Quiz2 -- Linear Regression and Classification Trees

Due Oct 14 at 11:59pm **Points** 50 **Questions** 24
Available Oct 2 at 11:59pm - Oct 31 at 11:59pm **Time Limit** None
Allowed Attempts Unlimited

Instructions

Dear students,

Answer all questions. You can consult other students. You can take your time until the due date. You have unlimited attempts.

All questions in this quiz are open book. You may work or discuss in groups.

Cheers,

:)

Jagan

This quiz was locked Oct 31 at 11:59pm.

Attempt History

	Attempt	Time	Score
KEPT	Attempt 42	9 minutes	50 out of 50
LATEST	Attempt 42	9 minutes	50 out of 50
	Attempt 41	250 minutes	40 out of 50 *
	Attempt 40	less than 1 minute	2 out of 50 *
	Attempt 39	3 minutes	2 out of 50 *
	Attempt 38	3 minutes	0 out of 50 *
	Attempt 37	less than 1 minute	2 out of 50 *
	Attempt 36	less than 1 minute	0 out of 50 *
	Attempt 35	less than 1 minute	0 out of 50 *

Attempt	Time	Score
Attempt 34	less than 1 minute	0 out of 50 *
Attempt 33	less than 1 minute	0 out of 50 *
Attempt 32	less than 1 minute	0 out of 50 *
Attempt 31	less than 1 minute	0 out of 50 *
Attempt 30	less than 1 minute	0 out of 50 *
Attempt 29	less than 1 minute	0 out of 50 *
Attempt 28	less than 1 minute	0 out of 50 *
Attempt 27	less than 1 minute	0 out of 50 *
Attempt 26	less than 1 minute	0 out of 50 *
Attempt 25	less than 1 minute	0 out of 50 *
Attempt 24	less than 1 minute	0 out of 50 *
Attempt 23	less than 1 minute	0 out of 50 *
Attempt 22	less than 1 minute	0 out of 50 *
Attempt 21	less than 1 minute	0 out of 50 *
Attempt 20	less than 1 minute	0 out of 50 *
Attempt 19	less than 1 minute	0 out of 50 *
Attempt 18	less than 1 minute	0 out of 50 *
Attempt 17	less than 1 minute	0 out of 50 *
Attempt 16	less than 1 minute	0 out of 50 *
Attempt 15	less than 1 minute	0 out of 50 *
Attempt 14	less than 1 minute	0 out of 50 *
Attempt 13	less than 1 minute	0 out of 50 *
Attempt 12	less than 1 minute	0 out of 50 *
Attempt 11	less than 1 minute	0 out of 50 *
Attempt 10	less than 1 minute	0 out of 50 *
Attempt 9	1 minute	5 out of 50 *
Attempt 8	less than 1 minute	4 out of 50 *
Attempt 7	less than 1 minute	3 out of 50 *

Attempt	Time	Score
Attempt 6	less than 1 minute	2 out of 50 *
Attempt 5	less than 1 minute	2 out of 50 *
Attempt 4	less than 1 minute	1 out of 50 *
Attempt 3	1 minute	1 out of 50 *
Attempt 2	5 minutes	1 out of 50 *
Attempt 1	150 minutes	26 out of 50 *

* Some questions not yet graded

Score for this attempt: **50** out of 50
Submitted Oct 7 at 9:48am
This attempt took 9 minutes.

Question 1

1 / 1 pts

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

Correct!

- (●) True
- ( ) False

Question 2

1 / 1 pts

If a branch separates all records into a single class, then the purity is very low.

- ( ) True
- (●) False

Correct!

Question 3

1 / 1 pts

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

Correct!

- (●) True
- ( ) False

Question 4

1 / 1 pts

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

Correct!

- (●) True
- ( ) False

Question 5

1 / 1 pts

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

- ( ) True

Correct!

- (●) False

Question 6

1 / 1 pts

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

- True
- False

Correct!

Question 7

1 / 1 pts

Linear Regression cannot not be applied on every dataset.

- True
- False

Correct!

Question 8

1 / 1 pts

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

- True
- False

Correct!

Question 9

1 / 1 pts

XOR function mappings can easily be classified by decision trees.

Correct!

- (●) True
- ( ) False

Question 10

1 / 1 pts

Correlations are never distorted if the data is not standardized.

- ( ) 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

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

Correct!

greedy

orrect Answers

greedy

local optima

Question 13

2 / 2 pts

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

Correct!

least

orrect Answers

the least

least

lowest

smallest

small

minimum

Question 14

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 15

2 / 2 pts

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

Correct!

Gain

orrect Answers

gain

Question 16

2 / 2 pts

The process of _____ data before calculating correlations is the best way to get good correlations.

Correct!

standardizing

orrect Answers

standardizing

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 becomes 0 :

Answer 1:

Correct!

0

Question 20

2 / 2 pts

The standard deviation of the new transformed data after standardization 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

2.5 / 2.5 pts

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

False

Correct!

True

Question 23

2.5 / 2.5 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.

Correct!

True

False

Question 24

10 / 10 pts

Read this article and provide your summary of the article:

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

Also discuss your understanding of the equation:

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

(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:

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.

*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.

Quiz Score: **50** out of 50