Homework 5 - Quiz Part

Started: Feb 27 at 7:11pm

Quiz Instructions

Homework 5 contains two parts: the quiz part and the programming part. You need to submit both parts.

Rules for Quiz Part:

- You can only attempt the quiz part one time. Make sure your answers are complete before you click the submit button.
- If you miss the due time, you still can submit within 24 hours after the due time, with a penalty of 40% points.
- Correct answers will be shown 24 hours after the due time.

Question 1

1 pts

Which of the following statement about simple linear regression is NOT correct?

- ○ The regression line always passes through the piont ($\bar{x}, \bar{y}$).
- ○ The simple linear regression can do statistical control of confounding factors.
- ○ Regression R squred equals to the sqaure of correlation coefficient.
- ◉ The number of predictors p = 1

Question 2

1 pts

The relationship between number of beers consumed (x) and blood alcohol content (y, %) was studied by using least squares regression. The following

regression equation was obtained from this study:

y= -0.0127 + 0.0180x

The above equation implies that:

- ○ Each beer consumed increases blood alcohol by 1.27%.
- ○ Each beer consumed increases blood alcohol by an average of amount of 1.8%.
- ○ On average it takes 1.8 beers to increase blood alcohol content by 1%.
- ◉ Each beer consumed increases blood alcohol by exactly 0.018.

Question 3

1 pts

Which of the following statements about regression analysis is NOT true?

- ○ Parametric regression models have no assumption regarding the functional form y=m(x) + e.
- ○ Regression intends to summarize observed data as simply and usefully as possible.
- ○ Regression is about estimating relationships between dependent and independent variables.
- ◉ All of the above are true.

Question 4

1 pts

Assume we are doing a multiple linear regression analysis:

$Y = beta_0 + beta_1X_1 + beta_2X_2 + beta_3X_3$

What can you say about their relationship if $beta_1$ is estimated as 2.0?

- ○ The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X_1, without regard to X_2 and X_3.
- ○ The relationship between X_1 and Y is significant.
- ◉ The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X_1, holding X_2 and X_3 constant.
- ○ The estimated average value of Y is 2 when X_1 equals to zero.

Question 5

1 pts

What is the purpose of using the set.seed() function in the following R section?

```
set.seed (200)

x <- rnorm(100)

eps <- rnorm(100, mean=0, sd=0.5)

y <- -1 + 0.5*x + eps
```

- It has no special purpose.
- To make the result reproducible.
- To set the mean of random numbers as 200.
- To generate 200 random numbers.

Question 6

1 pts

Below is a confusion matrix of a classification algorithm:

	Yes	No
Yes	9627	228
No	40	105

Rows are prediction and columns are reference or ground truth. Positive class is Yes. What is the sensitivity of the algorithm?

- 0.3153
- 0.9959
- 0.9732
- 0.7241

Question 7

1 pts

Below is a confusion matrix of a classification algorithm:

	Yes	No
Yes	100	900

No	0	0

Rows are prediction and columns are reference or ground truth. Positive class is No. What is the sensitivity of the algorithm?

- 0.90
- 0
- 1.00
- 0.10

Question 8

1 pts

Which of the following method is NOT appropriate to handle imbalanced datasets?

- Over-sample the majority class.
- Customize the cost function to assign larger penalty to misclassified minority class.
- Use different threshold for prediction.
- Use AUC rather than accuracy to measure performance.

Question 9

1 pts

Which of the following is a non-parametric method?

- kNN
- Linear regression
- Logistic regression

- LDA
- QDA

Question 10

1 pts

Assume we have y, X1, and X2 in a data frame df, where y is a binary variables containing values of 0 and 1. What is the correct R code to analyze the impact of X1 and X2 on y?

- model <- glm(y ~ X1 + X2, data=df)
- **model <- glm(y ~ X2 + X1, family=binomial(link='logit'), data=df)**
- model <- lm(y ~ X1 + X2, data=df)
- None of the above

Question 11

1 pts

Which of the following statement about LDA and QDA is NOT correct?

- If the Bayes decision boundary is non-linear, we expect QDA to performance better on the training set.
- If the Bayes decision boundary is linear, we expect LDA to performance better on the test set.
- If the Bayes decision boundary is non-linear, we expect QDA to performance better on the test set.
- **If the Bayes decision boundary is linear, we expect LDA to performance better on the training set.**
- All of the above are correct.

Question 12

1 pts

Which of the following statement about kNN is correct?

- ○ kNN cannot be used for regression problems.
- ◉ kNN performs poorly when the number of predictors p is large.
- ○ kNN is a parameteric method.
- ○ kNN with k=1 is less flexible than kNN with k=10.

Question 13

2 pts

Explain why a kNN classifier with k=1 usually has zero training error rate (in-sample error rate).

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

When K=1 you will chose the closest training sample to the test sample. Since the test sample is in the training data, it will choose itself as the closest and it will almost never make an error.

p 37 words </>

Question 14

1 pts

Suppose that a customer has a 15% chance of defaulting on her credit card payment. What are the odds that the customer will default?

Edit View Insert Format Tools Table

12pt Paragraph

The odds that the customer will default are 17.6%

p 9 words

Question 15

2 pts

Suppose we have collected a dataset from a machine learning class. There are three variables in the dataset:
x1: hours a student spent in study;
x2: undergrad GPA of the student;
y: a binary variable indicating if the student receives an A in the class.

We conduct a logistic regression be gressing y onto x1 and x2. The

coefficients are estimated as $\beta_0 = -5$, $\beta_1 = 0.1$, $\beta_2 = 1$.

(a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.8 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

Edit View Insert Format Tools Table

12pt Paragraph

1. 0.9427

2. 12

p 4 words

Quiz saved at 8:52pm Submit Quiz