2024_S2C_COMP4318_COMP5318_NE_FINALEXAM

● Graded

Student

Lihang Shen

Total Points

69 / 100 pts

Question 1

Question 1a

3 / 3 pts

✔ **+ 3 pts** A after normalisation:
[(2 - 2)/(4 - 2) , (8000 - 5000)/(9000 - 5000)] = [0, 0.75] (1 pt)
B after normalisation:
[(4 - 2)/(4 - 2) , (9000 - 5000)/(9000 - 5000)] = 1, 1
C after normalisation:
[(3 - 2)/(4 - 2) , (5000 - 5000)/(9000 - 5000)] = [0.5, 0] (1 pt)

+ 0 pts No answer or not corect

+ 0.5 pts Correct small part

Question 2

Question 1b

2 / 2 pts

✔ **+ 2 pts** Explanation of the Phenomenon (1pt): As the complexity of a supervised learning model increases, it fits the training data more closely, reducing the empirical risk. However, this often leads to overfitting, where the model learns noise in the data as patterns. Hence, while the model performs well on training data, its performance on new, unseen data worsens, causing the true risk to initially decrease but then increase.
Methods to Avoid Overfitting (1pt, list at least 2 methods to get full point)
Regularization
Cross-validation
Dropout, Early stopping
...

+ 1 pt Only explain why or only provide solution the phenomenon

+ 0.5 pts Only one solution are provided or just part of solution is correct

+ 0 pts No answer

Question 3

Question 2a

4 / 4 pts

+ 0 pts Wrong anser or no answer

+ 2 pts Correct computation of Manhattan distances for all training examples.
d = 7, 3, 3, 4, 2

+ 3 pts Correctly identifies the 3 nearest neighbors based on computed distances.
neighbour = (2,5,−1), (1,7,+1), (3,3,+1)

✔ **+ 4 pts** Correctly determines the majority class from the 3 nearest neighbors and provides the correct classification.
The test example (3,6) is classified as +1

Question 4

Question 2b

2 / 2 pts

+ 0 pts No answer/ wrong answer

+ 1 pt Correctly identifies the statement as false.

✔ **+ 2 pts** Provides a logical explanation highlighting why accuracy doesn't always increase as k grows.

Question 5

Question 3

2 / 5 pts

+ 0 pts No answer/ wrong answer

+ 1 pt Properly identifies the components from the problem description.

✔ **+ 2 pts** 1 correct:
Precision=80/100=0.8
Recall=80/140=0.57
F1=2.0.45/1.37=0.66
Acc=280/360=0.77

+ 3 pts 2 correct
Precision=80/100=0.8
Recall=80/140=0.57
F1=2.0.45/1.37=0.66
Acc=280/360=0.77

+ 4 pts 3 correct
Precision=80/100=0.8
Recall=80/140=0.57
F1=2.0.45/1.37=0.66
Acc=280/360=0.77

+ 5 pts all correct
Precision=80/100=0.8
Recall=80/140=0.57
F1=2.0.45/1.37=0.66
Acc=280/360=0.77

Question 6

Question 4a

2.5 / 4 pts

+ 4 pts The process is correct and the result is correct or the correct form of fraction

+ 3.5 pts The process is correct, but incorrect final result

✔ **+ 2.5 pts** Some minor mistakes in the process

+ 1 pt The backbone of the process is correct, but many mistakes of process.

+ 0 pts Didn't answer

Question 7

Question 4b

2 / 2 pts

✔ + 2 pts Correctly find there is a value not observed in the training data for a given class, mention using Laplace smoothing.

+ 1 pt Find the problem, but didn't use Laplace smoothing

+ 0 pts didn't answer or incorrect totally

Question 8

Question 5a

0.5 / 1 pt

+ 1 pt Correctly answer to ridge regression or l2 and the tole of lambda

✔ + 0.5 pts Only correctly answer one of them (name of loss function, the role of lambda)

+ 0 pts Didn't answer or incorrectly totally

Question 9

Question 5b

0.5 / 2 pts

+ 2 pts Correctly describe what is closed-form and gradient decent. And the 2 pros and 2 cons are correct.

+ 1 pt Some minor mistakes

✔ + 0.5 pts Most of answer is incorrect

+ 0 pts Totally incorrect

Question 10

Question 5c

0 / 3 pts

+ 3 pts Correct process and correct final result (
−1.9
0.7)

+ 2.5 pts The process is correct, but incorrect final result

+ 1 pt Only the backbone of the process is correct. With lots of mistakes.

✔ + 0 pts Totally incorrect or didn't answer

Question 11

Question 6a

1 / 2 pts

+ 2 pts Correctly answer the pros and cons of mini-batch GD and SGD

✔ + 1 pt Minor mistakes

+ 0 pts Totally incorrect or didn't answer

Question 12

Question 6b

2 / 2 pts

✔ **+ 2 pts** Correctly answer alpha is learning rate, and the affect of large and small learning rate.

If α is too large: (0.5 point)
The updates may overshoot the minimum, causing divergence.
The optimization process becomes unstable.
...
If α is too small: (0.5 point)
The convergence is very slow.
It may get stuck in poor local minima or plateaus.
...

+ 1 pt Minor mistakes, but correctly answer learning rate

+ 0 pts didn't answer or incorrect totally

Question 13

Question 6c

2 / 2 pts

✔ **+ 2 pts** Correctly answer the complexity and reasons of GD, MBGD and SGD.

+ 1 pt One of them are incorrect or major mistakes

+ 0 pts All of them are incorrect or didn't answer

Question 14

Question 7a

2 / 2 pts

+ 0 pts Incorrect answer

+ 0.5 pts Minor effort

+ 1 pt Incorrect output but true formula and great effort has been made

✔ **+ 2 pts** H(1) = -1/2 log(1/2)-1/2 log(1/2)=1
H(2) = -16/20 log(16/20)-4/20 log(4/20)=0.7219
H(3) = H(2) = 0.7219
E(1) = 1- 0.7219 = 0.2781

Question 15

Question 7b

0 / 2 pts

✔ **+ 0 pts** Incorrect answer or no answer

+ 0.5 pts Small effort has been made

+ 1 pt Case1: a = 8, b=2 --> Gain = 0

+ 1 pt Case2: a = 16,b=0: H(4) = H(5) = 0. So Gain = H(2) = 0.7219

Question 16

Question 7c

1 / 2 pts

+ 0 pts Incorrect answer

+ 0.5 pts Small effort has been made

✔ **+ 1 pt** Mention the goodness of split: distribution or attribute selection but not enough

+ 2 pts The information gain represents the goodness of split, if the split does not change the class distribution, the gain is zero. Otherwise, the gain is equal to the entropy of the node where the split occurs. It can also be used for attribute selection

Question 17

Question 8a

3 / 3 pts

+ 0 pts No answer or incorrect answer

+ 1 pt Partly correct

✔ **+ 1.5 pts** Dimensionality reduction is achieved by truncating the matrices U, Sigma,V to retain only the top K singular values and their corresponding vectors.

✔ **+ 1.5 pts** First K columns of U and first K rows of V_transpose will be kept

Question 18

Question 8b

1 / 2 pts

✔ **+ 0.5 pts** Partly correct for Advantages

+ 1 pt Advantages:

- Reduce model complexity
- Reduce overfitting

✔ **+ 0.5 pts** Parly correct for Disadvantages

+ 1 pt Disadvantages:

- Performing PCA can be computationally expensive.
- Loss of interpretability

+ 0 pts Incorrect or no answer

Question 19

Question 9a

1.5 / 3 pts

+ 3 pts The objective of SVM is to find the hyperplane that maximally separates data points of different classes while maximizing the margin between them.

+ 2 pts Mention maximizing the margin

+ 1.5 pts Mention about hyperplane to separate the data

+ 0.5 pts A small part is correct

+ 0 pts No answer or totally wrong answer

💬 **+ 1.5 pts** Point adjustment

Question 20

Question 9b

1 / 2 pts

+ 2 pts Support vectors are the data points closest to the decision boundary. They determine the position of the boundary, meaning removing them would alter the classifier.

+ 1 pt Support vectors are the data points closest to the decision boundary.

+ 1 pt They determine the position of the boundary, meaning removing them would alter the classifier.

+ 0.5 pts Small part is correct

+ 0 pts No or wrong answer

+ 1 pt Point adjustment

Question 21

Question 10a

2 / 2 pts

✔ + 2 pts The kernel allows SVM to transform data into a higher-dimensional space, making it possible to find a linear boundary for data that is not linearly separable in the original space.

+ 0 pts No or wrong answer

+ 1 pt Explanation is partially correct

Question 22

Question 10b

1 / 2 pts

+ 2 pts Regularization parameter (C) controls the trade-off between maximizing the margin and minimizing classification errors. A smaller C allows for a wider margin but more misclassified points, while a larger C tries to classify all points correctly with a narrower margin.

The following answers are also accepted to get full points:

- Small C means a very restricted model, where each point has a small influence. We can see this in the top left figure (the decision boundary is almost a line and the misclassified points do not have any influence on it)
- Big C means a less restrictive model and a bigger influence of all points. We can see how the decision boundary bends correctly to classify the previously misclassified points (compare the left top and left bottom figures).

+ 0 pts No or wrong answer

✔ + 1 pt Explanation is partially correct.

The provided answer is not sufficient.
Regularization parameter (C) controls the trade-off between maximizing the margin and minimizing classification errors. A smaller C allows for a wider margin but more misclassified points, while a larger C tries to classify all points correctly with a narrower margin.

Question 23

Question 11a

2 / 3 pts

+ 0 pts Incorrect

✔ **+ 1.5 pts** Random Forest as it uses bagging with decision trees + subset of features (1.5pt)

+ 1.5 pts Advantages: (1.5 pts, list at least 3 clear advantages to get full point)

- Robust to overfitting
- Ability to learn non-linear decision boundary
- Fast as only a subset of the features are considered
- Effective management of missing values

✔ **+ 0.5 pts** List only one correct advantage.

+ 0.5 pts Does not provide the correct name of the method. Lists some advantages, but they are not clearly articulated.

+ 1 pt Provides a list of some advantages, but the information is partially correct or overlapping.

+ 1 pt The name of the method is partially correct.

+ 1.5 pts Does not provide the correct name of the method. Lists some advantages, but they are not clearly articulated.

+ 0.5 pts Lists some advantages, but the information is partially correct and may include inaccuracies.

+ 1.5 pts Correct name of method provided, but the advantages listed are insufficient.

Question 24

Question 11b

2 / 2 pts

+ 0 pts Incorrect

✔ **+ 2 pts** It assigns the new example to the class corresponding to the leaf, then combine the decisions of the individual trees by majority voting.

+ 1.5 pts Partially correct.

+ 0.5 pts The answer is incorrect.

Question 25

Question 12

1 / 6 pts

+ 1.5 pts First Dense Layer (300 units):
Input size: 28 × 28 = 784 (since the input image is of size 28x28)
Parameters (weights): 784×300=235200
Biases: 300
Total parameters: 235200 + 300 =235500

+ 1.5 pts Second Dense Layer (200 units):
Input size: 300 (output of the previous layer)
Parameters (weights): 300 × 200 =60000
Biases: 200
Total parameters: 60000 + 200 = 60200 60000+200=60200

+ 1.5 pts Third Dense Layer (1 unit):
Input size: 200 (output of the previous layer)
Parameters (weights): 200 × 1=200
Biases: 1
Total parameters: 200+1=201

+ 1.5 pts Total Trainable Parameters:
First Dense Layer: 235500
Second Dense Layer: 60200 Third Dense Layer: 201
Total: 235,500 + 60200 + 201 = 295901 trainable parameters.

+ 0.5 pts Mentioning correctly small part

+ 0 pts No or totally wrong answer

+ 1 pt Point adjustment

Question 26

Question 13a

0 / 2 pts

+ 1 pt Ignore Subtle Features: Max pooling can discard smaller or subtle features that may still be relevant, as only the maximum value is considered.

+ 1 pt Sensitive to Noise: It might amplify noisy activations if the noise happens to have a higher value than other meaningful signals in the region.

+ 0 pts No answer or wrong answer.

+ 2 pts Full marks

Question 27

Question 13b

1 / 2 pts

+ 1 pt Loss of Prominent Features: Average pooling might dilute the impact of significant features (like edges or textures) by averaging them with less important ones.

+ 1 pt Less Effective for High-contrast Features: For tasks like object detection, it might not capture high-contrast, key features as effectively as max pooling.

+ 0 pts No or wrong answer

+ 2 pts Full marks

Question 28

Question 14

1.5 / 3 pts

+ 3 pts Full marks

✔ + 1.5 pts CNN leverages spatial hierarchies through local connectivity and shared weights.

+ 0 pts No or wrong answer

+ 1.5 pts CNNs use convolutional layers to focus on smaller regions of the image at a time, making them more efficient at detecting patterns like edges and textures

+ 1.5 pts Pooling layers in CNNs down sample feature maps, preserving essential information while reducing computational cost.

Question 29

Question 15a

2.5 / 2.5 pts

✔ + 2.5 pts LSTMs generally require more computational resources than RNNs due to their more complex architecture involving multiple gates and states. This additional complexity leads to increased memory usage and longer training times.

+ 1.5 pts Difference in terms of memory

+ 1.5 pts Difference in terms of gates

+ 0.5 pts Insufficient detail and explanation.

+ 0 pts No or wrong answer

Question 30

Question 15b

2.5 / 2.5 pts

✔ + 2.5 pts The gating mechanisms in LSTMs allow for selective information retention and updating, which improves their performance on tasks requiring long-term memory. RNNs lack these mechanisms, which can limit their ability to effectively learn and remember extended sequences.

+ 0 pts No or wrong answer

+ 1.5 pts The provided answer is partially correct. However, the explanation is insufficient

Question 31

Question 16a

1 / 2 pts

✔ + 1 pt Parly correct answer

+ 2 pts 1. Positional encoding in Transformers provides information about the position of tokens in a sequence.
2. Allows the self-attention mechanism to properly capture dependencies between words that are far apart in a sequence

+ 0 pts Incorrect or no answer

Question 32

Question 16b

1.5 / 2 pts

✔ **+ 0.5 pts** Parly correct answer

+ 1 pt 1. It allows the model to focus on different parts of the input sequence simultaneously.

✔ **+ 1 pt** 2. It can capture different types of dependencies and relationships between words or elements in a sequence.

+ 0 pts Incorrect answer

Question 33

Question 16c

1.5 / 2 pts

+ 1 pt Partly correct answer

✔ **+ 1.5 pts** 1. Greedy decoding: choose token with highest probability

+ 0.25 pts 2. Top-k sampling: restricts sampling to the top kkk tokens with the highest probabilities and normalizes their probabilities before sampling.

+ 0.25 pts 3. Top-p sampling: chooses from the smallest set of tokens whose cumulative probability exceeds a threshold ppp (e.g., 0.9).

+ 0 pts Incorrect answer

Question 34

Question 17

5 / 5 pts

+ 0 pts No answer/ wrong anser

✔ **+ 1 pt** Clear and concise explanation of the steps involved in the K-Means clustering algorithm.
initialization, assignment, update, repeat

+ 1 pt Correctly initializes centroids and starts calculations accordingly.
cluster 1: (1, 2)
cluster 2: (3, 4), (5, 6), (8, 8_

+ 2 pts Correct update:
centroid 1: (1, 2)
centroid 2: (5.33, 6)

+ 3 pts Correct assign in iteration 2:
cluster 1: (1, 2), (3, 4)
cluster 2: (5, 6), (8, 8)

✔ **+ 4 pts** Correct next centroids:
centroid 1: (2, 3)
centroid 2: (6.5, 7)

Question 35

Question 18a

3 / 3 pts

+ 0 pts Incorrect

+ 3 pts We calculate the probability of the observed sequence "Walk" on Day 1 and "Clean" on Day 2:
Day 1: Walk
Probability of Rainy on Day 1 and walking:
$\pi(\text{Rainy}) \times P(\text{Walk} | \text{Rainy}) = 0.6 \times 0.1 = 0.06$
Probability of Sunny on Day 1 and walking:
$\pi(\text{Sunny}) \times P(\text{Walk} | \text{Sunny}) = 0.4 \times 0.6 = 0.24$
Day 2: Clean
Given Day 1 was Rainy and walking:
Transition to Rainy and clean:
$P(\text{Rainy} \mid \text{Rainy}) \times P(\text{Clean} \mid \text{Rainy}) = 0.7 \times 0.5 = 0.35$
Transition to Sunny and clean:
$P(\text{Sunny} \mid \text{Rainy}) \times P(\text{Clean} \mid \text{Sunny}) = 0.3 \times 0.1 = 0.03$
Given Day 1 was Sunny and walking:
Transition to Rainy and clean:
$P(\text{Rainy} \mid \text{Sunny}) \times P(\text{Clean} \mid \text{Rainy}) = 0.4 \times 0.5 = 0.2$
Transition to Sunny and clean:
$P(\text{Sunny} \mid \text{Sunny}) \times P(\text{Clean} \mid \text{Sunny}) = 0.6 \times 0.1 = 0.06$
Total Probability for the Sequence "Walk" then "Clean":
Probability (Rainy on Day 1 and Walk, then Rainy on Day 2 and Clean) = $0.06 \times 0.35 = 0.021$
Probability (Rainy on Day 1 and Walk, then Sunny on Day 2 and Clean) = $0.06 \times 0.03 = 0.0018$
Probability (Sunny on Day 1 and Walk, then Rainy on Day 2 and Clean) = $0.24 \times 0.2 = 0.048$
Probability (Sunny on Day 1 and Walk, then Sunny on Day 2 and Clean) = $0.24 \times 0.06 = 0.0144$
Total Probability = $0.021 + 0.0018 + 0.048 + 0.0144 = 0.0852$

+ 1 pt Day 1: Walk
Probability of Rainy on Day 1 and walking:
$\pi(\text{Rainy}) \times P(\text{Walk} | \text{Rainy}) = 0.6 \times 0.1 = 0.06$
Probability of Sunny on Day 1 and walking:
$\pi(\text{Sunny}) \times P(\text{Walk} | \text{Sunny}) = 0.4 \times 0.6 = 0.24$

+ 1 pt Day 2: Clean
Given Day 1 was Rainy and walking:
Transition to Rainy and clean:
$P(\text{Rainy} \mid \text{Rainy}) \times P(\text{Clean} \mid \text{Rainy}) = 0.7 \times 0.5 = 0.35$
Transition to Sunny and clean:
$P(\text{Sunny} \mid \text{Rainy}) \times P(\text{Clean} \mid \text{Sunny}) = 0.3 \times 0.1 = 0.03$
Given Day 1 was Sunny and walking:
Transition to Rainy and clean:
$P(\text{Rainy} \mid \text{Sunny}) \times P(\text{Clean} \mid \text{Rainy}) = 0.4 \times 0.5 = 0.2$
Transition to Sunny and clean:
$P(\text{Sunny} \mid \text{Sunny}) \times P(\text{Clean} \mid \text{Sunny}) = 0.6 \times 0.1 = 0.06$

+ 1 pt Total Probability for the Sequence "Walk" then "Clean":
Probability (Rainy on Day 1 and Walk, then Rainy on Day 2 and Clean) = $0.06 \times 0.35 = 0.021$
Probability (Rainy on Day 1 and Walk, then Sunny on Day 2 and Clean) = $0.06 \times 0.03 = 0.0018$
Probability (Sunny on Day 1 and Walk, then Rainy on Day 2 and Clean) = $0.24 \times 0.2 = 0.048$
Probability (Sunny on Day 1 and Walk, then Sunny on Day 2 and Clean) = $0.24 \times 0.06 = 0.0144$
Total Probability = $0.021 + 0.0018 + 0.048 + 0.0144 = 0.0852$

✔ **+ 3 pts** Correct answer.

+ 2 pts Correct procedure followed, but the final result is incorrect.

+ 0.5 pts Incorrect calculations.

+ 1.5 pts Calculations are correct for Day 1, but incorrect for Day 2 and the final result.

Question 36

Question 18b

3 / 3 pts

+ 0 pts Incorrect

+ 3 pts Viterbi Algorithm Steps

Day 1: Observing "Walk"
For Rainy: π(Rainy)×P(Walk ∣ Rainy)=0.6×0.1=0.06
For Sunny: π(Sunny)×P(Walk ∣ Sunny)=0.4×0.6=0.24
The most likely state on Day 1 given the observation "Walk" is Sunny because 0.24 is greater than 0.06.
Day 2: Observing "Clean"
Transition from Rainy to Rainy and observing Clean:
0.06×P(Rainy ∣ Rainy)×P(Clean ∣ Rainy)=0.06×0.7×0.5=0.021
Transition from Sunny to Rainy and observing Clean:
0.24×P(Rainy ∣ Sunny)×P(Clean ∣ Rainy)=0.24×0.4×0.5=0.048
Transition from Rainy to Sunny and observing Clean:
0.06×P(Sunny ∣ Rainy)×P(Clean ∣ Sunny)=0.06×0.3×0.1=0.0018
Transition from Sunny to Sunny and observing Clean:
0.24×P(Sunny ∣ Sunny)×P(Clean ∣ Sunny)=0.24×0.6×0.1=0.0144
The most likely sequence for Day 2 given "Clean" is transitioning from Sunny on Day 1 to Rainy on Day 2 because 0.048 is the highest probability among the computed values. **The most likely sequence of weather conditions over the two days, given the observations "Walk" on Day 1 and "Clean" on Day 2, is Sunny on Day 1 followed by Rainy on Day 2.**

✔ **+ 3 pts** Answer is correct.

+ 1.5 pts The sequence is correct, but calculations are incorrect.

+ 1 pt Incorrect calculations and sequence.

+ 2 pts The calculations are correct, but the sequence is incorrect.

+ 0.5 pts The answer is incomplete.

+ 2.5 pts Correct calculations, but the sequence is incorrect.

+ 1.5 pts Incorrect calculations and sequence.

+ 2 pts The sequence is correct, but calculations are incorrect (minor).

Question 37

Question 19a

1.5 / 2 pts

+ 0 pts Incorrect

+ 2 pts Keywords: Task Objective, Data Labeling, Feedback Timing
Task Objective:

- RL: The goal is to learn a policy to maximize the cumulative reward over time, often through trial-and-error interactions with the environment.
- SL: The objective is to learn a mapping from inputs to outputs based on example input-output pairs, aiming to minimize error on this mapping.

Data Labeling:

- RL: Does not require labeled data as it learns from the rewards which are part of the environment's response.
- SL: Requires a dataset of labeled examples, where each example is pre-assigned a correct output by a supervisor.

Feedback Timing:

- RL: The feedback is delayed and given in the form of rewards, which may come after several steps or decisions (sparse and delayed rewards).
- SL: Immediate feedback is provided through labeled data, where the correct output (label) for each input is known at the time of training.

+ 1 pt **Task Objective**:

- RL: The goal is to learn a policy to maximize the cumulative reward over time, often through trial-and-error interactions with the environment.
- SL: The objective is to learn a mapping from inputs to outputs based on example input-output pairs, aiming to minimize error on this mapping.

+ 1 pt **Data Labeling**:

- RL: Does not require labeled data as it learns from the rewards which are part of the environment's response.
- SL: Requires a dataset of labeled examples, where each example is pre-assigned a correct output by a supervisor.

+ 0 pts The answer is insufficient.

✔ **+ 1.5 pts** List one sufficient difference between RL and SL.

+ 2 pts List two sufficient differences between RL and SL.

+ 2 pts List three sufficient differences between RL and SL.

+ 0.5 pts The differences between RL and SL are unclear.

Question 38

Question 19b

2 / 2 pts

+ 0 pts Incorrect

+ 2 pts Deep Q Learning with Experience Replay has significant advantages:

- **Increased Sample Efficiency**: By storing the agent's experiences and then randomly sampling from them to train the network, each piece of experience can be used in multiple updates, improving learning efficiency.
- **Enhanced Training Stability**: Experience replay helps in breaking the correlation between consecutive training samples by mixing old and new experiences, thus preventing the network from overfitting to recent trends in the environment and leading to more stable learning.

+ 1.5 pts The answer is partially correct.

+ 1 pt The answer is too general. Should provide a more specific explanation of how Experience Replay benefits Deep Q-Learning.

✔ **+ 2 pts** The answer is correct.

+ 1 pt The answer is partially correct.

Question 39

Question 19c

2 / 2 pts

+ 0 pts Incorrect

+ 2 pts Keywords: Exploration, Learning Rate, Network Architecture
If a Deep Q-learning agent's performance isn't improving as expected, consider the following issues and solutions:
- **Insufficient Exploration**: The agent might not be exploring the environment sufficiently, relying too much on exploiting known strategies. Solution: Adjust the exploration policy, e.g., by increasing the epsilon value in an epsilon-greedy policy or extending the duration of the exploration phase.

- **Inappropriate Learning Rate or Reward Signal**: If the learning rate is too high or too low, or if the reward signals are not informative enough, the agent might fail to converge to an optimal policy. Solution: Tune the learning rate and investigate the reward structure for effectiveness and adequacy.
- **Network Architecture or Overfitting**: The neural network architecture may not be appropriate, or the agent might be overfitting to the specific features of the training environment. Solution: Experiment with different network architectures or add regularization techniques.
  By addressing these potential issues, you can enhance the learning efficiency and overall performance of your Deep Q-learning agent.

✔ **+ 2 pts** The answer is basically correct.

+ 1.5 pts The answer is partially correct but lacks details.

+ 0.5 pts The answer is not clear.

THE UNIVERSITY OF SYDNEY

Room Number LG17

Seat Number 11

Student Number 490051481

ANONYMOUSLY MARKED

(Please do not write your name on this exam paper)

CONFIDENTIAL EXAM PAPER

This paper is not to be removed from the exam venue

Computer Science

EXAMINATION

Semester 2 – Final Exam, 2024

COMP4318/5318 Machine Learning and Data Mining

EXAM WRITING TIME: 2 hours

READING TIME: 10 minutes

EXAM CONDITIONS:
1. Closed book: no reference materials/resources are permitted

MATERIALS PERMITTED IN THE EXAM VENUE:
(No electronic aids are permitted e.g. laptops, phones)

Calculator – non-programmable

MATERIALS TO BE SUPPLIED TO STUDENTS:
None

INSTRUCTIONS TO STUDENTS:
Answer all questions using blue or black pen (not pencil), on this examination paper in the spaces provided.

For Examiner Use Only

Q	Mark	Q	Mark
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10			

Total _____

Please tick the box to confirm that your examination paper is complete. ☐

Question 1. General questions (5 points)

a) Given a dataset with only three data points A, B and C, each described by 2 features:

A = [2, 8000] B = [4, 9000] C = [3, 5000]

After normalisation, what is the value of A, B and C? (3 points)
The formula for normalisation is provided below, where x is the value of a feature.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

a_1 2
b_1 4
c_1 3

$\frac{2-2}{4-2} = 0$ for a_1 $\frac{3-2}{4-2} = 0.5$ for c_1

$\frac{4-2}{4-2} = 1$ for b_1

a_2 8000
b_2 9000 max
c_2 5000 min

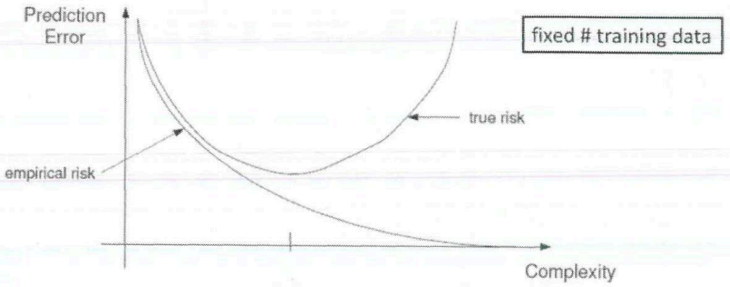
$\frac{8000-5000}{9000-5000} = \frac{3000}{4000} = 0.75$ for a_2

$\frac{9000-5000}{9000-5000} = 1$ for b_2

$\frac{5000-5000}{9000-5000} = 0$ for c_2

A = [0, 0.75] B = [1, 1] C = [0.5, ~~5~~ 0]

b) In Supervised Learning, given a fixed number of training examples, as the complexity of the model grows, the empirical risk decreases; however, the true risk first decreases and then increases (as shown in the figure below). (2 points)



Explain why and provide two different ways to avoid this phenomenon.

- This is called "overfitting", it means ~~rely~~ depend too much on the training set therefore the model only perform well for the training data set and perform badly in real data
- There are two ways to avoid it:
  ① in logistic regression, we can add a standardization term, ~~the coeffic~~
  ② in Decision Tree, we can prune tree
  ~~③ split the training data into 10 fold, and using~~
  ~~to~~

Question 2. kNN (6 points)

a) Consider a set of five training examples given as $((x_1^{(i)}, x_2^{(i)}), y^{(i)})$, $i = 1, \ldots, 5$, where $x_1^{(i)}$ and $x_2^{(i)}$ are the two attribute values (positive integers) and $y^{(i)}$ is the binary class label:

((1,1),-1), ((1,7),+1), ((3,3),+1), ((5,4),-1), ((2,5),-1).

Classify a test example (3, 6) using a k-NN classifier with k = 3 and Manhattan distance. Show your calculations. (4 points)
Manhattan distance is defined by: $d((u, v), (p, q)) = |u - p| + |v - q|$.

1	1	-1	①
1	7	1	②
3	3	1	③
5	4	-1	④
2	5	-1	⑤

k=3 Manhatan 3,6

①: $|3-1|+|6-1|=7$

② $|3-1|+|6-7|=3$ ✓ 1

③ $|3-3|+|6-3|=3$ ✓ 1

④ $|3-5|+|6-4|=4$

⑤ $|3-2|+|6-5|=2$ ✓ -1

Vote by majority: the answer is 1

b) Is the following statement TRUE or FALSE: "As the value of k used in a k-NN classifier is incrementally increased from 1 to n (n is the total number of training examples), the classification accuracy on the training set will always increase"? Explain your answer in 2 or 3 sentences. (2 points)

False. the proper number of k should less than $\sqrt{n}$. In addition, the example in the tut already prove this idea is incorrect. The reason behind this is that the majority class will always dominate the result when k=n. You can see in last question (Q3(a)), if n=k=5, then no matter what new data point is, it will always be classify as class "-1" since there are 3 "-1" and 2 "1"

Question 3. Performance Metrics. (5 points)

Given a dataset consisting of 360 images, with 220 images labeled as chickens and 140 images labeled as dogs. A k-Nearest Neighbors (kNN) algorithm is applied to classify these images. Out of the total images, the model predicts 100 images as dogs. Of these, 80 images are correctly labeled as dogs, while 20 images are misclassified as chickens. Calculate Precision, F1 Score, Recall, and Accuracy of the model.

ACTUAL VALUES	POSITIVE	NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

dog chick

Predict T F

dog Real T 80 0

chick F 20 0

$A = \frac{80}{100} = 0.8$

$P = \frac{80}{100} = .8$

$R = \frac{80}{80} = 1$

$F1 = \frac{0.8}{1.8} \times 2 = 0.889$

Question 4. Naïve Bayes (6 points)

Given the following dataset where loan default is the class.

	home owner	marital status	income (in K)	loan default
1	yes	single	120	yes
2	no	married	100	no
3	yes	single	110	no
4	no	married	100	yes
5	no	single	90	yes
6	yes	married	200	no
7	no	divorced	140	no
8	no	married	70	yes
9	yes	divorced	90	no
10	yes	single	85	yes

Bayes Theorem:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True

Probability density function for a *normal* distribution with mean μ and standard deviation σ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

where,

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}}$$

a) Predict the class of the following new example using Naïve Bayes:

E: home owner = yes, marital status = married, income= 130

Show your calculations. (4 points)

$\mu_{income} = \frac{120 + 100 + 90 + 70 + 85}{5} = 93$ [Yes] $n = 5$

$\sigma_{income} = \sqrt{\frac{(120-93)^2 + (100-93)^2 + (90-93)^2 + (70-93)^2 + (85-93)^2}{4}} = \sqrt{345} = 18.57$

$P(h_0 = Y | Yes) = \frac{2}{5}$ (1, 13) $P(ms = married | Yes) = \frac{2}{5}$ (4, 8) $P(income = 130 | Yes) =$

$\frac{1}{\sqrt{2\pi} \times 18.57} e^{-\frac{(130-93)^2}{2 \times 345}} = 2.95 \times 10^{-3}$

$P(Yes | E_i) = \frac{\frac{5}{10} \times \frac{2}{5} \times \frac{2}{5} \times 2.95 \times 10^{-3}}{P(E)}$ # $= \frac{59}{25000}$

$\mu_{income} = \frac{100 + 110 + 200 + 140 + 90}{5} = 128$ [No] $n = 5$

$\sigma_{income} = \sqrt{\frac{(100-128)^2 + (110-128)^2 + (200-128)^2 + (140-128)^2 + (90-128)^2}{4}} = 44.38$

$P(h_0 = Y | No) = \frac{3}{5}$ (3, 9) $P(ms = M | No) = \frac{2}{5}$ (2, 6) $P(income = 130 | No) =$

$\frac{1}{\sqrt{2\pi} \times 44.38} e^{-\frac{(130-128)^2}{2 \times 44.38^2}} = 8.78 \times 10^{-3}$

$P(No | E_i) = \frac{\frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} \times 8.78 \times 10^{-3}}{P(E)} = 1.0776 \times 10^{-3}$

$1.0776 \times 10^{-3} - \frac{59}{25000} > 0 \rightarrow P(No | E_i) > P(Yes | E_i) \rightarrow$ Hence It is [No]

b) Given a new example: *home owner = yes, marital status = divorced, income= 130.*

What is the probability *P(marital status|yes)*? What could be the issue with that probability value and how to address it? (2 points)

We need to use "laplace correction" method.

$$P = \frac{0+1}{5+3} = \frac{1}{8}$$

↳ Since marital status have 3 different classes single, married, Divorced

Question 5. Linear Regression (6 points)

In Linear Regression, given the following cost function:

$$f(w) = min_w \frac{1}{2}\|y - Xw\|^2 + \lambda\|w\|^2,$$

where $X \in R^{n\times(m+1)}$, $y \in R^n$, $w \in R^{m+1}$, n is the number of samples, m is the number of features.

a) What is the name of the loss function above, and what is the role of λ? (1 points)

. ~~Cross Entropy~~ loss function,
least square residual

. λ is smaller means weak regularization term, so easy to overfit

λ is bigger means stronger regularization term, so easy to underfit

. we need to choose a ~~proper~~ λ

b) To update the weight of the model, we can use gradient descent methods or the closed-form solution $w = (X^T X + \lambda I_m)^{-1} X^T y$. What are they? List two pros and cons of using gradient descent compared with the closed-form solution. (2 points)

. Closed form derive from matrix representation. The time complexity is $O(nk^2 + k^3)$

. Gradient descent time complexity is $O(nk)$

Cons:

① GD (gradient descent) cost less with $O(nk)$ time complexity

② easier to use

c) Assume that $\lambda = 0$ and the Gradient Descent update for the above loss function is:

$$w_{t+1} = w_t - \alpha \nabla f(w_t), \text{ where } \nabla f(w) = (X^T(Xw - y)),$$

and the dataset includes only 2 samples: $X = \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}, y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$.

If the initial weight $w_0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, what is w_1 after one iteration update? (Assume that $\alpha = 0.1$) (3 points)

$W_1 = \frac{1}{2} - 0.1 \times ($

Question 6. Gradient Descent (6 points)

Given the following update rules for gradient descent-based optimization methods:

- GD Update: $w_{t+1} = w_t - \alpha \sum_{j=1}^{n} \left(w_t^\top x^{(j)} - y^{(j)}\right) x^{(j)}$
- Mini-batch GD Update: $w_{t+1} = w_t - \alpha \sum_{j=1}^{b} \left(w_t^\top x^{(j)} - y^{(j)}\right) x^{(j)}$
- SGD Update: $w_{t+1} = w_t - \alpha \left(w_t^\top x^{(j)} - y^{(j)}\right) x^{(j)}$ for any $j = 1, \ldots, n$

a) List two pros and cons of using Mini-batch GD and SGD. (2 points)

pros: ① cost less than Gradient descent

cons: ① may get wrong direction when use SGD

② may need more iterations

b) What is α in the above formulas? What if it is too big or too small? (2 points)

· It is the learing rate

· Big learning rate may lead to hard converg

small α may lead to converg slowly

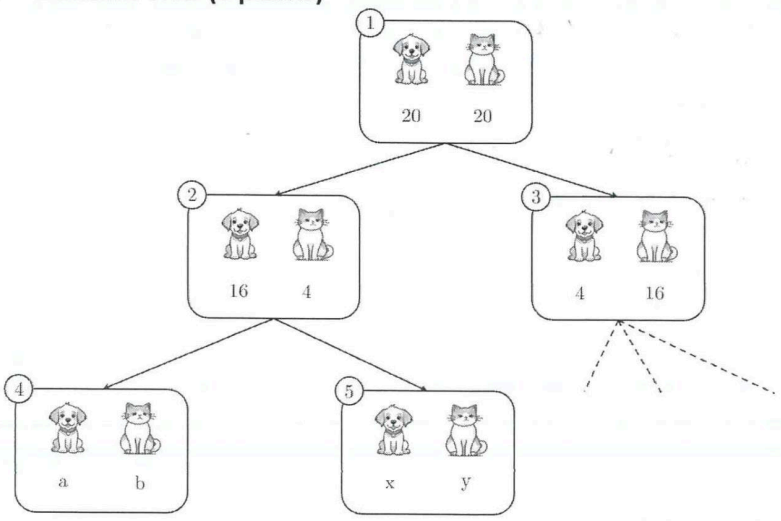
c) Given GD's complexity is O(nk), where n is the number of data samples and k is the number of features. What is the complexity of Mini-batch GD and SGD? (2 points)

SGD: $O(1<)$

Minibatch : $O(bk)$

↳ "b" is the batch size

Question 7. Decision Tree (6 points)



The figure above is an example of the decision tree for a classification problem. The numbers in each node represent the number of samples in each class (Dog/Cat) at that node. For example, in the original dataset (node number (1)), there are 20 Dogs and 20 Cats.

A split at node (1) gives us nodes (2) and (3) with a given number of samples in each class. In fact, the values in node (3) can be inferred based on the values in node (2) and vice versa. Suppose we have some split (D) at node (2), which gives us nodes (4) and (5), and the model decides that no split will be implemented at (4) and (5).

Entropy H(S) of node S is computed by:

$$H(S) = -\sum_i P_i . log_2 P_i$$

With P_i be the proportion of class i in node S

Information Gain of the split at the node S is computed by:

$$Gain(S) = H(S) - \frac{n_k}{n_S}\sum_k H(k)$$

a) Compute the information gain for the split at node (1) (2 points)

$H(S) = -\frac{20}{40}\log_2\frac{20}{40} - \frac{20}{40}\log_2\frac{20}{40} = 1$

$H(S_{left}) = -\frac{16}{20}\log_2\frac{16}{20} - \frac{4}{20}\log_2\frac{4}{20} = 0.723$

$H(S_{right}) = -\frac{16}{20}\log_2\frac{16}{20} - \frac{4}{20}\log_2\frac{4}{20} = 0.723$

$Gain = 1 - 0.723 \times \frac{1}{2} - 0.723 \times \frac{1}{2} = 0.277$

b) Compute the information gain for the split (D) at node (2), with
a = 8, b =2
a = 16, b = 0
(2 points)

$H(S) = 0.723$

$H(S_1) = -\frac{8}{10}\log\frac{8}{10} - \frac{2}{10}\log\frac{2}{10} = 0.7219$

$H(S_2) = 0$

$Gain = 0.723 - \frac{10}{26} \times 0.7219 = 0.44533$

c) From the values obtained in question b), derive your conclusion about the meaning of information gain. (2 point)

Information gain means: whether we succefully divide the data into different classes. As you may notice, "$H(S)$" is the Entropy of parent node. "$-\sum H(K)$" is the sum of Entropy for children nodes. If ~~parent~~ information Gain is small, it means we do not split the data correctly. If Gain is big, then it means we successfully split the data. By the way, small Entropy means there are less different label in a dataset.

$H(S) - \underbrace{\sum H(K)}$

need this to become small

Question 8. Dimensionality Reduction and PCA (5 points)
Consider principal component analysis (PCA) as a technique for dimensionality reduction and answer the following questions:

a) The principal components can be obtained by applying singular value decomposition (SVD), i.e. $X = U \times \Lambda \times V^T$ where X is the data matrix, U and V are orthogonal matrices, and Λ is a diagonal matrix. Explain how dimensionality reduction can be achieved using SVD decomposition. (3 points)

It is very simple. ~~refer~~refer to matrix multiplication rule, when

$$X_{n\times m} = U_{n\times m} \Lambda_{m\times m} V_{m\times m} \Longrightarrow X_{n\times k} = U_{n\times k} \Lambda_{k\times k} V_{k\times k}$$

X dimention will decrease when $k < m$. By the way, k is chosen by using Elbow method.

b) What are the advantages and disadvantages of using PCA for image compression? (2 points)

·Pro: can significantly reduce cost when there are too many attributes.

Cons: It may lead to the loss of some interesting attributes. (since we usually only keep 95% of principle components)

Question 9. Support Vector Machine (5 points)
Given the SVM algorithm in the following figure:

SVM Algorithm

Our separating hyperplane is *H*

H is in the middle of 2 other hyperplanes, *H1* and *H2*, defined as:

$H_1 : \mathbf{w} \cdot \mathbf{x} + b = 1$

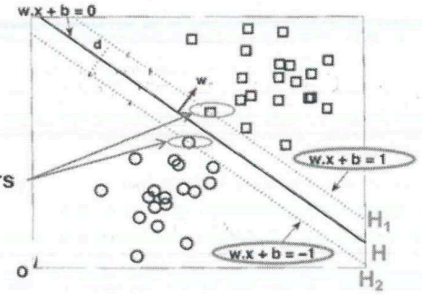
$H2 : \mathbf{w} \cdot \mathbf{x} + b = -1$

The points laying on *H1* and *H2* are the support vectors

d is the margin of *H*

$$d = \frac{2}{\|\mathbf{w}\|}$$

Support vectors



Given N training examples $(\mathbf{x}_i, y_i), i = 1,..,N$

$\mathbf{x}_i = (x_{i1},...,x_{im})^T, y_i = \{-1,1\}$

training vector class

Minimize $\frac{1}{2}\|\mathbf{w}\|^2$ (1)

Subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \forall i$ (2)

a) What is the objective of SVM in classification problems (1) given the constraint (2)? (2.5 points)

~~It means we need to classify items.~~

"$(W \cdot X_i + b)$" means the predict value

"Y_i" means the target value

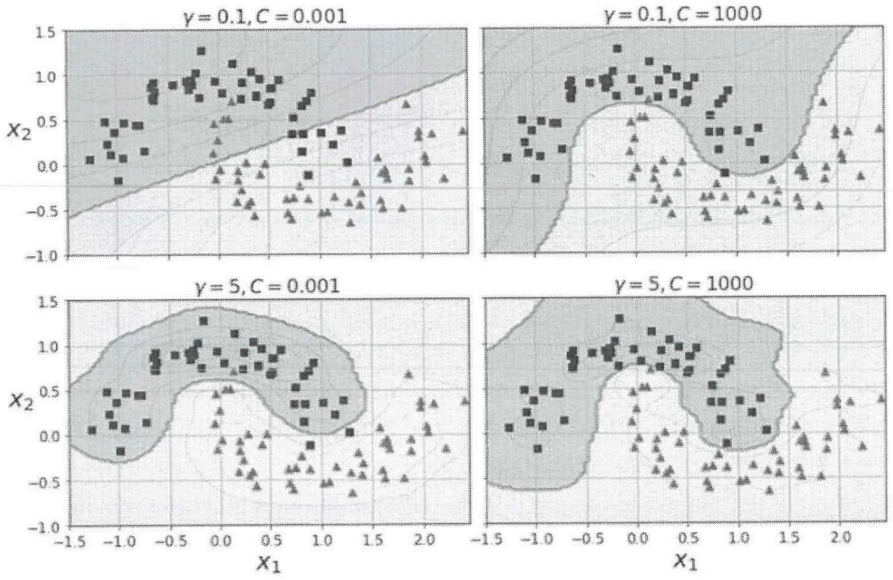
Since this is a binary classification problem (~~either 0 or 1~~)

So it means we need to ~~classify~~ successfully classify every data points.

b) Why are support vectors important in SVMs? Explain in 2 or 3 sentences (2 points)

- "W" is a linear combination of support vectors.
- good support vectors can give us max margin ~~which~~ can ~~increase the ability to help model have robust ability for minor change in data set~~
- $W_k = \sum \lambda_i y_i x_{ik}$
  → x_{ik} is the support vector when λ_i is not 0

Question 10. The figure of running non-linear SVM on the Moon dataset with Radial-Basis Function (RBF) kernel with different values of C and gamma are given in the following: **(4 points)**



30

a) What are the purposes of using kernel in SVM? Explain in 2 or 3 sentences. (2 points)

- Sometimes we need to do some non-linear ~~regres~~ classification
- We do not want to do dot product in heigher dimension space since it is cost or C.o.t
- Hence we use kernel trick to simplify it

b) What role does the regularization parameter (C) play in SVM? Explain in 2 or 3 sentences (2 points)

~~Large~~ Larger C means allow less data points to be misclarified

Smaller C means more data points can be misclarrified.

Question 11. Ensemble Methods (5 points)
The pseudocode below describes the training process of an ensemble model:

Parameters:

n - number of training examples, *m* – number of all features, *k* – number of features to be used by each ensemble member (k<m), *M* – number of ensemble members

Model generation:

For each of *M* iteration

1. Generate a bootstrap sample:
   Sample *n* instances with replacement from training data

2. Random feature selection for selecting the best attribute:
   Grow decision tree without pruning. At each step, select the best feature to split on by considering only *k* randomly selected features and calculating information gain.

Based on the information given in the pseudocode, answer the following questions.

a) What is the exact name of the method described in the pseudocode, and what are its advantages (list 2-3 advantages)? (3 points)

· Random Forest

Pros:

① It usually give us better prediction when the ~~base~~ weak classifier have error rate less than 50%

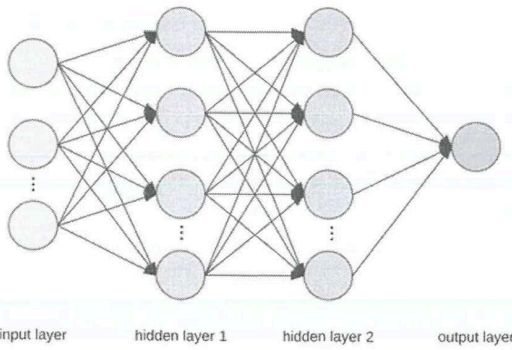
② it significantly reduce the correlation among weak learners.

b) How to employ the method described in the pseudocode to classify a new example? Given that the new example will be applied to each of the built decision trees starting from the root. (2 points)

We generate M weak learners by using M different bootstraps training data. In addition, for each weak learner, we choose different subset of attributes to train. Finally use each weak decision tree to predict the result, and we use majority vote to make the final decision.

Question 12. Neural Networks (6 points)

Given the following Keras Sequential model:




```
model = keras.models.Sequential([
keras.layers.Flatten(input_shape=[28, 28]),
keras.layers.Dense(300, activation="tanh"),
keras.layers.Dense(200, activation="tanh"),
keras.layers.Dense(1, activation="sigmoid") ])
```

28x28 = 784 attributes

Calculate the total number of trainable parameters in the model.

From input layer, we have

$28^2 = 784$

Fully connected:

$784 \times 300 \times 200 \times 1 = 47,040,000$

Question 13. Convolutional Neural Network (4 points)

a) What are the drawbacks of Max Pooling? (2 points)

It works well in the dark background since white is 255

b) What are the drawbacks of Average Pooling? (2 points)

It works well in lighter background.

Question 14. Why are Convolutional Neural Networks (CNNs) more effective than traditional Deep Neural Networks (DNNs) for image processing tasks? (3 points)

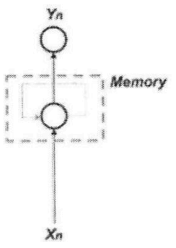
CNN is good at identifying shift image

1 ≠= 1 for CNN

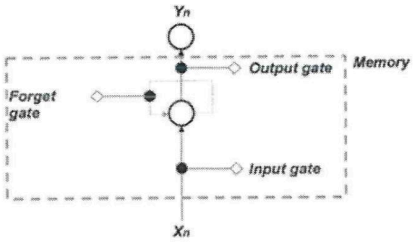
since it has Conv layer

Question 15. Recurrent Neural Networks (5 points)

The two figures below are standard RNN and LSTM units, respectively.



Standard RNN



Long Short-Term Memory

a) How do RNNs and LSTMs differ in their computational requirements? (2.5 points)

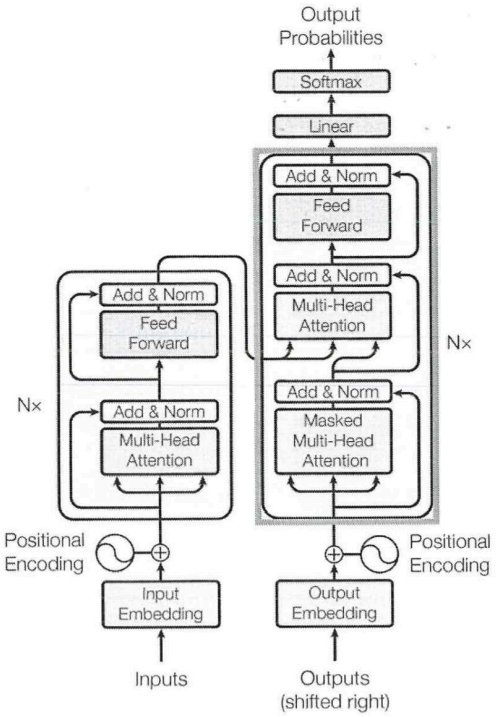
- RNN does not forget during recursion time, require memory space. Using hmm
- LSTM forget ~~&~~ information when pass through forget gate. ~~Enable~~ only using output gate, C_t, Sigmoid(h_{t-1}, X_t)

b) What impact does the gating mechanism in LSTMs have on their performance compared to RNNs? (2.5 points)

- LSTM performs well in long term memory since it will forget unimportant information.
- Besides LSTM does not need to worry about gradient vanishing problem because it does not need to multiply h_{mm} ~~in every &~~ every times.

Question 16. Transformer (6 points)

Given the architecture of a Transformer model as follows.



a) What is the purpose of positional encodings in the Transformer model? Explain in 2 or 3 sentences. (2 points)

It just describe the ~~place~~ location of each token. we will use it when we output.

Such as "I am happy" will not become "I happy am"

b) How does multi-head attention within the Transformer model enhance its ability to process input sequences compared to single attention heads? Explain in 2 or 3 sentences. (2 points)

- Multihead just parallelly run several single head attention.
- ~~We~~ the model conclude the output of all single head attention to get the final result. (I think it likes ensemble)

Vote by majority

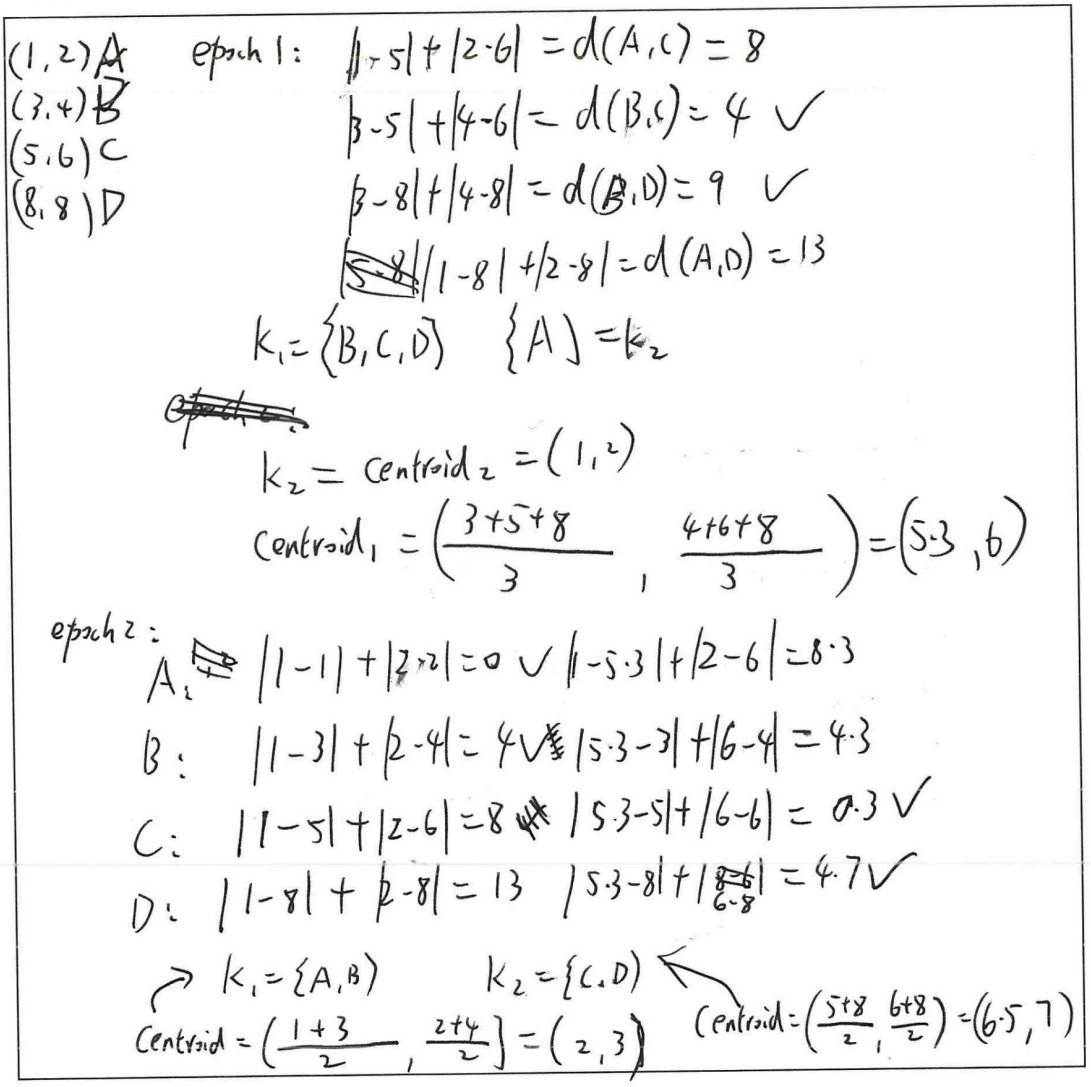
c) In the inference phase of a Transformer model, how is the next token chosen based on the softmax vector output? Explain in 2 or 3 sentences. (2 points)

The ~~heigher~~ token with heigher probability will be chose.
Say [0.1, 0.1, 0.8] then the token with 0.8 will be chose.

Question 17. K-mean clustering (5 points)

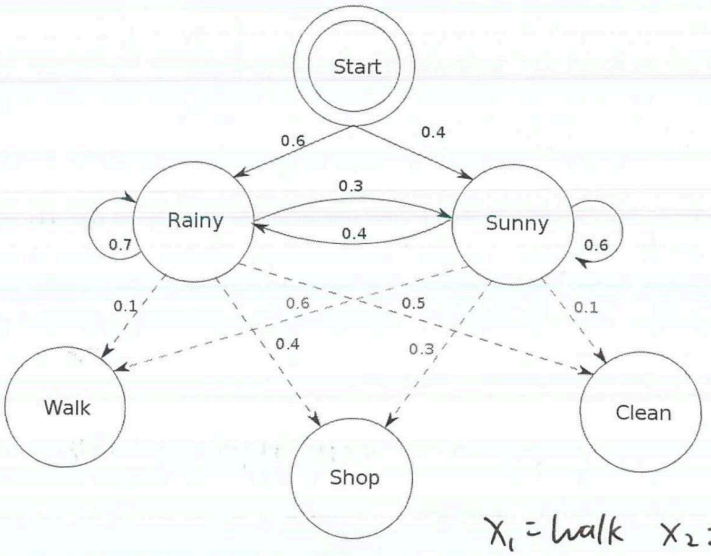
Explain the K-means clustering algorithm and perform clustering using K-means for the following dataset of 2-dimensional four data points: (1, 2), (3, 4), (5, 6), (8, 8). Follow these steps: Choose k = 2 and initialize the centroids with (1, 2), (3, 4). Assign each point to the nearest centroid. Update the centroids based on the new clusters. Repeat the process until the centroids stabilize. Provide the resulting clusters and the positions of the centroids after 2 iterations. Use the Manhattan distance formula to calculate the distances between points.

Manhattan distance is defined by: $d((u,v),(p,q)) = |u - p| + |v - q|$.



Question 18. Hidden Markov Models (6 points)

We use the following Hidden Markov Model (HMM) to understand the impact of weather on Alice's daily activities. Assume the model has two hidden states: "Rainy" and "Sunny," which influence Alice's activities. The activities we observe include "Walking," "Shopping," and "Cleaning."



X_1 = walk X_2 = clean

Consider the following observations: On the first day, Alice was observed walking, and on the second day, she was observed cleaning.

a) Calculate the probability of observing this sequence of activities. Show your calculations. (3 points)

Time 1:

$P_R(0) = A^*(R)\, e_R(x_1) = 0.6 \times 0.1 = 0.06$

$P_S(0) = A^*(S)\, e_S(x_1) = 0.4 \times 0.6 = 0.24$

$P_R(1) = e_R(x_2)\left[P_R(0) \times a_{R\to R} + P_S(0) \times a_{S\to R}\right] = 0.5 \times \left[0.06 \times 0.7 + 0.24 \times 0.4\right]$

$= 0.069$

$P_S(1) = e_S(x_2)\left[P_S(0) \times a_{S\to S} + P_R(0) \times a_{R\to S}\right] = 0.1 \times \left[0.24 \times 0.6 + 0.06 \times 0.3\right]$

$= 0.0162$

$P = P_S(1) + P_R(1) = 0.069 + 0.0162 = 0.0852$

b) Determine the most likely sequence of underlying weather conditions (hidden states) for these two days. Show your calculations. (3 points)

$V_R(0) = 0.7 \quad V_S(0) = 0.24$

$V_R(1) = e_R(x_2) \max\{0.4 \times 0.7, 0.24 \times 0.4\} = 0.5 \times 0.096 = 0.048$

0.042 0.096

$Ptr_R(1) = Sunny$

$V_S(1) = 0.1 \times \max\{0.24 \times 0.6, 0.4 \times 0.3\} = 0.0144$

0.144 0.018

$Ptr_S(1) = Sunny$

$\pi_2 = Argmax\{V_R(1), V_S(1)\} = Rainy$

$\pi_1 = Ptr_{Rainy}(1) = Sunny$

sunny → Rainy

Question 19. Reinforcement Learning (6 points)

a) What is the main difference between reinforcement learning and supervised learning in terms of learning objectives? List 2 or 3 main differences. (2 points)

Re

- make decision
- do not need labeled data

Super

- classify data
- need labeled data

b) What are the advantages of Deep Q Learning with Experience Replay over the typical Deep Q Learning? Explain in 2-3 sentences. (2 points)

- It will not face divergence problem after apply Experience replay. Since original data is highly related, but after using this method, it's correlationship will decrease

c) You have used a Deep Q-learning approach to train a game-playing agent, but the agent's performance is not improving as expected. What could be the potential issues causing this problem, and how might you address them to improve the agent's learning efficiency and performance? Explain in 2-3 sentences. (2 points)

- The reward function may be not good enough
- To overfitting
- We can reset our reward, such as every step it will have -1 reward + immediate reward.
- we can also set some ε to avoid overfitting.

END OF EXAMINATION

THIS PAGE LEFT INTENTIONALLY BLANK.

Use these pages for additional writing space if you run out of space in an answer area. If you want any writing on this page to be marked, you MUST write "see additional page" in the relevant answer area and clearly indicate the question number you are answering on this page. Any writing outside of this booklet will not be examined.