Online Test 1 (10%)

Started: 1 Oct at 11:53

Quiz instructions

Due Date: Week 5 Lab Session

Weighting: 10%

About this test

1. **You must participate in this test in person on campus.**
2. This is an online test worth 10% of the total marks for this unit.
3. It consists of 20 questions on theories and applications. Using KNIME in the quiz is essential to find answers.
4. You may **only attempt this test once**.
5. **Do not navigate away from the test (close the browser tag) before you have completed it and submitted it.**
6. You have **60 minutes** to complete the test; exceeding this will result in a loss of marks and penalties.
7. This is a **close-book test**.
8. Once you begin the quiz, a summary will appear in the upper right corner, showing the number of questions (completed and total) and the time remaining.
9. **Please bring your student ID to the test and put it on the desk in front of you so the tutor can verify your identity.**
10. **Please use Chrome to access the quiz. The second options are Firefox and Microsoft Edge. Please do not use Safari in the test.**
11. **The multi-answering questions have a deduction mechanism. Be careful about what you select in the answers.**

Question 1 5 pts

Select the description(s) that is (are) TRUE.

- [ ] Logistic regression is as the same as linear regression but with different names.
- [ ] Linear regression model can only be used to process numerical data.
- [x] Logistic regression model can be used to process categorical data.
- [x] Linear regression model can be used to process categorical data.

Question 2 5 pts

Which description(s) of the following Python code is (are) correct?

```
numbers = [1, 2, 3, 4, 5]
for i in range(len(numbers)):
    numbers[i] *= 2
print(numbers)
```

- [x] The len() function returns the length of variable "numbers," which equals to 5 in this given example.
- [ ] The output value is 30.
- [ ] The printed output would be "numbers" as a string.
- [x] The statement "range(len(numbers))" is equal to "range(0, len(numbers))"

Question 3 5 pts

Which of the definition of Bayes' Theorem is TRUE?

(i) $P(A \mid B) = \frac{P(B|A) \cdot P(B)}{P(A)+P(B|A)}$

(ii) $P(B \mid A) = \frac{P(B|A) \cdot P(A)}{P(B)}$

(iii) $P(A \mid B) = \frac{P(B|A)}{P(B)}$

(iv) $P(B \mid A) = \frac{P(A|B) \cdot P(B)}{P(A)}$

Please note that the order of the answer can be different than what is listed in the question.

Look at the answers carefully.

- (i)
- (ii)
- (iii)

○
(iv)

Question 4 5 pts

Which description(s) of the following Python code is (are) correct?

```
number = 7
if number > 5:
    print("Greater than 5")
else:
    print("Less than or equal to 5")
number = 2
```

☑ The value of variable "number" will be 2 after the program is executed.

☐ The printed output is "Less than or equal to 5"

☐ The variable number is an array with single element.

☑ The printed output is "Greater than 5"

Question 5 5 pts

Select the description(s) that is (are) TRUE.

☐ Predicting a categorical attribute of objects are usually solved as clustering problems.

☑ Predicting a categorical attribute of objects are usually solved as classification problems.

☑ Most classification models are categorised as supervised models.

☐ A predictive model doesn't require the class variable in the training process.

Question 6 5 pts

Select the description(s) that is (are) TRUE.

☐ Precision is also known as the overall success rate.

- [x] Type II error rate is also known as the miss rate.
- [x] Type I error rate is also known as the false alarm rate.
- [ ] A good model should have a low precision value.

Question 7 5 pts

Which answer(s) will be the least square regression line(s) of the given dataset?

x	y
5	7
2	1
4	3
6	5

- [ ] x = 0.941 y - 1
- [x] y = 0.941 x - 1
- [x] y = 1.257 x - 1.343
- [ ] x = 0.55 y + 2.05

Question 8 5 pts

Select the description(s) that is (are) TRUE.

- [x] Each cluster may have different number of data points after the clustering in DBSCAN.
- [ ] Each cluster should have the same number of data points after the clustering in DBSCAN.
- [x] Users can control the number of clusters output by k-means, even if it is an unsupervised learning method.
- [ ] Users can control the number of clusters output by DBSCAN, even if it is an unsupervised learning method.

Question 9 5 pts

Select the description(s) that is (are) TRUE.

- [x] "KNIME Explorer" is the window for you to allocate the existing KNIME workflows in the workspace.
- [x] You can use "Node Repository" to configure nodes in KNIME.
- [ ] You can use "KNIME Explorer" to find the desired node in KNIME.
- [ ] You can use "Node Repository" to find the desired node in KNIME.

Question 10 5 pts

Which definition of Accuracy in the predictive model evaluation is TRUE?

(i) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$

(ii) $Accuracy = \frac{TP+TN+FP+FN}{TP+TN} \times 100$

(iii) $Accuracy = \frac{TP+TN}{FP+FN} \times 100$

(iv) $Accuracy = \frac{TP \times TN}{TP+TN+FP+FN} \times 100$

TP, TN, FP, and FN stand for the true positive, the true negative, the false positive, and the false negative, respectively.

Please note that the order of the answer can be different than what is listed in the question.

Look at the answers carefully.

- ○ (iii)
- ◉ (i)
- ○ (iv)
- ○ (ii)

Question 11 5 pts

Select the description(s) that is (are) TRUE.

- [x] In phase 4 of the data analytics lifecycle (DAL), whether the existing tools are sufficient to run the selected model should be considered.
- [ ] Phase 4 of the data analytics lifecycle (DAL) is mainly for model planning.
- [ ] Linear regression is a model that doesn't require training.
- [x] The collected dataset should be partitioned into training and test sets if linear regression is selected to be the model.

Question 12 5 pts

Given the example of tossing two coins.

The possible sample space is {HH, HT, TH, TT}, where H and T represent the head and tail, respectively.

Select the description(s) that is (are) TRUE.

- [ ] P(HH) = 0.5
- [x] P(Getting two heads given the first coin is H) = 0.5
- [ ] P(The second coin being T given the first coin is H) = 0.25
- [x] P(At least one H) = 0.75

Question 13 5 pts

Match the contents to their correct groups.

True positive rate (TPR)	Recall
True negative rate (TNR)	Selectivity
False positive rate (FPR)	Fall-out
False negative rate (FNR)	Missing rate

Question 14 5 pts

Given the data points below, which will be the least square regression line(s) for this dataset?

x	y
3	5
1	2
2	3
-7	-9
6	10

- [ ] x = 0.693 y - 0.525
- [x] y = 1.436 x + 0.764
- [ ] y = 0.764 x - 1.436
- [ ] y = 0.693 x - 0.525

Question 15 5 pts

John flies frequently and likes to upgrade his seat to first class. He has determined that if he checks in for his flight at least two hours early, the probability that he will get an upgrade is 0.65; otherwise, the probability that he will get an upgrade is 0.3. With his busy schedule, he checks in at least two hours before his flight only 45% of the time.

Suppose John did not receive an upgrade on his most recent attempt, what is the probability that he did not arrive two hours early?

- (●) About 70.97%.
- ( ) About 54.25%.
- ( ) About 50.55%.
- ( ) About 30%.

Question 16 5 pts

Match the contents to their correct groups.

True positive rate (TPR)

Sensitivity

True negative rate (TNR) — Specificity

Precision — Calculated by the true positi

Area under the curve (AUC) — The plasma concentration-ti

Given the vehicle dataset below, try to process the data with KNIME or any preferred tool to answer the questions.

Dataset: **car data_v02.csv (https://swinburne.instructure.com/courses/62961/files/33245650?wrap=1)** **(https://swinburne.instructure.com/courses/62961/files/33245650/download?download_frd=1)** **(https://swinburne.instructure.com/assessment_questions/5281188/files/23802173/download?verifier=EP8C6IvRt9W6OuhaBGyTjFsHicgUTRNKPWCYLezO&download_frd=1)**

Preparation steps:

1. Exclude the tuples of fuel type "CNG" from your data. (The CNG class can still be kept on the label.)
2. Partition the data with 85% and 15% for the training and test sets, respectively, **with the "Draw randomly" method and use "3122" as the seed**. The "draw randomly" method cuts the data after shuffling the data. This is important for getting the exact result.

Question 17 5 pts

After removing the "CNG" fuel-type vehicles, which node in KNIME should be used to partition the data?

The parameter setting of this node must follow the given values:

1. Use the "Draw randomly" method to split the data.

2. Use "3122" to be the random seed.

The goal is to split 85% of the input data into the training set, and the remaining goes into the test set.

- Row Splitter
- Numeric Row Splitter
- Shuffle

◉
Partitioning

Question 18 5 pts

The given dataset has been cleaned; thus, there is no duplicated data.

The first step in preparing the data is removing tuples whose "Fuel_Type" is "CNG".

Which node(s) in KNIME can be used to complete this task?

☐
Rule-based Row Splitter

☐
Nominal Value Row Filter

☑
Row Filter

☐
Rule-based Column Filter

Question 19 5 pts

Use the "Year", "Present_Price", and "Kms_Driven" attributes with the default parameter setting to build a regression model to predict the "Selling_Price".

Select the descriptions that is(are) TRUE.

☑
The R-square value is 0.865.

☑
The Numeric Scorer node should be used for finding the model performance.

☑
The Regression Predictor node should take the data input from the Regression Learner node.

☐
The Scorer node should be used for finding the model performance.

Question 20 5 pts

Use the "Year", "Selling_Price", "Present_Price", and "Owner" attributes to build a regression model for predicting the "Seller_Type".

The maximal number of epochs should be set to 3000, the Epsilon value should be set to 1.0E-3, "LineSearch" should be used as the learning rate strategy, and the seed should be set to "3122". The remaining parameters should be kept by default.

Select the descriptions that is(are) TRUE.

- [x] The Scorer node should be used to find the model performance.
- [x] All tuples of the "dealer" seller_type in the test set are predicted correctly.
- [ ] All tuples of the "individual" seller_type in the test set are predicted correctly.
- [ ] The Numeric Scorer node should be used to find the model performance.

Saved at 9:51 Submit quiz