X

(https://swayam.gov.in) (https://swayam.gov.in/nc_details/NPTEL)

santosh44kumar@yahoo.com ˅

NPTEL (https://swayam.gov.in/explorer?ncCode=NPTEL) » **Python for Data Science (course)**

Announcements (announcements) **About the Course (https://swayam.gov.in/nd1_noc20_cs36/preview)**

Ask a Question (forum) Progress (student/home) Mentor (student/mentor)

Unit 6 - Week 4

Course outline

How does an NPTEL online course work?

Week 0

Week 1

Week 2

Week 3

Week 4

- Introduction to Classification Case Study (unit?unit=4&lesson=28)
- Case Study on Classification Part I (unit?unit=4&lesson=29)
- Case Study on Classification

Assignment 4

The due date for submitting this assignment has passed. **Due on 2020-02-26, 23:59 IST.**

Assignment submitted on 2020-02-26, 00:09 IST

Click here (https://drive.google.com/open?id=1WPCZLhqOsDble9CBeBs07GMPlOxaYJSc) to download the Data sets.

1) Identify which one of the following methods(s) is/are used to solve the given problem. ***1 point***

Problem statement: Mr. John is going to sell his house and wants to predict the right asking price using Machine Learning on previous data.

He has collected data on the square footage, location, age of the house, numbers of bedrooms and bathrooms and price (in Rs) of the house.

☑ Linear Regression

☑ Random Forest

□ Logistic Regression

☑ Decision Tree

Yes, the answer is correct.
Score: 1

Accepted Answers:

Linear Regression

Random Forest

Part II (unit?unit=4&lesson=30)

○ Introduction to Regression Case Study (unit?unit=4&lesson=31)

○ Case Study on Regression Part I (unit?unit=4&lesson=32)

○ Case Study on Regression Part II (unit?unit=4&lesson=33)

○ Case Study on Regression Part III (unit?unit=4&lesson=34)

● Data sets (unit?unit=4&lesson=89)

○ Quiz : Practice Assignment 4 (assessment?name=69)

● **Quiz : Assignment 4 (assessment?name=95)**

● Week 4 Feedback (unit?unit=4&lesson=99)

● Solution - Assignment 4 (unit?unit=4&lesson=101)

Supporting material for Week 4

Download Videos

Decision Tree

2) Which of the following statement (s) is/are not true about **supervised learning?** ***1 point***

○ Modeling the relationship between measured features of data and some label associated with the data

● Modeling the features of a dataset without reference to any label

○ In classification, the labels are discrete categories

○ In regression, the labels are continuous quantities

Yes, the answer is correct.
Score: 1

Accepted Answers:
Modeling the features of a dataset without reference to any label

3) Which of the following metric (s) is/are used for the classification problem? ***1 point***

□ R-Squared

□ Adjusted R-Squared

☑ Confusion matrix

☑ Accuracy score

Yes, the answer is correct.
Score: 1

Accepted Answers:
Confusion matrix
Accuracy score

4) In logistic regression, if the posterior probability **Pr(Class=k | X=x)** is linear in ‘x’ then ***1 point***

○ x is not related to Y

● decision boundary is linear

○ decision boundary is non-linear

○ there is no decision boundary

Yes, the answer is correct.
Score: 1

Accepted Answers:
decision boundary is linear

5) How do you find the optimum value of ‘K’ in a K-Nearest Neighbor classifier? ***1 point***

○ Larger K-value (greater than 10 neighbors)

○ By considering only the closest 2 to 4 neighbors

● By observing an error Vs K plot and that value of K corresponding to the lowest error

○ Only the nearest single neighbor, i.e. K=1

Yes, the answer is correct.
Score: 1

Accepted Answers:
By observing an error Vs K plot and that value of K corresponding to the lowest error

6) What is the logit function when ‘p’ refers to probability of occurrence of an event? ***1 point***

○

$\log(Y|X)$

○ $\exp(p(x)/(1-p(x)))$

○ exp(odd)

● $\log\left(\frac{p(X)}{(1-p(X)}\right)$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$log\left(\frac{p(X)}{(1-p(X)}\right)$

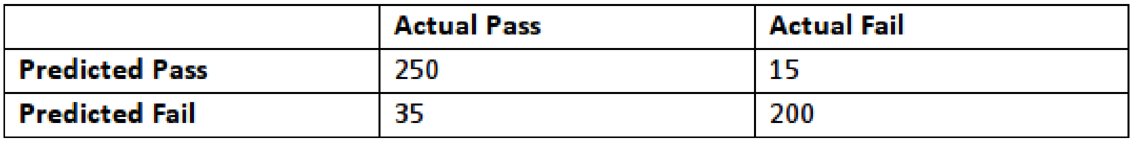
7) During which of the following situations, you may not consider using KNN as a method of solving a classification problem? ***1 point***

○ When there are more than two classes to classify
● When there is very large number of input variables (p) in the data matrix (N x p)
○ While imputing missing values in the categorical variables/features
○ When all the variables are categorical

Yes, the answer is correct.
Score: 1

Accepted Answers:
When there is very large number of input variables (p) in the data matrix (N x p)

8) Consider the following confusion matrix and calculate the number of samples that has been wrongly classified as Fail ***1 point***



	Actual Pass	**Actual Fail**
Predicted Pass	250	15
Predicted Fail	35	200

○ 15
● 35
○ 50
○ 200

Yes, the answer is correct.
Score: 1

Accepted Answers:
35

9) Which of the following method you will use to find the best fit line in logistic regression? ***1 point***

○ Ordinary Least Square
● Maximum Likelihood Estimator
○ Weighted Least Square

○ Lasso Method

Yes, the answer is correct.
Score: 1

Accepted Answers:
Maximum Likelihood Estimator

10) Which of the following statement(s) is/are true about errors in regression? ***1 point***

● Error values of linear regression must be normally distributed but not in the case of logistic regression

○ Error values of logistic regression must be normally distributed but not in the case of linear regression

○ Both linear regression and logistic regression error values must be normally distributed

○ Both linear regression and logistic regression error values need not to be normally distributed

Yes, the answer is correct.
Score: 1

Accepted Answers:
Error values of linear regression must be normally distributed but not in the case of logistic regression

11) State which of the following statements are true/false about Random Forest. ***1 point***

i. Random Forest can be adapted to classification or numeric prediction problems
ii. It classifies the data based on voting or average method

○ True, False

○ False, True

● True, True

○ False, False

Yes, the answer is correct.
Score: 1

Accepted Answers:
True, True

Given the datasets - CrashTest_TrainData.csv, CrashTest_TestData.csv read them as two separate data frames named Train_Data and Test_Data respectively.

Data description:

- A crash test is a form of destructive testing that is performed in order to ensure high safety standards for various cars
- Several cars have rolled into an independent audit unit for crash test and they are being evaluated on a defined scale {poor (-10) to excellent (10)}
- However, with this data in future they should be able to predict the type of the car

Answer questions from 12 to 20.

12) To predict the type of the car, how many valid input variables are available in the **Train_Data**? ***1 point***

○ 7

○ 5

● 6

○ 4

No, the answer is incorrect.
Score: 0

Accepted Answers:
5

13) What is the difference between third quartile values of the variable **ManBI** from **Train_Data** and **Test_Data**? ***1 point***

○ 1.2858

○ 2.3856

● 0.9175

○ 0.0156

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.9175

14) How many distinct car types are there in the **Train_Data**? ***1 point***

○ 4

○ 3

● 2

○ 1

Yes, the answer is correct.
Score: 1

Accepted Answers:
2

15) How many missing values are there in **Train_Data**? ***1 point***

○ 0

○ 1

● 3

○ 5

Yes, the answer is correct.
Score: 1

Accepted Answers:
3

16) What is the proportion of car types in the **Test_Data**? ***1 point***

○ 60,40

○ 20,80

● 50,50

○ 90,10

Yes, the answer is correct.
Score: 1

Accepted Answers:
50,50

Follow the steps given below to build the classifier models:

- Drop the missing values
- Ensure the datatypes of the columns are appropriate
- Map the categorical variables into integers

17) What is the accuracy score of the K-Nearest Neighbor model (model_1) with 3 neighbors using **Train_Data** and **Test_Data**? ***1 point***

○ 0.89
○ 0.65
● 0.70
○ 0.88

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.70

18) Identify the list of indices of misclassified samples from the '**model_1**'. ***1 point***

○ 1, 2, 8, 9,11, 15
● 0, 1, 7, 8, 10, 14
○ 2, 3, 4, 8, 11, 15
○ 0, 3, 9, 12, 15, 17

Yes, the answer is correct.
Score: 1

Accepted Answers:
0, 1, 7, 8, 10, 14

19) Rebuild the model (model_2) with 2 neighbors keeping the other modelling steps constant. Compare results of the two models (model_1 & model_2). ***1 point***

Choose the correct option.

○ Performance of model 1 is better than the model 2
● Performance of model 2 is better than the model 1
○ There is no difference in the performance of two models
○ None of the above

Yes, the answer is correct.
Score: 1

Accepted Answers:
Performance of model 2 is better than the model 1

20) Build a logistic regression model (model_3) keeping the modelling steps constant. The accuracy of the model_3 is: - ***1 point***

○ 0.65
○ 0.82
○ 0.92
● 1.0

Yes, the answer is correct.
Score: 1

Accepted Answers:
1.0