

Duration: 120 minutes
Aids Allowed: None

Student Number: _____

Family Name(s): _____

Given Name(s): _____

Lecture Section: Afternoon Section
 Evening Section

*Do not turn this page until you have received the signal to start.
In the meantime, please read the instructions below carefully.*

This test consists of 6 questions on 22 pages (including this one), printed on both sides of the paper. *When you receive the signal to start, please make sure that your copy of the test is complete, fill in the identification section above, and write your name on the back of the last page.*

Answer each question directly on the test paper, in the space provided, and use the reverse side of the pages for rough work. If you need more space for one of your solutions, use the reverse side of a page and *indicate clearly the part of your work that should be marked.*

Write up your solutions carefully! If you are giving only one part of an answer, indicate clearly what you are doing. Part marks might be given for incomplete solutions where it is clearly indicated what parts are missing.

You **must** write the test in pen if you would like to potentially request for the test to be regraded.

MARKING GUIDE

1: _____/15

2: _____/20

3: _____/10

4: _____/15

5: _____/10

6: _____/20

TOTAL: _____/90

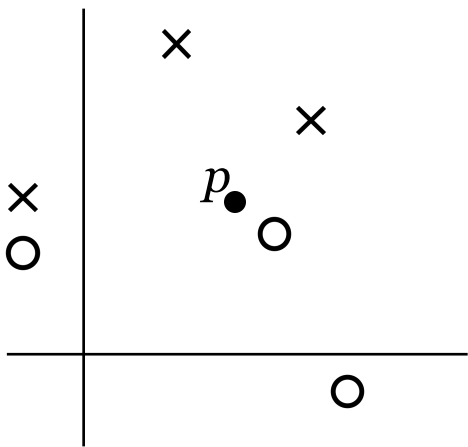
Use this page for rough work—clearly indicate any section(s) to be marked.

Question 1. [15 MARKS]

Part (a) [3 MARKS]

We would like to use 1-Nearest Neighbour to classify the point p as either X or O using the training data shown below. What is the prediction if cosine distance (i.e, negative cosine similarity) is used as the distance measure? What is the prediction if Euclidean distance is used?

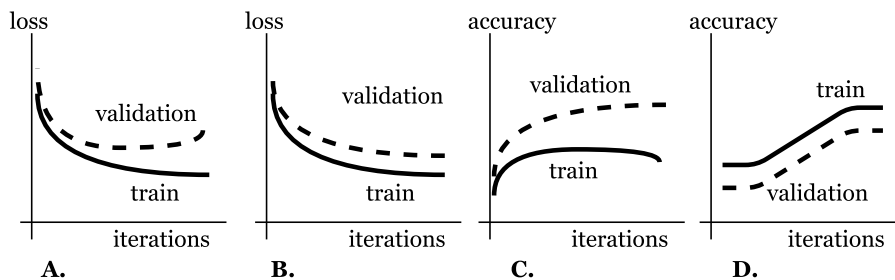
- (A) Cosine distance: O, Euclidean distance: X
- (B) Cosine distance: O, Euclidean distance: O
- (C) Cosine distance: X, Euclidean distance: X
- (D) Cosine distance: X, Euclidean distance: O



Part (b) [3 MARKS]

Which of the following learning curves demonstrates overfitting? **Circle one of choices 1-5.**

- 1. A
- 2. B
- 3. C
- 4. D
- 5. A and C



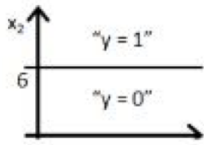
Use this page for rough work—clearly indicate any section(s) to be marked.

Part (c) [3 MARKS]

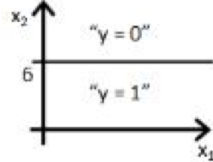
Suppose you train a logistic regression classifier and the learned hypothesis function is

$$h_{\theta}(x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2),$$

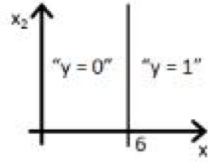
where $\theta_0 = 6, \theta_1 = 0, \theta_2 = -1$. Which of the following represents the decision boundary for $h_{\theta}(x)$?



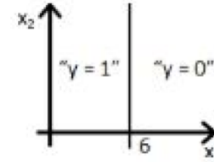
A



B



C



D

Part (d) [3 MARKS]

Alice and Bob, two students from the CSC411/2515 night section, walk home after lecture, and see what appears to be an alien spaceship floating in the sky. Alice concludes that aliens are real. How can we use Bayes' rule to explain why Bob might not reach the same conclusion?

- (A) Bob's reasoning is more like gradient descent with momentum, which can give different results than gradient descent without momentum.
- (B) Alice is using MAP inference, whereas Bob is using Maximum Likelihood inference.
- (C) Bob's prior beliefs about aliens' existence are different from Alice's.
- (D) None of the above.

Part (e) [3 MARKS]

Which of the following is (are) true about optimizers?

- (A) We can speed up training by using an optimizer that uses a different learning rate for each weight.
- (B) Dropout should not be used alongside momentum.
- (C) Reducing the batch size when using Stochastic Gradient Descent always improves training.
- (D) It does not make sense to use Stochastic Gradient Descent to train a linear regression model because linear regression is convex.
- (E) All of the above.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 2. [20 MARKS]

Your training set is $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$. Assume that your model for the data is

$$y^{(i)} \sim \text{Laplace}(\theta^T x^{(i)}, 1).$$

The probability density function of the Laplace distribution with mean μ and scale parameter b is

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(\frac{-|x - \mu|}{b}\right).$$

Part (a) [5 MARKS]

Write down the formula for the likelihood of the training set D .

Part (b) [15 MARKS]

Suppose you would like to learn the model with penalized maximum likelihood, with an L1 regularization/penalty term. Derive the Gradient Descent update for this setting. Briefly justify every step. **Draw a rectangle around the formula that you derived.**

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 3. [10 MARKS]

Suppose we are interested in a neuron in the 4-th layer of a convolutional neural network. What are two different methods of visualizing what the job of that neuron is? Do not just give names of algorithms or methods; give enough detail for the reader to be able to implement the methods you are describing.

Method 1: _____
Details:

Method 2: _____
Details:

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 4. [15 MARKS]

Bob would like to classify emails as spam or non-spam. He would like to estimate the probability that a new email e containing the keywords (w_1, w_2, \dots, w_n) is spam by taking all the emails in the training set with those keywords, and then computing the proportion of those emails that are spam. Specifically, he estimates the probability using

$$P(\text{spam}|\text{new email } e) = \frac{\text{num. of spam emails with keywords } w_1, w_2, \dots, w_n \text{ in the training set}}{\text{num. of total emails with keywords } w_1, w_2, \dots, w_n \text{ in the training set}}.$$

Part (a) [3 MARKS]

Explain why Bob's plan will generally **not** work.

Part (b) [5 MARKS]

Describe the datasets for which Bob's plan might work. Be specific: state which properties are required of the datasets.

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (c) [5 MARKS]

Describe how to estimate the probability that an email is spam using the Naive Bayes assumption. Your answer should include formulas.

Part (d) [2 MARKS]

Why does the Naive Bayes assumption address a problem with Bob's plan?

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 5. [10 MARKS]

Explain why a one-hidden-layer neural network is less likely to overfit if it has fewer units in the hidden layer. You should assume the reader is familiar with neural networks, but you should not assume that the reader is familiar with other concepts from CSC411/2515.

Use this page for rough work—clearly indicate any section(s) to be marked.

Question 6. [20 MARKS]

The Poisson distribution is used to model data that consists of non-negative integers. Its probability mass function is $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$. For the dataset $\{k_1, k_2, k_3, \dots, k_n\}$, the maximum likelihood estimate for λ is $\frac{1}{n} \sum_{j=1}^n k_j$.

Suppose you observe m integers in your training set. Your model assumption is that each integer is sampled from one of two different Poisson distributions. You would like to learn this model using the EM algorithm.

Part (a) [2 MARKS]

List all the parameters of the model.

Part (b) [4 MARKS]

What is the likelihood of the training set under the model?

Part (c) [7 MARKS]

Derive the E-step for this model. Your answer should include a mathematical justification. **Draw a rectangle around the formula that you derived.**

Use this page for rough work—clearly indicate any section(s) to be marked.

Part (d) [7 MARKS]

Derive the M-step for this model. Your answer should include a mathematical justification. **Draw a rectangle around the formula that you derived.**

Use this page for rough work—clearly indicate any section(s) to be marked.

Additional page for answers

On this page, please write nothing except your name.

Family Name(s): _____

Given Name(s): _____