## Part 1 (Lec1-3)

- A discrete random variable takes a finite number of values, while a continuous random variable can only take infinite number of values.

Answer: False.
Reason: Imagine now I have a random variable X which denotes the number of stars in the universe, this variable is discrete but may indeed take infinite number of possible values.

- Causality is a deterministic relationship; suppose we know A and B have causal relation, if A occurs, B is for sure to take place.

Answer: False.
Reason: Causality is a statistical relationship. Please see Page 19 of Lec 3.

- One key step in Data Cleaning is to check the missing features of data samples. When we have insufficient number of training samples in our dataset, we may consider removing the examples with missing features.

Answer: False.
Reason: We may consider removing samples only when we have sufficient number of samples. Please see Page 25 of Lec 2.

- Three balls are drawn from three urns sequentially, one ball from each urn. The first urn contains 1 blue and 7 red balls, the second urn contains 2 blue and 6 red balls, and the third urn contains 3 red and 5 green balls. Find the probability that 2 red balls are chosen.

Answer: 270/512
Reason:

Three possible cases where two red balls are chosen:

First draw – Second draw – Third draw
Red – Red – Green Probability: 7/8 * 6/8 * 5/8 = 210/512
Red – Blue – Red Probability: 7/8 * 2/8 * 3/8 = 42/512
Blue – Red – Red Probability: 1/8 * 6/8 * 3/8 = 18/512

So the probability is (210+42+18)/512 = 270/512

- A machine learning algorithm takes the temperature as one of its input features. The temperature is measured in Celsius. Please select the correct option.

  a) The temperature in Celsius is considered as interval data.

b) We can calculate the mean and standard deviation of temperature.
c) The temperature in Celsius is considered as ratio data.
d) None of the rest.
e) (a), and (b)
f) (a), and (c)

Answer: e) None of the rest.
a) Is correct; b) is correct (see Page 13 or Lec 2); c) is not correct, since Celsius is interval.

- A person draws 2 cards from a deck of 52 cards, one after another without replacing the previous card back. What is the probability of drawing two Queens in a row?

Answer: 1/221

Since there are 4 queens in the cards, the probability of drawing 2 queens in a row without replacement is: 4/52 * 3*/51 = 1/13*1/17 = 1/221/

- Which of the following task is likely to be achieved via supervised learning?
  a) Using historical data for weather forecast.
  b) Grouping together users with similar viewing patterns in order to recommend similar content.
  c) Grouping a number of oranges by their size.
  d) None of the rest.

Answer: a)
b) and c) are clustering tasks, which are unsupervised.

- A machine learning algorithm takes the letter grade of students as one of its input features. The letter grade can take any element from {A+, A, A−, B+, B, B−, C+, C, D+, D, F}, subject to some distribution curving. For example, A+ corresponds to 95%, A corresponds to 85%, and A- corresponds to 80%. Which of the following statements is/are true?

  a) The letter grade is an example of nominal variable.
  b) The letter grade is an example of ordinal variable.
  c) The letter grade is an example of interval variable.
  d) The letter grade is an example of discrete variable.
  e) (a), and (c)
  f) (b), and (c)
  g) (b), and (d)

Answer: g)
Reason: the letter grade is discrete. It is also ordinal since there is no equal split.

- We have a collection of 1,000 images from three classes: cat, bird, and dog. With these images, we would like to train an image classifier that categorizes an input image into one of the three classes.

  To ensure the 1,000 images are of good quality for training the classifier, we ask a well-trained human inspector to go through all the images, to label the images and remove noisy ones. Eventually, we revmoved 200 images of low quality suggested by the inspector, and use the remaining 800 images to train the classifier; the 800 images comprise 200 cat images, 300 bird images, and 300 dog images.

  a) The human inspection process can be considered as a data cleaning step.
  b) If we are to use one-hot encoding for the labels of the three classes, we can set
  Cat = [1 1 1]
  Dog = [0 1 0]
  Bird = [0 0 0].

  c) The image classification conducted here is an unsupervised-learning task.
  d) If we keep the 200 noisy images (suggested by the human inspector), we will end up having more training images and hence a better-performed image classifier.

  e) (a) and (b)
  f) (a), (b), and (d)
  g) None of others is correct.

Answer: a).
Option b) is not correct, since at mostly one 1 may appear in the one-hot encoding.
Option c) is not correct, since classification is supervised learning.
Option d) is not correct, since adding the noisy images into training may end up a worse classifier.

- Suppose the random variable X has a probability mass function (pmf) given in the table below.

| X | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Pr[X] | 0.1 | (BLANK1) | 0.2 | 0.4 | (BLANK2) |

We also know that the expected value of X is 3.5.

1) What is the probability of Pr[X=5]? [1] (2 Marks)

2) What is the probability of Pr[X<=2]? [2] (2 Marks)

Answer:
Blank 1: 0.2
Blank 2: 0.2

Reason:
Let the probability of (BLANK1) to be x, and probability of (BLANK2) to be y.

We know that

1) the probability must add up to 1, so that we have 0.1+x+0.2+0.4+y = 1
2) the expectation is 3.5, so that we have 1*0.1+2*x+3*0.2+4*0.4+5*y = 3.5

We solve these two questions together, there for we have x = 0.1, and y = 0.2.

As such,

Pr[X=5] = 0.2;
Pr[X<=2] = 0.1 + 0.1 = 0.2.

## Part 2 (Lec4-5)

This question is related to the understanding of modelling assumptions. f(x) = 5x - 3 is a linear function. (2 marks)

| Correct | Answer Choice |
|---|---|
| ○ | True |
| ◉ | False |

A: linear function does not have bias

This question is related to the understanding of linear systems and partial derivatives. Which of the following statements below is correct?

| Seq | Answer Choice |
|---|---|
| a) | In over-determined linear systems, the number of parameters is greater than the number of unknown equations. |
| b) | The system $\begin{bmatrix} 1 & 4 \\ 2 & 7 \\ -3 & 11 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2.5 \\ 4 \end{bmatrix}$ has no *exact* solution but an approximated solution is available using the left inverse. |
| c) | If **f**(**x**) is a vector-valued function of *sizep* x1 and **x** is an *m* x1 vector, then differentiation of **f**(**x**) with respect to **x** is an *m x p* matrix. |
| d) | A linear function needs to satisfy the properties of homogeneity only. |
| e) | None of the other options. |

A: b) is correct.

a) In over-determined linear systems, the number of parameters is less than the number of unknown equations.
c) If f(x) is a vector-valued function of size p x1 and x is an m x1 vector, then differentiation of f(x) with respect to x is an p x m matrix
d) A linear function needs to satisfy the properties of homogeneity and additivity

A set of linear equations is written as $\mathbf{w}^T\mathbf{X} = y^T$ where $\mathbf{X} \in \mathcal{R}^{3\times 2}$ and $\mathbf{y} \in \mathcal{R}^{2\times 1}$. How many simultaneous equations are there in this set of equations?

| Seq | Answer Choice |
|---|---|
| a) | 1 |
| b) | 2 |
| c) | 3 |
| d) | 4 |
| e) | 5 |

A: b)
We have w Transpose X = Y Transpose. Therefore, the number of columns in X, represents the number of samples, i.e., the number of simultaneous equations. Therefore, there are 2 simultaneous equations.

The values of feature x and their corresponding values of target y are shown in the table below.

| *x* | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| y | 5 | 4 | 3 | 2 | 1 |

Find the least square regression line y = a *x* + b and then estimate the value of y when *x* = 8.

| Seq | Answer Choice |
|---|---|
| a) | y = 8 |
| b) | y = +1 |
| c) | y = 0 |
| d) | y = −1 |
| e) | None of the above |

A: c)
- Remember to add bias to X
- Do left inverse with the formula
- You will obtain weight:

```
[[ 8.]
 [-1.]]
```

- If you are using python, y when x=8, will be an arbitrarily small value very close to zero.
It is basically due to the nature of python code – generally output the decimals, instead of 0.
But if you do the manual calculation of your w and your Xnew -> 1*8 + 8 *(-1) =  0
You can also do the round – it will result in 0.
In the following exam, we will set the precision to avoid confusion.

You are given a collection of 5 training data points of two features $(x_1, x_2)$ and their target output $(y)$ which are packed as follows:

Feature matrix: $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 0 & 6 \\ 1 & 0 \\ 0 & 5 \\ 1 & 7 \end{bmatrix}$, Target output: $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$.

Predict the output (up to 4 decimal places) of $(x_1, x_2) = (1, 3)$ using the linear regression model. (4 marks)

1) What is the mean of squared error of the estimated model? [1] (up to 4 decimal places, 2 mark)

2) The prediction for y is [2] (up to 4 decimal places, 2 marks).

| Number | Correct Answer |
|---|---|
| 1 | Range - Min: 1.3886 Max: 1.3888 |
| 2 | Range - Min: 2.9999 Max: 3.0001 |

A:
- remember to add bias to X
- this is an over-determined system, therefore, we need to use left-inverse and corresponding equation: $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- Tips for MSE: can use “mean_squared_error” function, from “sklearn.metrics”, as shown in lecture python demo