 X_1 X_2^B + \beta_7 X_1 X_3^B + \beta_0 + \epsilon$

i and ii model are different for each fish type

i) $Y = \beta_1 X_1 + \beta_2 + \beta_0 + \epsilon$ when fish = A

$Y = \beta_1 X_1 + \beta_3 + \beta_0 + \epsilon$ when fish = B

$Y = \beta_1 X_1 + \beta_4 + \beta_0 + \epsilon$ when fish = C

ii) $Y = (\beta_1 + \beta_5) X_1 + \beta_2 + \beta_0 + \epsilon$ when fish = A

$Y = (\beta_1 + \beta_6) X_1 + \beta_3 + \beta_0 + \epsilon$ when fish = B

$Y = (\beta_1 + \beta_7) X_1 + \beta_4 + \beta_0 + \epsilon$ when fish = C

c) Interpretation for without interaction:

If we only include the $\beta_j X_2^{A|B|C}$, we are only changing the constant level of the fit, therefore our fit is only giving an extra constant term for each type of fish.

Interpretation for with interaction:

Things get more complicated in the case of fitting with interaction term since our interaction term is also capable to change slope of the fit for each fish type.

β_1 is obtained when we do regression without X_2 variable. $\beta_2, \beta_3, \beta_4$ are changes made to the constant term by X_2 variable. $\beta_5, \beta_6, \beta_7$ are changes made to the slope term by X_2 variable. Lastly, β_0 is obtained as a constant term when we do regression without X_2 variable.

β_2 and β_5 are added when the fish type is A.

β_3 and β_6 are added when the fish type is B.

β_4 and β_7 are added when the fish type is C.

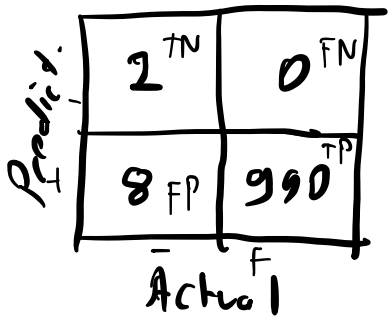
5)

990 → 1

10 → 0

a) accuracy is not a good metric to use when there is class imbalance.

Suppose after training a classifier, we have obtained the following conf. matrix.



Accuracy is $\frac{992}{1000} = 0.992$ which is very high.

However this value is misleading when we check precision, precision score is 0.2 which is very low.

Therefore, when there is class imbalance one should check other metrics such as precision, recall.

F2 score is the harmonic mean of precision and recall. F2 score can be used to evaluate the model performance in the case of class imbalance since it captures both properties of precision and recall. F2 score eliminates the possibility of observing low recall and high precision or high recall and low precision.

For the given example F2 score can be computed as:

precision = 0.2
recall = 1

$$\frac{2(0.2 * 1)}{1.2} = \frac{1}{3} = 0.\bar{3}$$

F2 score is $\frac{1}{3}$ which is low indicating model performance is bad.

6) As mentioned before, the problem of the dataset is the unequal number of labels which is called as class imbalance. The greatest method will be introducing new negative samples. However, if we are not able to add new negative samples, weighted sampling can be performed. The class with less samples will be sampled more, whereas the class with more samples will be sampled less.