EE2211 Midterm Trial AY2526-S1

Dr Feng LIN
feng_lin@nus.edu.sg

Question #: 1

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

A. True
✓B. False

Lecture 4

Linear and Affine Functions

A linear function plus a constant is called an affine function

A linear function $f: \mathcal{R}^d \rightarrow \mathcal{R}$ is **affine** if and only if it can be expressed as $f(\mathbf{x}) = \boldsymbol{a}^T\mathbf{x} + b$ for some d-vector $\boldsymbol{a}$ and scalar b, which is called the offset (or bias)

Example:

$$f(\mathbf{x}) = 2.3 - 2x_1 + 1.3x_2 - x_3$$

This function is affine, with $b = 2.3$, $\boldsymbol{a}^T = [-2,\ 1.3,\ -1]$.

Question #: 2

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

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 $\mathbf{f}(\mathbf{x})$ is a vector-valued function of *sizep* x1 and $\mathbf{x}$ is an *m* x1 vector, then differentiation of $\mathbf{f}(\mathbf{x})$ with respect to $\mathbf{x}$ is an *m x p* matrix.

D. A linear function needs to satisfy the properties of homogeneity only.

E. None of the other options.

Derivative and Gradient

Lecture 5

Partial Derivatives

Differentiation of a vector function w.r.t. a vector

If $\mathbf{f}(\mathbf{x})$ is a vector function of *size h* x1 and $\mathbf{x}$ is a *d* x1 vector. Then differentiation of $\mathbf{f}(\mathbf{x})$ results in a *h* x *d* matrix

$$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_h}{\partial x_1} & \cdots & \frac{\partial f_h}{\partial x_d} \end{bmatrix}$$

The matrix is referred to as the **Jacobian** of $\mathbf{f}(\mathbf{x})$

Lecture 4

$\mathbf{X}$ is Square	Even-determined	m = d	One unique solution in general	$\hat{\mathbf{w}} = \mathbf{X}^{-1}\mathbf{y}$
$\mathbf{X}$ is Tall	Over-determined	m > d	No exact solution in general; An approximated solution	$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ Left-inverse
$\mathbf{X}$ is Wide	Under-determined	m < d	Infinite number of solutions in general; Unique constrained solution	$\hat{\mathbf{w}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$ Right-inverse

Question #: 3

A set of linear equations is written as

$$\mathbf{w}^T\mathbf{X} = \mathbf{y}^T$$

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?

A. 1
✓B. 2
C. 3
D. 4
E. 5

Lecture 4

Systems of Linear Equations

A set of linear equations can have no solution, one solution, or multiple solutions:

$$\mathbf{Xw} = \mathbf{y}$$

Where

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,d} \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

$$w^T X = y^T$$

$$\Rightarrow (w^T X)^T = (y^T)^T$$

$$\Rightarrow X^T w = y$$

$X^T \in R^{2\times 3}$

Question #: 4

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.

Lecture 5 Linear regression

A. y = 8
B. y = +1
✓C. y = 0
D. y = ?1
E. None of the above

```python
import numpy as np
from numpy.linalg import inv


X = np.array([[1, 3], [1, 4], [1, 5], [1, 6],
[1, 7]])


y = np.array([5, 4, 3, 2, 1])


w = inv(X.T @ X) @ X.T @ y


y_pre = np.array([1, 8]) @ w


print(f"y_pre: {y_pre}")
```

Question #: 5

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:

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

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

Lecture 5

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

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).

1. Range - Min:1.3886 Max:1.3888
2. Range - Min:2.9999 Max:3.0001

```python
import numpy as np
from numpy.linalg import inv


X = np.array([[1, 1, 2], [1, 0, 6], [1, 1, 0], [1, 0, 5], [1, 1, 7]])
y = np.array([1, 2, 3, 4, 5])


w = inv(X.T @ X) @ X.T @ y


# (1)
y_train = X @ w
MSE_train = np.mean(np.power((y - y_train), 2))
print(f"MSE_train: {MSE_train}")


# (2)
X_test = np.array([1, 1, 3])
y_test = X_test @ w
print(f"y_test: {y_test}")
```

Question #: 6

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.

A. True
✓B. False

Lecture 2

Data Cleaning: Handling missing features

1. Removing the examples with missing features from the dataset
   – Can be done if the dataset is big enough so we can sacrifice some training examples

2. Using a learning algorithm that can deal with missing feature values
   – Example: random forest

week 9.

3. Using a data imputation technique

Question #: 7

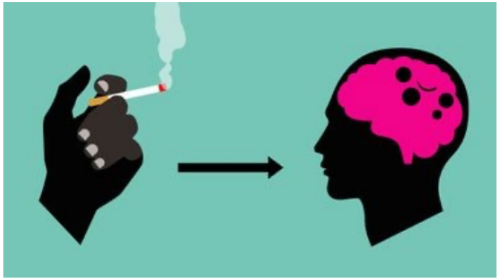
Causality is a deterministic relationship; suppose we know A and B have causal relation, if A occurs, B is for sure to take place.

A. True

✓B. False

Causality is a statistical relationship

- Decades of data show a clear causal relationship between smoking and cancer.
- If one smokes, it is a sure thing that his/her risk of cancer will increase.
- But it is not a sure thing that one will get cancer.
- The relationship is not deterministic.



Question #: 8

A discrete random variable takes a finite number of values, while a continuous random variable can only take infinite number of values. (2 marks)

A. True

✓B. False

Question #: 9

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.

Please select the correct option.

✓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 unsuperivsed-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.

Lecture 2

Formatting Data

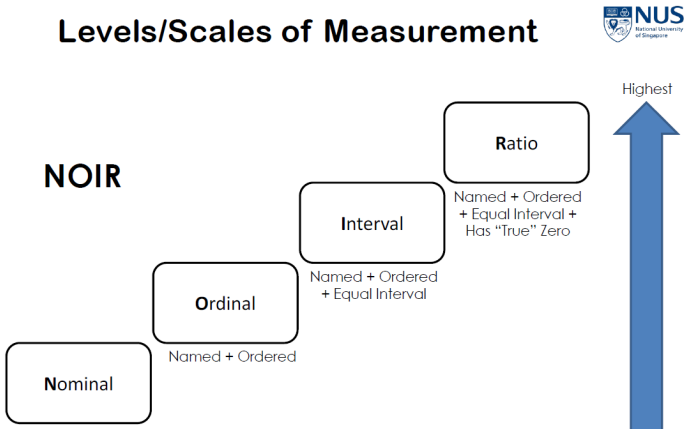
- **Binary Coding** to convert categories into binary form
  - One-hot encoding: unify several entities within one vector
    - Example: the color of a pixel can be red, yellow, or green
    - Very common in classification tasks!

$$\text{red} = [1, 0, 0]$$
$$\text{yellow} = [0, 1, 0]$$
$$\text{green} = [0, 0, 1]$$

Question #: 10

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)



Question #: 11

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.

Question #: 12

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?

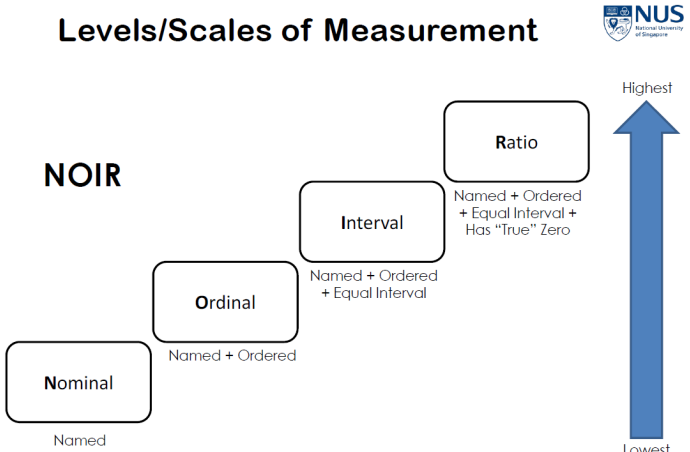
A. 2/52
B. 4/52
✓C. 1/221
D. 3/51

$$\frac{4}{52} \times \frac{3}{51} = \frac{1}{221}$$

Question #: 13

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

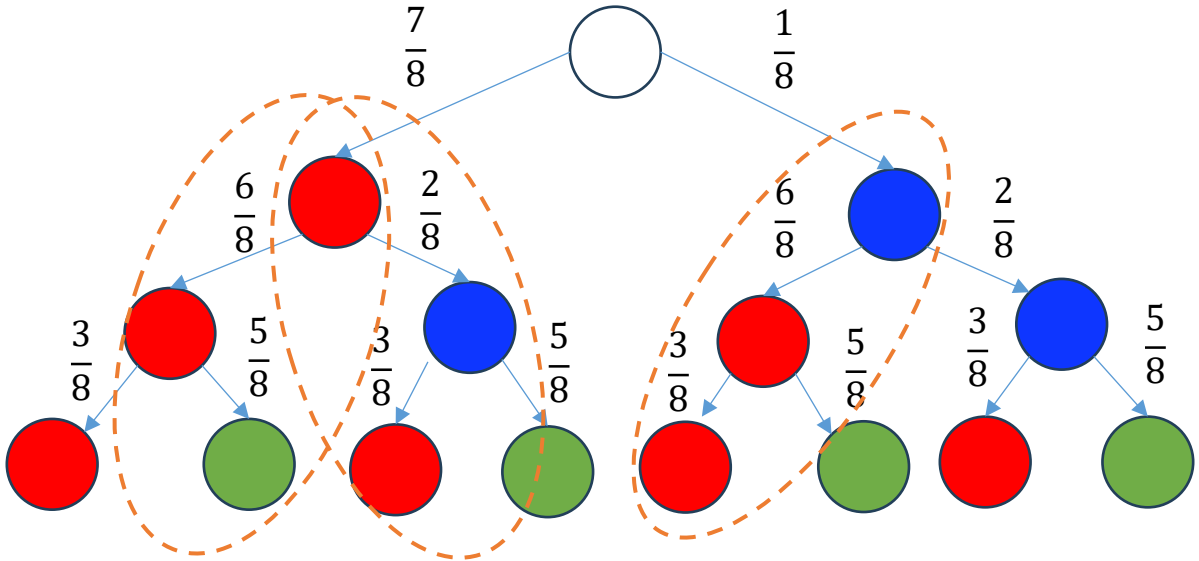
A. The temperature in Celsius is considered as interval data.
B. We can calculate the mean and standard deviation of temperature.
C. The temperature in Celsius is considered as ratio data.
D. None of the rest.
✓E. (a), and (b)
F. (a), and (c)



Question #: 14

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. (3 marks)

A. 226/64
B. 226/512
✓C. 270/512
D. 270/1024
E. None of the rest.



$$\frac{7}{8}\times\frac{6}{8}\times\frac{5}{8}+\frac{7}{8}\times\frac{2}{8}\times\frac{3}{8}+\frac{1}{8}\times\frac{6}{8}\times\frac{3}{8}=\frac{270}{512}$$

Question #: 15

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) P_2	0.2	0.4	(BLANK2) P_5

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)

1. Range - Min:0.19999 Max:0.20001
2. Range - Min:0.19999 Max:0.20001

$$\begin{cases} 0.1 + P_2 + 0.2 + 0.4 + P_5 = 1 \\ 1 \times 0.1 + 2 \times P_2 + 3 \times 0.2 + 4 \times 0.4 + 5 \times P_5 = 3.5 \end{cases}$$

$$\Rightarrow \begin{cases} P_2 = 0.1 \\ P_5 = 0.2 \end{cases}$$

THANK YOU