Midterm

Due Oct 31 at 11:59pm **Points** 100 **Questions** 29
Available Oct 17 at 12am - Dec 2 at 11:59pm **Time Limit** None

Instructions

Dear students,

Answer all questions. You are required to work individually on the midterm. The exam is due on October 31st. There is no time limit.

As it is an individual exam, you are (strictly) not allowed to consult with anyone. Also, if you require any clarification(s) on any question, you must consult with the instructor and not another student. Any collaboration will be viewed strictly as honoring the university honor code is paramount!

You can step in and out of Canvas midterm without hitting the submit button, since Canvas saves your previous answers. However, once you hit the submit button, it is final, Canvas will not let you edit your answers.

You have time until the due date.

All questions in this quiz are open book, open notes. You are allowed to search the internet and consult the python tutorial code.

Enjoy!

Cheers,

:)

Jagan

This quiz was locked Dec 2 at 11:59pm.

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	8,504 minutes	98 out of 100

Score for this quiz: **98** out of 100
Submitted Oct 29 at 4:02pm

This attempt took 8,504 minutes.

Question 1

2 / 2 pts

XOR function mappings can easily be classified by decision trees.

Correct!

- (•) True
- ( ) False

Question 2

2 / 2 pts

Discretized values in a decision tree may be combined into a single branch if order is not preserved.

- ( ) True

Correct!

- (•) False

Question 3

2 / 2 pts

Correlations are distorted if the data is standardized.

- ( ) True

Correct!

- (•) False

Question 4

2 / 2 pts

Assume an attribute (feature) has a normal distribution in a dataset. Assume the standard deviation is S and the mean is M. Then the outliers usually lie below -3*M or above +3*M.

- ○ True
- ◉ False

Correct!

Question 5

2 / 2 pts

Jaccard coefficient ignores 00 combinations since it is meant to eliminate skewness when 00 combinations are common and irrelevant.

- ◉ True
- ○ False

Correct!

Question 6

2 / 2 pts

Bias toward selecting an attribute at a node of the decision tree may happen if the attribute has many branches.

- ◉ True
- ○ False

Correct!

Question 7

2 / 2 pts

Linear Regression cannot be applied on every dataset, it is prudent to apply linear regression if the correlation is greater than 0.5 or less than -0.5.

Correct!

- (●) True
- ( ) False

Question 8

2 / 2 pts

Dimensionality reduction helps to eliminate irrelevant attributes or reduce possible noise.

Correct!

- (●) True
- ( ) False

Question 9

2 / 2 pts

Higher level aggregations may have more variations than lower level aggregations.

- ( ) True

Correct!

- (●) False

Question 10

2 / 2 pts

For non-linear relationships, correlations can give correct results.

- True
- Correct! False

Question 11

5 / 5 pts

Correct! **Gini Index**	1- (∑[P(j|t)]^2 for all j)
Correct! **Interactions**	Cannot classify properly
Correct! **Dividing Gain by SplitINFO**	Can overcome disadvar
Correct! **Misclassification Error**	1- (max(P(i|t)) for all i)
Correct! **Underfitting**	Model too simple

Other Incorrect Match Options:

- Causes Variance

Question 12

2 / 2 pts

A continuous attribute range may be split at the point where the GINI index values is _____.

Correct!

least

orrect Answers

minimum

least

the least

lowest

smallest

small

Question 13

2 / 2 pts

The loss function for linear regression is the square of the difference between the original Y value and the _____ Y value.

Correct!

predicted

orrect Answers

predicted

estimated

Question 14

2 / 2 pts

_____ = GINI measure before splitting - GINI measure after splitting.

Correct!

Gain

orrect Answers

gain

Question 15

0 / 2 pts

The process of _____ data before calculating correlations is a prudent way to get good correlations.

ou Answered

standardizing

orrect Answers

removing outliers

Question 16

2 / 2 pts

Decision trees use a _____ approach which often is unable to find the best tree.

Correct!

greedy

orrect Answers

greedy

local optima

Question 17

2 / 2 pts

BoxPlots are centric to median

Answer 1:

Correct!

median

Question 18

2 / 2 pts

Standardization transformation is centric to Mean

Answer 1:

Correct!

Mean

Question 19

2 / 2 pts

The Mean of the transformed data after standardization (z-score calculation) becomes 0 :

Answer 1:

Correct!

0

Question 20

2 / 2 pts

The standard deviation of the new transformed data after standardization (z-score calculation) is 1 :

Answer 1:

Correct!

1

Question 21

2 / 2 pts

Outliers are values outside the range between Q1 - 1.5 * IQR and this Q3 + 1.5 * IQR :

Answer 1:

Correct!

Q3 + 1.5 * IQR

Question 22

2 / 2 pts

Is this statement true? When outliers are important then it is important not to change the current minimum and maximum for normalization.

- False
- True

Correct!

Question 23

2 / 2 pts

Is this statement true? When outliers are not significant then it is important to change the maximum and minimum by subtracting outlier end points from minimum and maximum to get the new minimum and maximum.

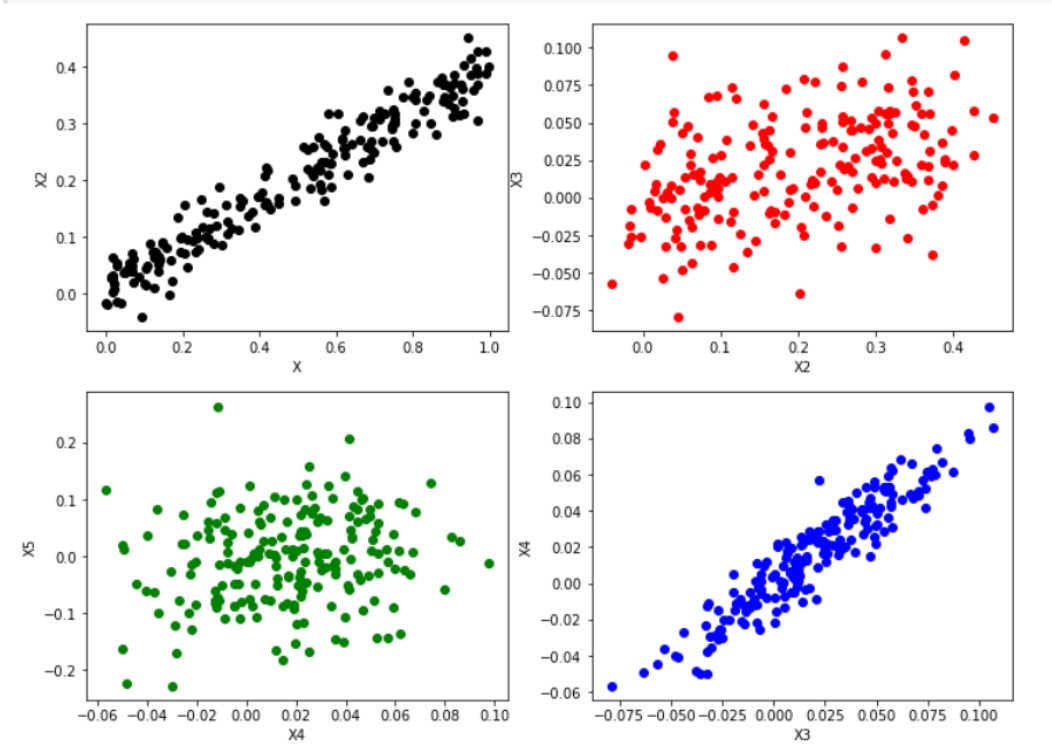
- False
- True

Correct!

Question 24

5 / 5 pts

Analyze the figure and state which one of the following pairs have the highest co-relation and why?



Your Answer:

The pair X4 and X3, or the graph in the lower-right corner with the blue dots is the one with the highest correlation. This graph is positive, in that as the input (*x) increases, the output (*y) also increases. It appears to be linear in that its pattern closely resembles a line, and it is denser when compared to other graphs.

Question 25

10 / 10 pts

Explain your understanding of the equation:

R^2 = SSR/SST = 1 - SSE/SST

what does R^2 (R-squared) describe?

Read this article and provide your summary of the article:

https://statisticsbyjim.com/regression/interpret-r-squared-regression/

(Note: All of you will get full points for this question for answering. Do bit worry about quality. The purpose is: the paper gives you a new perspective of how to look at things.)

Your Answer:

*R2 = SSR/SST = 1 - SSE/SST

R-squared equals the sum of squares regression divided by the sum of squares total, which is also equal to the sum of squares error divided by the sum of squares total.

Basically, we are saying that we can determine r-squared by using either the mean of the response variable or the observed data points interchangeably with the sum of squared differences between individual data points and the mean of the response variable.

*Article Summary:

R-squared is a goodness-of-fit measure for linear regression models, which helps to indicate the percentage of variance in the dependent variable that the independent variables explain collectively. It measures the strength of the relationship between your model and the dependent variable on a 0 – 100% scale, where 0% represents a model that does not explain any of the variation in the response variable around its mean and 100% represents a model that explains all the variation in the response variable around its mean.

Linear regression identifies the equation that produces the smallest difference between all the observed values and their fitted values, and a regression model fits the data well if the differences between the observations and the predicted values are small and unbiased.

Residual plots can expose a biased model by displaying problematic patterns in the residuals. If your model is biased, you cannot trust the results.

R-squared evaluates the scatter of the data points around the fitted regression line. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values.

R-squared is the percentage of the dependent variable variation that a linear model explains.

Usually, the larger the R2, the better the regression model fits your observations.

When a regression model accounts for more of the variance, the data points are closer to the regression line. Please note that r-squared does not indicate if a regression model provides an adequate fit to your data. A good model can have a low R2 value, and a biased model can have a high R2 value!

Regression models with low R-squared values can be perfectly good models for several reasons as some fields of study have an inherently greater amount of unexplainable variation. In these areas, your R2 values are bound to be lower. For example, studies that try to explain human behavior generally have R2 values less than 50%. People are just harder to predict than things like physical processes.

When the regression line consistently under and over-predicts the data along the curve, there is bias, which generally occurs when your linear model is underspecified. To produce random residuals, try adding terms to the model or fitting a nonlinear model.

A variety of other circumstances can artificially inflate your R2, but to get the full picture, you must consider R2 values in combination with residual plots, other statistics, and in-depth knowledge of the subject area.

Question 26

10 / 10 pts

From the given dataset find the mean, standard deviation, 25% - the 25% percentile, 50% - the 50% percentile, 75% - the 75% percentile for the 'Math' and 'English' attribute.

dictionary_data={'Name':['A','B','C','D','E'],'Math':[87,90,51,25,98],'English':[50,68,45,88,14]}

dataframe = pandas.DataFrame(dictionary_data)

Your Answer:

For the 'Math' attribute:

The mean is:

(87 + 90 + 51 + 25 + 98) / 5 = 70.2

The standard deviation is:

(87 - 70.2)^2 = 282.24, (90 - 70.2)^2 = 392.04, (51 - 70.2)^2 = 368.64

, (25 - 70.2)^2 = 2043.04, (98 - 70.2)^2 = 772.84

282.24+...772.84 = 3858.8, 3858.8/(5 - 1) = 964.7

SQRT(964.7) = 31.0596

The 25th percentile is:

0.25 * (5 + 1) = 1.5 ≈ 2

If we order the values, we get:

25, 51, 87, 90, 98

2nd value is 51, so the 25th percentile is 51!

The 50th percentile is:

0.5 * (5 + 1) = 3

If we order the values, we get:

25, 51, 87, 90, 98

3rd value is 87, so the 50th percentile is 87!

The 75th percentile is:

(*This should be 4, since the 75th percentile is the middle of the upper half of the data)

If we order the values, we get:

25, 51, 87, 90, 98

4th value is 90, so the 75th percentile is 90!

For the 'English' attribute:

The mean is:

(50 + 68 + 45 + 88 + 14) / 5 = 53

The standard deviation is:

(50 - 53)^2 = 9, (68 - 53)^2 = 225, (45 - 53)^2 = 64

, (88 - 53)^2 = 1225, (14 - 53)^2 = 1521

9+...1521 = 3044, 3044/(5 - 1) = 761

SQRT(761) = 27.5862

The 25th percentile is:

0.25 * (5 + 1) = 1.5 ≈ 2

If we order the values, we get:

14, 45, 50, 68, 88

2nd value is 45, so the 25th percentile is 45!

The 50th percentile is:

0.5 * (5 + 1) = 3

If we order the values, we get:

14, 45, 50, 68, 88

3rd value is 50, so the 50th percentile is 50!

The 75th percentile is:

(*This should be 4, since the 75th percentile is the middle of the upper half of the data)

If we order the values, we get:

14, 45, 50, 68, 88

4th value is 68, so the 75th percentile is 68!

Use dataframe.describe()

Question 27

10 / 10 pts

Explain Lasso and Ridge Regression. Compare and Contrast Lasso and Ridge Regression.

Your Answer:

Ridge regression is a variant of MLR designed to fit a linear model to the dataset by minimizing the following regularized least-square loss function:

$L\text{ridge}(y, f(X,w)) = \sum_{i=1}^{N} \| y_i - X_i w - w_0 \|^2 + \alpha \left[\| w \|^2 + w_0^2 \right]$

where α is the hyperparameter for ridge regression. Ridge regression can be thought of as a type of regularization technique to reduce model complexity and to prevent over-fitting which may result from simple linear regression. By setting an appropriate value for the hyperparameter, α, we can control the sum of absolute weights, thus producing a test error that is quite comparable to that of MLR without the correlated attributes. In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients. When many predictor variables are significant in the model and their coefficients are roughly equal then ridge regression tends to perform better because it keeps all of the predictors in the model.

Another variation of MLR is lasso regression, which is designed to produce sparser models by imposing 1 regularization on the regression coefficients as shown below:

$L\text{lasso}(y, f(X,w)) = \sum_{i=1}^{N} \| y_i - X_i w - w_0 \|^2 + \alpha \left[\| w \|_1 + | w_0 | \right]$

Like ridge regression, lasso regression can also be thought of as a type of regularization technique to reduce model complexity and to prevent over-fitting which may result from simple linear regression. In lasso regression, the cost function is altered by adding a penalty equivalent to absolute value of the magnitude of coefficients. In cases where only a small number of predictor variables are significant, lasso regression tends to perform better because it's able to shrink insignificant variables completely to zero and remove them from the model.

They are both known as regularization methods because they both attempt to minimize the sum of squared residuals (RSS) along with some penalty term. They both constrain or regularize the coefficient estimates of the model. Ridge regression includes all (or none) of the features in the model. Thus, the major advantage of ridge regression is coefficient shrinkage and reducing model complexity. While lasso regression, along with shrinking coefficients, will feature selection as well.

Question 28

6 / 6 pts

Given the observation table for student who pass/ do not pass the test according to their studying style, concentration level and sleeping habits

below: Draw the decision tree for the below table (you can draw it on paper and upload the picture of the solution).

Test	Study	Sleep	Concentrate	Time	Pass ?
1	Hard	No	High	Long	Yes
2	Less	Yes	High	Less	Yes
3	Don’t study	No	No	Long	No
4	Don’t study	No	High	Long	No
5	Hard	Yes	No	Long	Yes
6	Hard	No	No	Less	No
7	Less	Yes	No	Long	No
8	Less	Yes	High	Long	Yes
9	Less	Yes	No	Less	No
10	Don’t study	No	High	Less	No

Q28.pdf (https://csus.instructure.com/files/15672284/download)

Question 29

10 / 10 pts

Given the dataset of cars with their mileage and cost for cost prediction:

Build a decision tree and attach the picture of you decision tree formed as the solution to the question using hints from tutorial_6. What do you think is special about the data type of the values being predicted in this problem? Do a search and find out why a DecisionTreeRegressor is used in this case (instead of the general purpose DecisionTreeClassifier)? Both examples are in the tutorial code.

```
array([['Toyota Corolla', '40', '20175'],
       ['Ford', '45', '25000'],
       ['Dodge', '62', '35782'],
       ['Chevrolet', '50', '30000'],
       ['Canoo', '57', '34750'],
       ['Tesla', '113', '54000'],
       ['BMW', '70', '36400']], dtype='<U21')
```

Q29.pdf (https://csus.instructure.com/files/15672293/download)

Quiz Score: **98** out of 100