UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2022/23

ADVANCED MACHINE LEARNING

Duration 120 mins (2 hours)

This paper is a WRITE-ON examination paper.

You ***must*** write your Student ID on this Page and must not write your name anywhere on the paper.

All answers should be written within the designated boxes in this examination paper and sufficient space is provided for each question.

If, for some reason, space is required to complete or correct an answer to a question, use the “Additional Space” provided on the facing or adjacent page to the question. Clearly indicate which question the answer corresponds to.

No credit will be given for answers presented elsewhere and without clear indication of to what question they correspond. Blue answer books may be used for scratch; they will be discarded without being looked at.

Answer all parts of the question in section A (40 marks)
and ALL three questions from section B (20 marks each)

Student ID:

Question	Mark	*Arithmetic checked*	*Double Marked*
A1	/40		
B2	/20		
B3	/20		
B4	/20		
Total:	/100		

University approved calculators MAY be used.

A foreign language translation dictionary (paper version) is permitted provided it contains no notes, additions or annotations.

16 page examination paper

 Version 1

Section A

A 1

(a) Explain why in boosting it is important to use as decorrelated trees as possible and explain how this is achieved in random forest. [5 marks]

$\overline{5}$

(b) Multilayer perceptrons (MLPs) are famously universal approximators (they can approximate any smooth function up to arbitrary precision). This should make them susceptible to massively overfitting a training set. Explain why, despite this, MLPs often have good generalisation performance. [5 marks]

$\overline{5}$

(c) Show that if $\lambda > 0$ is an eigenvalue of $\mathbf{C} = \mathbf{X}\,\mathbf{X}^{\mathsf{T}}$ then it is also an eigenvalue of $\mathbf{D} = \mathbf{X}^{\mathsf{T}}\mathbf{X}$, where $\mathbf{X}$ is a matrix. [5 marks]

$\overline{5}$

(d) If $\|\boldsymbol{x}\|$ is a proper norm, use the triangular inequality ($\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$), linearity of a norm ($\|a\,\boldsymbol{x}\| = a\,\|\boldsymbol{x}\|$) and the definition of convexity, to show that the norm is convex. [5 marks]

$\overline{5}$

(e) Use the fact norms are convex to argue that an elastic net with a loss function

$$L(\boldsymbol{w}) = \sum_{i=1}^{m} \left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i - y_i\right)^2 + \alpha\|\boldsymbol{w}\|_{L_1} + \beta\,\|\boldsymbol{w}\|_{L_2}^2$$

has a unique minimum. [5 marks]

$\overline{5}$

(f) Explain the main advantage and disadvantage of stochastic gradient descent (SGD) compared to full gradient descent. [5 marks]

$\overline{5}$

(g) The multinomial distribution $\mathrm{Multi}(\boldsymbol{k}|n, \boldsymbol{p})$ describes the likelihood of observed counts $\boldsymbol{k} = (k_1, k_2, \ldots, k_d)$ for d possible outcomes, where the total number of counts is $\sum_{i=1}^{d} k_i = n$. The vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_d)$ is a vector of probabilities (summing to 1) where p_i is the probability of outcome i occuring. Show that the Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha})$ is a conjugate prior to the multinomial likelihood $\mathrm{Multi}(\boldsymbol{k}|n, \boldsymbol{p})$ where

$$\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \Gamma(\alpha_0) \prod_{i=1}^{d} \frac{p_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} \qquad \mathrm{Multi}(\boldsymbol{k}|n, \boldsymbol{p}) = n! \prod_{i=1}^{d} \frac{p_i^{k_i}}{k_i!}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a vector of parameters that controls the Dirichlet distribution with $\alpha_0 = \sum_{i=1}^{n} \alpha_n$. Derive the update equation for the parameters $\boldsymbol{\alpha}$. [5 marks]

$\overline{5}$

(h) Describe how the minimum description length formalism is used for model selection.

[5 marks]

$\overline{5}$

End of question A1

(a) $\frac{}{5}$	(b) $\frac{}{5}$	(c) $\frac{}{5}$	(d) $\frac{}{5}$	(e) $\frac{}{5}$	(f) $\frac{}{5}$	(g) $\frac{}{5}$	(h) $\frac{}{5}$	Total $\frac{}{40}$

TURN OVER

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Section B

B 2

(a) Show that the expected generalisation given by

$$\mathbb{E}_{\mathcal{D}}\big[E(\mathcal{D})\big] = \mathbb{E}_{\mathcal{D}}\left[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right]$$

can be written as the sum of a bias term, B, and variance term V where

$$B = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2, \quad V = \mathbb{E}_{\mathcal{D}}\left[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)^2\right]$$

where $\hat{f}_m(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[\hat{f}(\boldsymbol{x}|\mathcal{D})\right]$ is the prediction made by averaging over all machines.

[10 marks]

$\overline{10}$

(b) Explain in words (1) the *bias* and (2) the *variance* terms and (3) the dilemma.
[6 marks]

1

2

3

$\overline{6}$

(c) Consider a classification problem where $\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})$ is the softmax output of a learning machine trained on a dataset $\mathcal{D}$ for class c. Let $\hat{m}_c(\boldsymbol{x}) = \mathbb{E}_\mathcal{D}\left[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\right]$ be the output of the mean machine for class c (i.e. the outputs averaged over machines trained on all possible datasets). Consider the cross-entropy loss

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \sum_{c\in\mathcal{C}} [\![y = c]\!] \log\left(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\right)$$

where $[\![y = c]\!]$ is an indicator function equal to 1 if the target y is equal to class c, and 0 otherwise. Show that the expected loss over all training sets can be written as the expected loss for the mean machine (a bias) plus an additional term (a variance). Use Jensen's inequality ($\mathbb{E}\left[\log(X)\right] \leq \log\left(\mathbb{E}\left[X\right]\right)$) to show the second term is non-negative. [4 marks]

$\overline{4}$

End of question B2

(a) $\overline{10}$ (b) $\overline{6}$ (c) $\overline{4}$ Total $\overline{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

B 3

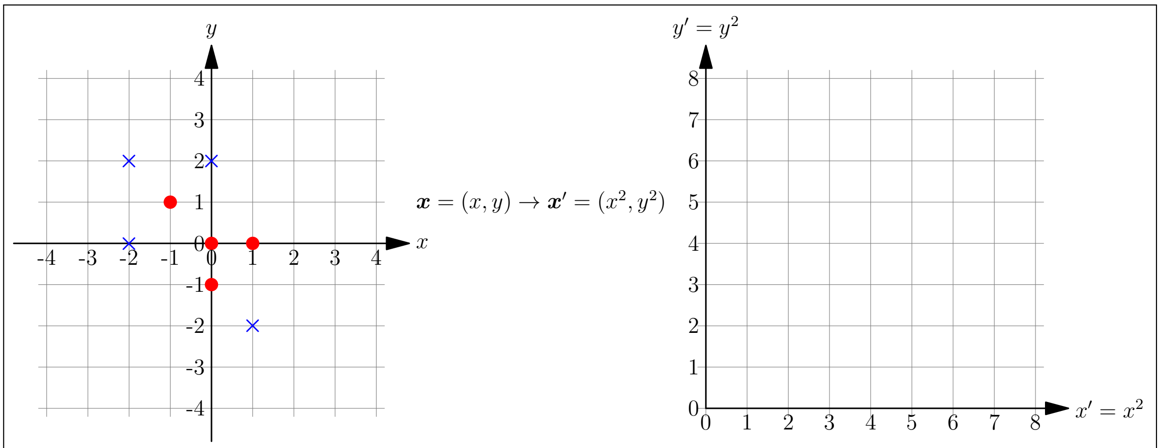
(a) Show that for the mapping

$$\boldsymbol{x} = (x_1, x_2, x_3)^{\mathsf{T}} \to \vec{\phi}(\boldsymbol{x}) = (x_1^2, x_2^2, x_3^2, \sqrt{2}\, x_1\, x_2,\ \sqrt{2}\, x_1\, x_3,\ \sqrt{2}\, x_2\, x_3)^{\mathsf{T}}$$

the kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \left\langle \vec{\phi}(\boldsymbol{x}), \vec{\phi}(\boldsymbol{y}) \right\rangle$ is equal to $(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y})^2$. [5 marks]

$\overline{5}$

(b) Show how the data points $\{\boldsymbol{x}_i = (x_i, y_i) | i = 1, 2, \ldots\}$ shown below transform under the mapping $\boldsymbol{x} = (x_i, y_i) \to \boldsymbol{x}' = (x_i^2, y_i^2)$ and sketch the position of the maximal margin dividing plane in the new feature space. [5 marks]



$\overline{5}$

(c) Give three conditions that any positive semi-definite kernel, $K(\boldsymbol{x}, \boldsymbol{y})$, should satisfy.

[6 marks]

1

2

3

$\overline{6}$

(d) Using properties of positive semi-definite kernels to show that

$$K^{(3)}(\boldsymbol{x}, \boldsymbol{y}) = K^{(2)}(\boldsymbol{x}, \boldsymbol{y})\, K^{(1)}(\boldsymbol{x}, \boldsymbol{y})$$

is positive semi-definite if $K^{(1)}(\boldsymbol{x}, \boldsymbol{y})$ and $K^{(2)}(\boldsymbol{x}, \boldsymbol{y})$ are positive semi-definite.

[4 marks]

$\overline{4}$

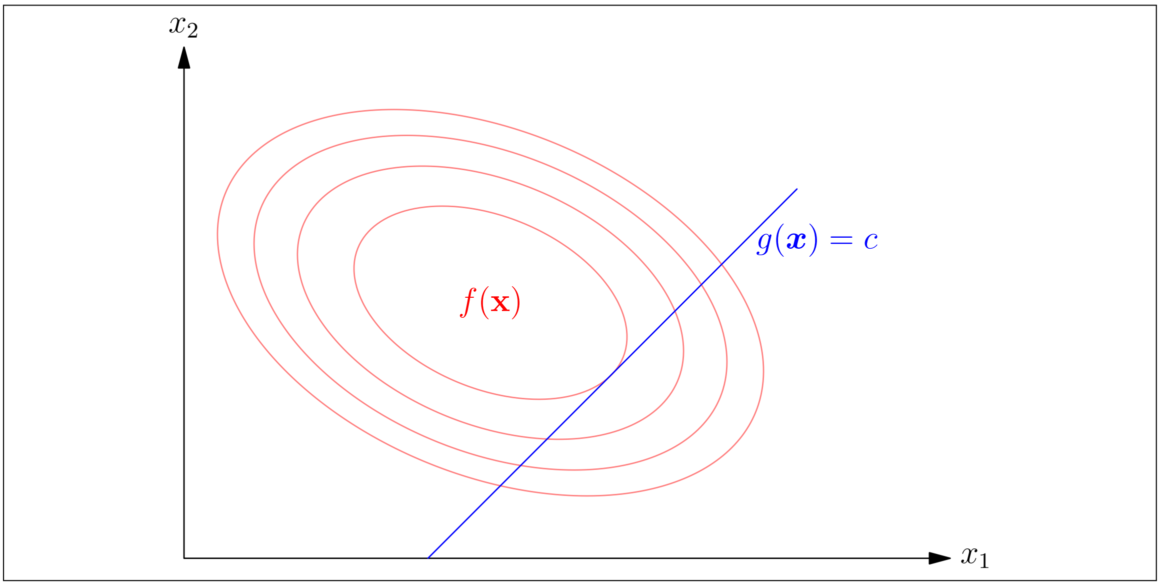
End of question B3

(a) $\overline{5}$ (b) $\overline{5}$ (c) $\overline{6}$ (d) $\overline{4}$ Total $\overline{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

B 4

(a) Below we show contour lines for a quadratic minimum $f(\boldsymbol{x})$ and a constraint $g(\boldsymbol{x}) = x_2 - x_1 = c$. Plot the gradient $\boldsymbol{\nabla} f(\boldsymbol{x})$ at various points along the contour lines and $\boldsymbol{\nabla} g(\boldsymbol{x})$ at various points along the constraint. Mark the point that minimises $f(\boldsymbol{x})$, subject to the constraint $g(\boldsymbol{x}) = c$. Write down the condition for the minimum points. [10 marks]



$\overline{}$
10

(b) Consider for a dataset $\{\boldsymbol{x}_i | i = 1, 2, \ldots, n\}$. Subtracting the mean and projecting onto a vector $\boldsymbol{v}$ gives us a number $z_i = \boldsymbol{v}^\mathsf{T}(\boldsymbol{x}_i - \boldsymbol{\mu})$. Show that the direction $\boldsymbol{v}$, with $\|\boldsymbol{v}\|^2 = 1$, that maximises the variance of the set of numbers $\{z_i | i = 1, 2, \ldots, n\}$, is given by the eigenvector of the covariance matrix with the largest eigenvalue. [10 marks]

$\overline{}$
10

End of question B4

(a) $\frac{}{10}$ (b) $\frac{}{10}$ Total $\frac{}{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

