DATA/MSML/BIOI602 Principles of Data Science

Midterm Exam

Fall 2024

Heng Huang

University of Maryland College Park

1. Personal information:

   Name:

   E-mail address:

2. There should be 10 numbered pages in this exam.
3. This exam is open book, open notes. **NO internet access is allowed**.
4. You do not need a calculator.
5. If you need more room to answer a question, please use the back of the page and clearly mark on the front of the page if we are to look at the back.
6. You have 75 minutes.
7. Good luck!

Questions	Topics	Maximum scores	Scores
I	Multiple choices	36	
II	Q&A	37	
III	Pandas	5	
IV	Regression	12	
V	PCA	10	
	Total	100	

I. Multiple Choice Questions [36 pts, each question 3pts]

1. Which of the following is one of the key data science skills?
    a) Data Visualization
    b) Machine Learning
    c) Statistics
    d) All of the mentioned

D

2. What does the term "feature engineering" refer to in data science?
    a) The process of transforming raw data into meaningful features
    b) The process of gathering more data
    c) The process of applying machine learning algorithms
    d) The process of splitting data into training and testing sets

A

3. Which of the following is a common method for data preprocessing?
    a) Data normalization
    b) Data storage
    c) Data aggregation
    d) Data visualization

A

4. Which of the following is the top most important thing in data science?
    a) data
    b) question
    c) answer
    d) none of the mentioned

B

5. What type of data is considered unstructured?
    a) Data in relational databases
    b) Data in spreadsheets
    c) Data in CSV files
    d) Text documents and images

D

6. Which of the following testing is concerned with making decisions using data?
    a) Hypothesis
    b) Probability
    c) Causal
    d) None of the mentioned

A

7. All pandas data structures are ___ mutable but not always _____mutable.
   a) size, value
   b) semantic, size
   c) value, size
   d) none of the mentioned

C

8. Which of the following statement will import pandas?
   a) import pandas as pd
   b) import panda as py
   c) import pandaspy as pd
   d) all of the mentioned

A

9. Which of the following step is performed by data scientist after acquiring the data?
   a) Data Integration
   b) Data Replication
   c) Data Cleansing
   d) All of the mentioned

C

10. What are the conditions in which Type-I error occurs?
   a) The null hypotheses get rejected even if it is true
   b) The null hypotheses get accepted even if it is false
   c) Both the null hypotheses as well as alternative hypotheses are rejected
   d) None of the options

A

11. Heights of college women have a distribution that can be approximated by a normal curve with a mean of 65 inches and a standard deviation equal to 3 inches. About what proportion of college women are between 65 and 67 inches tall?
   a) 0.50
   b) 0.17
   c) 0.75
   d) 0.25

D

12. The statement "If there is sufficient evidence to reject a null hypothesis at the 10% significance level, then there is sufficient evidence to reject it at the 5% significance level" is: Please select the best answer of those provided below.
   a) Always True
   b) Not Enough Information; this would depend on the type of statistical test used
   c) Never True
   d) Sometimes True; the p-value for the statistical test needs to be provided for a conclusion

D

II. Short Questions [37 pts]

Answer the following 7 questions. **Explain your reasoning** in 1 sentence.

1. [5 pts] Half of all kangaroos in the zoo are angry, and 2/3 of the zoo is comprised of kangaroos. Only 1 in 10 of the other animals are angry. What's the probability that a randomly chosen animal is an angry kangaroo?

   1/2*2/3= 1/3

2. [5 pts] (True/False) Suppose you are given a dataset of cellular images from patients with and without cancer. The dataset had 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, it is it a good classifier.

   FALSE. This is not a good accuracy on this dataset, since a classifier that outputs "cancer-free" for all input images will have better accuracy (90%).

3. [5 pts] (True/False) A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

   FALSE. The second classifier has better test accuracy which reflects the true accuracy, whereas the first classifier is overfitting.

4. [7 pts] Boolean random variables A and B have the joint distribution specified in the table below.

A	B	$P(A, B)$
0	0	0.32
0	1	0.48
1	0	0.08
1	1	0.12

Given the above table, please compute the following five quantities: (1 pt for each result)

P (A = 0) = P(A = 0;B = 0) + P(A = 0;B = 1) = 0.32 + 0.48 =0.8

P (A = 1) =1 -P(A = 0) = 0.2

P (B = 1)= P(B = 1;A = 0) + P(B = 1;A = 1) = 0.48 + 0.12 = 0.6

P (B = 0) = 1 -P(B = 1) = 0.4

P (A = 1|B = 0) = P(A = 1;B = 0)=P(B = 0) = 0.08/0.4 = 0.2

Are A and B independent? Justify your answer. (1pt for Yes, 1pt for answer)

YES. Using the calculations above,

P(A = 0)P(B = 0) = 0.8 * 0.4 = 0.32 = P(A = 0;B = 0)
P(A = 0)P(B = 1) = 0.8 * 0.6 = 0.48 = P(A = 0;B = 1)
P(A = 1)P(B = 0) = 0.2 * 0.4 = 0.08 = P(A = 1;B = 0)
P(A = 1)P(B = 1) = 0.2 * 0.6 = 0.12 = P(A = 1;B = 1)

5. [5 pts] (True/False) Support vector machines, like logistic regression models, give a probability distribution over the possible labels given an input example.

False

6. [5 pts] In online learning, we can update the decision boundary of a classifier based on new data without reprocessing the old data. Now for a new data point that is an outlier, which of the following classifiers are likely to be affected more severely? Naive Bayes, Logistic Regression, SVM? Please give a one sentence explanation to your answer.

   False (There are no guarantees that the support vectors remain the same. The feature vectors corresponding to polynomial kernels are non-linear functions of the original input vectors and thus the support points for maximum margin separation in the feature space can be quite different.)

7. [5 pts] When comparing the difference between two population proportions, a pooled estimate of the population proportion can be used for two-tail tests where the null hypothesis assumes that the population proportions are equal. What is the alternate hypothesis?

   H1: p1 not equal to p2

III. [5 pts] Throughout this question, we are dealing with pandas DataFrame and Series objects. All code for this question, where applicable, must be written in Python. You may assume that pandas has been imported as pd.

The following DataFrame cars contains the names of car models from 1970 to 1982. The name column is the primary key of the table.

The first five rows are shown below.

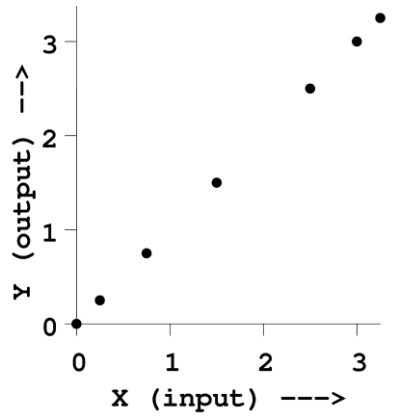
name	mpg	horsepower	weight	acceleration	year	origin	brand
toyota corolla 1200	32.0	65	1836	21.0	1974	Japan	toyota
buick skylark 320	15.0	165	3693	11.5	1970	USA	buick
fiat 128	29.0	49	1867	19.5	1973	Europe	fiat
ford mustang gl	27.0	86	2790	15.6	1982	USA	ford
ford torino	17.0	140	3449	10.5	1970	USA	ford

Write a line of Pandas code that creates a **Series** of the **names** of cars created by brand "carbrand" with greater than mpgnum mpg. The resulting Series should be assigned to the variable varname.

varname = cars[(cars["brand"]=="carbrand") & (cars["mpg"] > mpgnum)]["name"]

IV. [12 pts] Regression

1. [6 pts] Consider the following data with one input and one output.



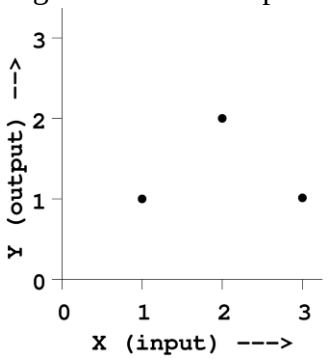
(a) What is the mean squared training set error of running linear regression on this data (using the model $y = w_0 + w_1 x$)?

0

(b) What is the mean squared test set error of running linear regression on this data, assuming the rightmost three points are in the test set, and the others are in the training set.

0

2. [6 pts] Consider the following data with one input and one output.



X	Y
1	1
2	2
3	1

(a) What is the mean squared training set error of running linear regression on this data (using the model $y = w_0 + w_1 x$)? (Hint: by symmetry it is clear that the best fit to the three datapoints is a horizontal line).

```
SSE = (1/3)^2 + (2/3)^2 + (1/3)^2 = 6/9
MSE = SSE/3 = 2/9
```

(b) What is the mean squared leave-one-out cross-validation (LOOCV) error of running linear regression on this data?

```
1/3 * (2^2 + 1^2 + 2^2) = 9/3 = 3
```

V. [10 pts] Basic PCA

Given 3 data points in 2-d space, (1, 1), (2, 2) and (3, 3),

(a) [4 pts] what is the first principle component?

$$pc = (1/\sqrt{2}, 1/\sqrt{2})' = (0.707, 0.707)', \text{ (the negation is also correct)}$$

(b) [3 pts] If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

$$4/3 = 1.33$$

(c) [3 pts] For the projected data in (b), now if we represent them in the original 2-d space, what is the reconstruction error?

0