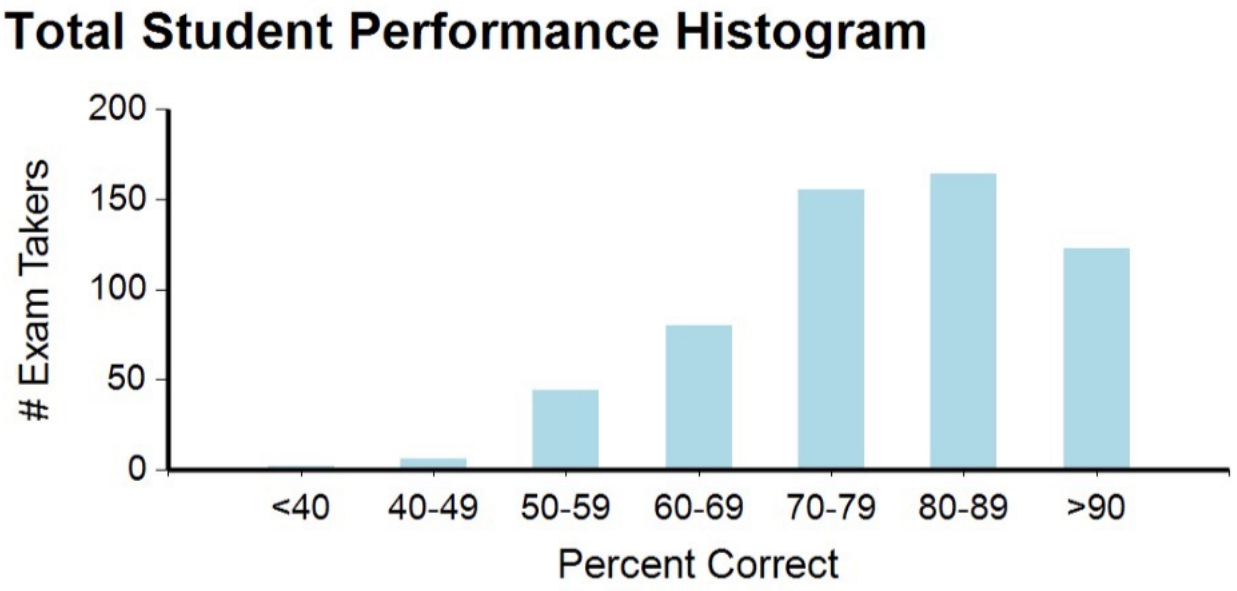
EE2211 AY22/23-Sem1

Midterm Summary
Most Wrongly-answered Questions

Some Statistics

- Average Score: 52.7/67 = 79%
- Highest Score: 67/67 = 100%
  - 11 students with full marks – Congrats!
- Expect for the questions listed here, for all other questions, we have a correct rate >= 75%!



Question1: FIB

We would like to design a linear regression function to predict cognitive decline in the elderly. We have collected two clinical features of 5 participants (x_1, x_2) that are related to future cognitive decline. We also have the five-year follow-up cognitive scores y of these 5 participants (see below).

$$\mathbf{X} = \begin{bmatrix} 50 & 10 \\ 40 & 7 \\ 65 & 12 \\ 70 & 5 \\ 75 & 4 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 9 \\ 6 \\ 5 \\ 3 \\ 2 \end{bmatrix}.$$

i) Perform least square estimation based on the existing training data. What is the mean of squared error of the estimated model? [1] (up to 4 decimal places)

ii) If we have a new participant with the two clinical features (x_1, x_2) = (62, 8). What is his/her five year follow-up cognitive score? [2] (up to 4 decimal places)

Number	Correct Answer
1	Range - Min: 1.226 Max: 1.227
2	Range - Min: 4.916 Max: 4.917

Correct Rate: 56%

Solution

- Over-determined system $\hat{\mathrm{w}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y}$
- Applying the formula of least square estimation
- Remember to add the bias term to X

```
import numpy as np

X = np.array([[1, 50, 10], [1, 40, 7], [1, 65, 12], [1, 70, 5], [1, 75, 4]])
Y = np.array([[9], [6], [5], [3], [2]])
w = np.linalg.inv(X.T @ X) @ X.T @ Y
Y_pred = X @ w
err = ((Y_pred - Y) ** 2).mean()
print(err)
X_test = np.array([[1, 62, 8]])
Y_test_pred = X_test @ w
print(Y_test_pred)
```

```
1.2267303102625298 MSE
[[4.91646778]] Y_test_pred
```

$\hat{\mathrm{w}}$ = [[9.14677804]
[-0.11634845]
[0.37291169]]

Question2: MCQ

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	4.5	4	3.5	3

Based on the least square regression line y = a x + b, please select the correct option below.

I) The estimated a* and b* minimize the sum of squared errors between the predicted y' and the target output y for all training samples.

II) The sum of squared errors between the predicted y' and the target output y for all training samples is called the loss function.

III) The estimation of y at x = 100 may not be accurate as the prediction range is far beyond the given data range.

IV) The estimation of y at x = 100 should be accurate as the prediction is linear.

Seq	Answer Choice	Correct
a)	I & III	☑
b)	III	☐
c)	I, II & III	☐
d)	I & IV	☐
e)	None of the above	☐

Correct Rate: 42%

Most selected option: c)

Solution

- 1) Right. Lecture 5 page 13
- 2) Wrong. The sum of squared errors between the predicted y' and the target output y for all training samples can be viewed as an **objective or cost function**. Lecture 5 page 13-15

  $(f_{\mathbf{w}}(\mathbf{x}_i) - \mathrm{y}_i)^2$ is called the **loss function**: a measure of the difference between $f_{\mathbf{w}}(\mathbf{x}_i)$ and y_i or a penalty for misclassification of example *i*.

  - All model-based learning algorithms have a **loss function**
  - What we do to find the best model is **to minimize the objective** known as the **cost function**
  - **Cost function** is a sum of **loss functions** over training set plus possibly some model complexity penalty (regularization)

- 3) Right. The range of x in the training set is too narrow.
- 4) Wrong. See point 3.
- Correct answer: a

Question3: True-False

A vector function has the mapping of $f(\mathbf{x}): \mathcal{R}^3 \rightarrow \mathcal{R}^2$. The gradient of this function with respect to **x** is a matrix of partial derivatives, which is a 2×3 matrix.

Correct	Answer Choice
◉	True
○	False

Correct Rate: 47%

A vector function has the mapping of $\boldsymbol{f}(\mathbf{x}): \mathcal{R}^3 \rightarrow \mathcal{R}^2$. The gradient of this function with respect to $\mathbf{x}$ is a matrix of partial derivatives, which is a 2×3 matrix.

- Correct. Vector-to-vector derivative requires the (partial) derivatives of each component of the output vector with respect to each component of the input vector, which in this case will contain 2 × 3 values
- Lecture 5 page 9

Derivative and Gradient

NUS National University of Singapore

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

h=2

d=3

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

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

Question4: MCQ

Please select the correct option.

Seq	Answer Choice	Correct
a)	Reinforcement learning concerns learning with a sequence of states, actions, and delayed rewards.	☑
b)	Classification is one type of supervised learning with continuous input.	☐
c)	Clustering is one type of unsupervised learning with continuous output.	☐
d)	(a), (b), and (c)	☐
e)	(a), and (b)	☐
f)	(a), and (c)	☐

Correct Rate: 42%

Most selected option: a)

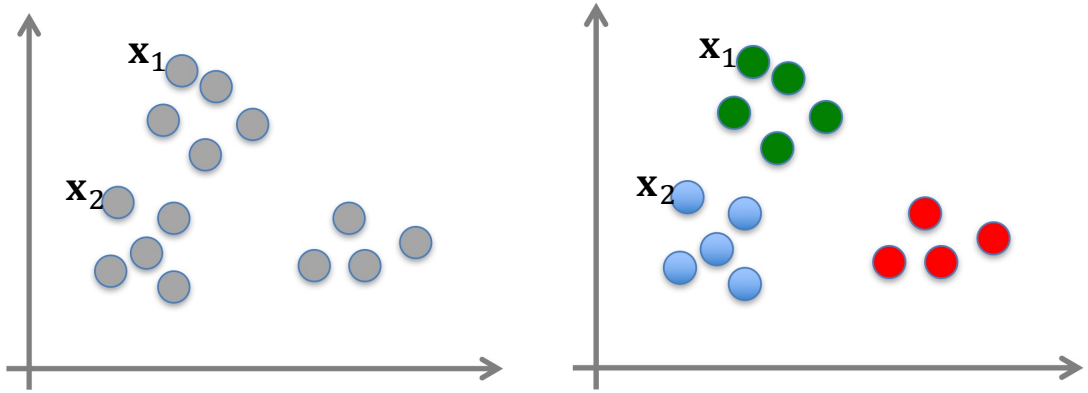
Solution

- Classification can also deal with discrete input.

Feature Extraction

Color	Size	Shape	Label
Blue	Large	Ring	Yes
Red	Large	Triangle	Yes
Orange	Large	Diamond	Yes
Green	Small	Circle	Yes
Yellow	Small	Arrow	No
Blue	Large	Rectangle	No
Red	Large	Circle	No
Green	Small	Diamond	No
Yellow	Large	Triangle	?

- Typically, the output format of clustering is a discrete clustering ID.



Question5: True-False

A machine-learning-based face detector is performing deductive reasoning to make a decision, i.e., to decide an image region corresponds to a human face or not.

Correct	Answer Choice
○	True
◉	False

Correct Rate: 57%

Solution

- Machine-learning-based face detectors require a set of training images for learning. The test images are typically not included in the training set. The model has to make inference on the new test images. The process is known as **inductive reasoning**.