COMP 2211 Final Exam (Part A) - Spring 2022 - HKUST

Date: May 27, 2022 (Friday)
Time Allowed: 1 hour 15 minutes, 12:45–2:00 pm
Instructions:

1. This is a closed-book, closed-notes examination.
2. In this part, there are **5** questions (The last one is a dummy problem) on **13** pages.
3. Write your answers in the space provided.

Student Name	SOLUTIONS AND MARKING SCHEME
Student ID	
Email Address	

For T.A. Use Only

Problem	Topic	Score
1	True/False Questions	/ 10
2	Naïve Bayes and K-Nearest Neighbors	/ 14
3	Multilayer Perceptron (MLP)	/ 14
4	Digital Image Processing	/ 15
5	Dummy Question	/ 0
	Total	/ 53

Problem 1 [10 points] True/False Questions

Indicate whether the following statements are true or false by putting **T** or **F** in the given table. You get 1 point for each correct answer.

(a) One drawback of the K-means algorithm is that one needs to specify exactly how many clusters the algorithm should find.

(b) Increasing the number of hidden layers always increases the model performance.

(c) When handling a binary classification task, both softmax and sigmoid functions can be used as the activation function in the output layer.

(d) Validation accuracy must be lower than training accuracy.

(e) We cannot train/inference deep learning networks using CPU.

(f) Otsu's thresholding method and affine transformations are point-based image operations.

(g) In the convolutional layer of a CNN, the number of weights depends on the depth of the input volume and the number of biases is equal to the number of kernels.

(h) After training a neural network, you observe a large gap between the training accuracy (100%) and the task accuracy (40%). Dropout is commonly used to reduce this gap.

(i) In a minimax-based 3×3 tic-tac-toe game, an AI player will definitely win because it knows all possible moves of the game.

(j) The alpha-beta pruning algorithm is preferred to minimax because it computes the same answer as minimax while usually doing so without examining as much of the game tree.

Answer:

Question	Answer (T/F)
(a)	T
(b)	F
(c)	T
(d)	F
(e)	F
(f)	F
(g)	T
(h)	T
(i)	F
(j)	T

Marking scheme:

- 1 point for giving each correct answer. 10 points in total.

Problem 2 [14 points] Naïve Bayes and K-Nearest Neighbors

Given the training data in the table below.

No.	CGPA	Interest in Computing Subjects	Practice-oriented Learner?	COMP 2211 Grade	Select COMP as Major
1	≤ 3	High	No	B	No
2	≤ 3	High	No	A	No
3	> 3 AND ≤ 4	High	No	B	Yes
4	> 4	Medium	No	B	Yes
5	> 4	Low	Yes	B	Yes
6	> 4	Low	Yes	A	No
7	> 3 AND ≤ 4	Low	Yes	A	Yes
8	≤ 3	Medium	No	B	No
9	≤ 3	Low	Yes	B	Yes
10	> 4	Medium	Yes	B	Yes
11	≤ 3	Medium	Yes	A	Yes
12	> 3 AND ≤ 4	Medium	No	A	Yes
13	> 3 AND ≤ 4	High	Yes	B	Yes
14	> 4	Medium	No	A	No

(a) [6.5 points] Given a new example with the following attribute values. Predict the value of its "Select COMP as Major" using Naïve Bayes classifier. Show all the steps.

- CGPA ≤ 3
- Interest in Computing Subjects = Medium
- Practice-oriented Learner = Yes
- COMP 2211 Grade = B

Answer: Let

- E be CGPA ≤ 3, Interest in Computing Subjects = Medium, Practice-oriented Learner = Yes, COMP 2211 Grade = B
- E_1 be CGPA ≤ 3
- E_2 be interest in computing subjects = medium
- E_3 be practice-oriented learner = yes
- E_4 be COMP 2211 grade = B

$$P(Yes|E) = \frac{P(E_1|Yes)P(E_2|Yes)P(E_3|Yes)P(E_4|Yes)P(Yes)}{P(E)}$$

$$P(Yes) = 9/14 = 0.643$$
$$P(E_1|Yes) = 2/9 = 0.222$$
$$P(E2|Yes) = 4/9 = 0.444$$
$$P(E3|Yes) = 6/9 = 0.667$$
$$P(E4|Yes) = 6/9 = 0.667$$
$$P(Yes|E) = \frac{(0.222)(0.444)(0.667)(0.668)(0.443)}{P(E)} = \frac{0.028}{P(E)}$$
$$P(No|E) = \frac{P(E_1|No)P(E_2|No)P(E_3|No)P(E_4|No)P(No)}{P(E)}$$

$$P(No) = 5/14 = 0.356$$
$$P(E_1|No) = 3/5 = 0.6$$
$$P(E2|No) = 2/5 = 0.4$$
$$P(E3|No) = 1/5 = 0.2$$
$$P(E4|No) = 2/5 = 0.4$$
$$P(No|E) = \frac{(0.6)(0.4)(0.2)(0.4)(0.357)}{P(E)} = \frac{0.007}{P(E)}$$

Hence, the Naïve Bayes classifier predicts "Select COMP as Major" = Yes for the new example.

Marking scheme:

- 0.5 for giving each conditional and prior probability. 6 points in total.
- 0.5 for giving the correct prediction.

(b) [7.5 points] Similar to the above, but this time, predict the value of its "Select COMP as Major" using K-nearest neighbor for K = 5. Complete the following table and state the prediction result based on the data in the completed table. For similarity measure, use a simple match of attribute values:

$$S(a_i, b_i) = \sum_{i=1}^{4} w_i * \texttt{distance}(a_i, b_i)$$

where $\texttt{distance}(a_i, b_i)$ is 0 if a_i equals b_i, and 1 otherwise. a_i and b_i are either CGPA, interest in computing subjects, practice-oriented learner or COMP 2211 grade. Weights, w_i, are all 1 except for interest in computing subjects, it is 2.

No.	Class	Distance to New Example
1	No	$0 + 2 + 1 + 0 = 3$
2	No	$0 + 2 + 1 + 1 = 4$
3	Yes	$1 + 2 + 1 + 0 = 4$
4	Yes	$1 + 0 + 1 + 0 = 2$
5	Yes	$1 + 2 + 0 + 0 = 3$
6	No	$1 + 2 + 0 + 1 = 4$
7	Yes	$1 + 2 + 0 + 1 = 4$
8	No	$0 + 0 + 1 + 0 = 1$
9	Yes	$0 + 2 + 0 + 0 = 2$
10	Yes	$1 + 0 + 0 + 0 = 1$
11	Yes	$0 + 0 + 0 + 1 = 1$
12	Yes	$1 + 0 + 1 + 1 = 3$
13	Yes	$1 + 2 + 0 + 0 = 3$
14	No	$1 + 0 + 1 + 1 = 3$

Among the 5 nearest neighbors, 4 are from class Yes, and 1 from class No. Hence, the KNN classifier predicts "Select COMP as Major" = Yes for the new example.

Marking scheme:

- 0.5 point for giving each correct answer. 7 points in total.
- 0.5 point for giving the correct prediction.

Problem 3 [14 points] Multilayer Perceptron (MLP)

This problem is about multilayer perceptron (MLP). Answer all the questions below.

(a) [3 points] State when using the F1 metric is better than using accuracy as an evaluation metric. Also use a confusion matrix to illustrate the stated situation.

Answer:

- When the dataset is unbalanced.
- When false-negative and false-positive matter a lot.

Actual/Predicted	Infectious disease=yes	Infectious disease=no
Infectious disease=yes	1	20
Infectious disease=no	5	100

Accuracy = 0.8016
F1 score = 0.0741

Marking scheme:

- 1 point for stating the situation
- 2 points for the confusion matrix

(b) [2 points] Suppose we are dealing with binary classification tasks using MLP. Explain why it is inappropriate to use ReLU as the activation function in the output layer.

Answer:
We cannot determine the cut-off threshold to distinguish between the output classes when there is an unbounded output range.

Marking scheme:

- 2 points for explaining by the "unbounded" range of ReLU so cannot determine the cut-off
- 1 point if mentioning the range of function (*not describing what ReLU do, but stating the range) but no further explanation
- 0 point if only stating ReLU can output value more than 0 & 1 without mentioning it has an unbounded range (bounded function beyond the range from 0 to 1 can work, just do the mapping)

(c) [2 points] Design the output layer of an MLP for handling a multilabel classification problem with n classes by stating the number of neurons and the activation function.
Remark: Multilabel classification is a supervised learning problem where an instance may be associated with multiple labels.

Answer:
n neurons and sigmoid function.

Marking scheme:

- 1 point for n neurons
- 1 point for sigmoid

(d) [3 points] Explain why it is not good to initialize all weights of an MLP to zero.
Hint: Refer to the following updating rules of weights and biases for MLP.

- $\delta_k = (O_k - T_k)O_k(1 - O_k)$
- $\delta_j = O_j(1 - O_j)\sum_{k \in K} \delta_k w_{jk}$
- $w_{jk} \leftarrow w_{jk} - \eta \delta_k O_j$
- $w_{ij} \leftarrow w_{ij} - \eta \delta_j O_i$
- $\theta_j \leftarrow \theta_j - \eta \delta_j$
- $\theta_k \leftarrow \theta_k - \eta \delta_k$

Answer:
If a network is initialized with all zeros, all the neurons will propagate on the same gradient, making different neurons learn the same features. Thus, this leads to poor performance.

Marking scheme:

- 3 points if "the same update/propagate/feature learned/symmetry problem"
- 2 points if "zero updates" as it isn't always true for all MLP design
- 2 points if stating only gradient computation are the same for neurons
- 1 point if only mentioning deltaj will become zero / stating gradient "may" not update
- 0 point for low efficiency / poor performance

(e) [2 points] Explain what will happen if the learning rate η of an MLP is

(i) Too large

(ii) Too small

Answer:

(i)
- Cause the model to coverage too quickly to a sub-optimal solution.
- Unstable training like oscillations.

Marking scheme:
- 1 point for explaining what will happen if the learning rate is too large.
- not accept learn faster/overfit/low accuracy

(ii) Learning will be slow.

Marking scheme:
- 1 point for explaining what will happen if the learning rate is too small.
- not accept underfit

(f) [2 points] Describe a way to avoid overfitting in MLP. Explain why it works.

Answer:
Adding regulations helps to keep the weights small, such that it is less likely for the model to have a large variance (i.e. be sensitive to noise and fluctuations in data).

Marking scheme:
- 1 point for method
- 1 point for explanation
- only valid explanation (but not stating the definition or recalling some rule of thumbs) get the explanation point

Problem 4 [15 points] Digital Image Processing

(a) Assume we apply the following kernel to an 8-bit grayscale image.

$$K = \begin{bmatrix} -1 & -2 & 0 \\ -2 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix}$$

(i) [2 points] Determine the maximum and minimum possible values that a pixel to which the given kernel is applied can have. Do not perform any normalization.

Answer:

Since an 8-bit grayscale image is assumed, the highest value that each pixel can have is 255, and the lowest value is 0. After applying the kernel, the maximum value is achieved when the negative values of the kernel multiply 0s and the positive values of the kernel multiply 255s. Then, the maximum achievable value is $v_{max} = (2 + 2 + 1)(255) = 1275$. Following the same reasoning, the minimum value will be $v_{min} = (-2 - 2 - 1)(255) = -1275$.

Marking scheme:

- 1 point for stating the maximum possible value.
- 1 point for stating the minimum possible value.

(ii) [2 points] Suggest a grey-level transformation function to ensure that any output of this kernel will be within the legal range of a standard 8-bit grayscale image.

Answer:

A 8-bit image must have a range from 0 to 255. Since the maximum and minimum possible values for the gray levels are $v_{max} = 1275$ and $v_{min} = -1275$, respectively, the function will be

$$\begin{aligned} I_{output} &= 255 \left(\frac{I_{input} - v_{min}}{v_{max} - v_{min}} \right) \\ &= 255 \left(\frac{I_{input} - (-1275)}{1275 - (-1275)} \right) \\ &= 255 \left(\frac{I_{input} + 1275}{2550} \right) \end{aligned}$$

where I_{input} and I_{output} are the input and output images, respectively.

Marking scheme:

- 2 points for stating a transformation function.
- -1 point if the function is merely returning legal range.

(b) Consider the following 2×2 image:

0	9
18	27

Apply the following image operations sequentially.

(i) [3 points] Show the resulting image of size 4×4 after adding reflection padding.

Answer:

0	0	9	9
0	0	9	9
18	18	27	27
18	18	27	27

Marking scheme:

- 0.25 point for each correct value. 3 points in total.

(ii) [4 points] Apply a 3×3 averaging kernel to the resulting image of part (b)(i). Assume the output image after image averaging is in the same shape as the input image by doing zero padding. Round the number to integer if needed.

Answer:

0	2	4	4
4	9	12	10
8	15	18	14
8	14	16	12

Marking scheme:

- 0.25 point for each correct value. 4 points in total.

(iii) [4 points] Compute the optimal threshold after applying ONE ITERATION of Otsu's method on the resulting image of part (b)(ii). Assume the initial threshold is set to the mean pixel intensity of the resulting image.

Answer:

- Initial threshold $= (0+2+4+4+4+9+12+10+8+15+18+14+8+14+16+12)/16 = 9.375$
- $\mu_1 = (0+2+4+4+4+9+8+8)/8 = 4.875$
- $\mu_2 = (12+10+15+18+14+14+16+12)/8 = 13.875$
- Optimal threshold $= (4.875 + 13.875) = 9.375$

Marking scheme:

- 1 point for the answer of initial threshold.
- 1 point for the answer of μ_1.
- 1 point for the answer of μ_2.
- 1 point for the optimal threshold after 1 iteration.
- -1 point for incorrect input from part b(ii). I.e. incorrect calculation based on result of part b(ii.)

Problem 5 [0 points] Dummy Question

Are you sure you have finished all the questions?

---------- END OF PAPER ----------