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.

15 page examination paper

Version 1

Section A

A 1

(a) A popular method for fitting a dataset $\big\{(\boldsymbol{x}_k, y_k) | k = 1, 2, \ldots, m\big\}$, where $\boldsymbol{x}_k \in \mathbb{R}^p$ are p-dimensional feature vectors and $y_k \in \mathbb{R}$ are targets, is to minimise a loss function, $\mathcal{L}(\boldsymbol{w})$, with an L_1 regulariser (the so called Lasso method)

$$\mathcal{L}(\boldsymbol{w}) = \sum_{k=1}^{m} \left(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x}_k - y_k\right)^2 + \nu \sum_{i=1}^{p} |w_i|,$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)^{\mathsf{T}}$. Explain why this regulariser is used and either graphically or otherwise explain how it works. [5 marks]

$\overline{5}$

(b) Explain why a convolutional operator acting on an image is approximately translationally equivariant and explain why it is not fully equivariant.

[5 marks]

$\overline{5}$

(c) Explain the kind of problems where the following learning machines excel

(i) Random Forest

(ii) Support Vector Machines

(iii) Convolutional Neural Networks

(iv) Hierarchical Bayesian Models

[5 marks]

1

2

3

4

$\overline{5}$

(d) Explain how Gradient Boosting works. [5 marks]

$\overline{5}$

(e) Consider the loss function

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\boldsymbol{w} + \nu\|\boldsymbol{w}\|$$

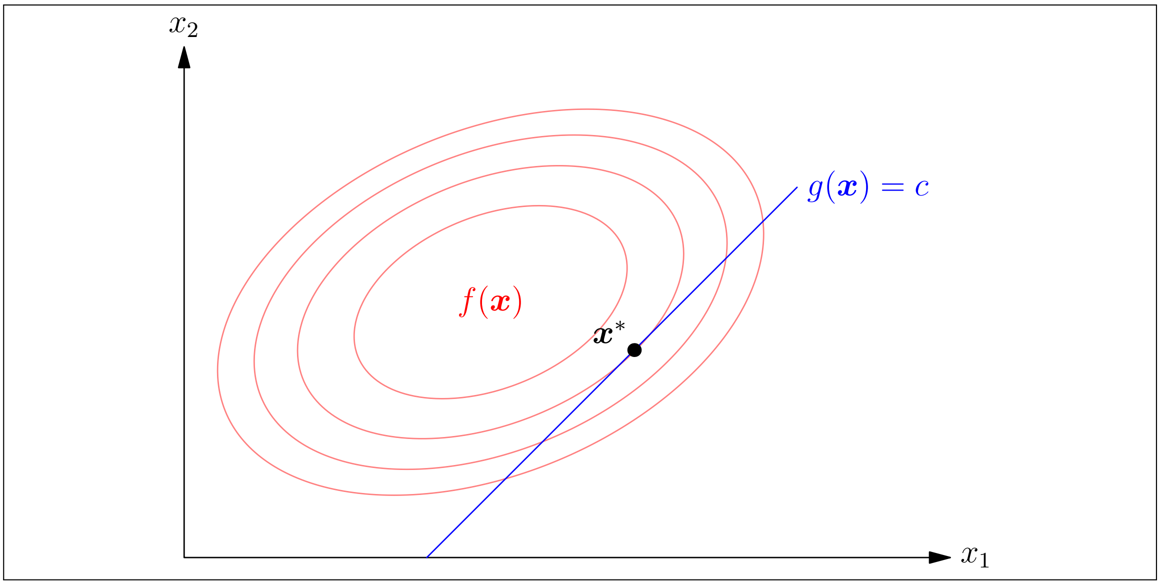
where $\mathbf{X}$ is the design matrix, $\boldsymbol{w}$ is a set of weights and $\nu > 0$ is a scalar. Explain why $\mathcal{L}(\boldsymbol{w})$ is convex? [5 marks]

$\overline{5}$

(f) Explain how to do model selection in Bayesian inference and explain why this is especial useful when using Gaussian Processes. [5 marks]

$\overline{5}$

(g) Below is shown contour lines for a function $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\mathbf{Q}\boldsymbol{x}$ that we wish to minimise subject to a linear constraint $g(\boldsymbol{x}) = c$. At the point $\boldsymbol{x}^*$ indicated, sketch the gradient $\boldsymbol{\nabla} f(\boldsymbol{x}^*)$ and $\boldsymbol{\nabla} g(\boldsymbol{x}^*)$. Use the diagram to explain why $\boldsymbol{\nabla}\mathcal{L} = 0$ finds the solution to the constrained optimisation problem where $\mathcal{L} = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$. [5 marks]



$\overline{}$
5

(h) Describe in words what the Wasserstein distance measures. [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) We can write the loss function for ridge regression as

$$\mathcal{L}(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \nu\|\boldsymbol{w}\|_2^2,$$

where $\mathbf{X}$ is the design matrix and $\boldsymbol{y}$ is a vector of target values.

(i) Calculate the weights, $\boldsymbol{w}^*$ that minimises the loss function.

(ii) Using the singular value decomposition $\mathbf{X} = \mathbf{USV}^\mathsf{T}$ write the optimal weight vector $\boldsymbol{w}^*$ in terms of $\mathbf{V}$, $\mathbf{U}$ and $\mathbf{S}$.

(iii) Hence show that $\boldsymbol{w}^* = \mathbf{V}\hat{\mathbf{S}}^+ \mathbf{U}^\mathsf{T}\boldsymbol{y}$, where $\hat{\mathbf{S}}^+$ is a diagonal matrix with non-zero elements $\hat{S}^+_{ii} = s_i/(s_i^2 + \nu)$, and s_i are the singular values of $\mathbf{X}$ (i.e. $s_i = S_{ii}$).

(5 marks for each sub-part) [15 marks]

$\overline{}$
15

(b) Use the result that you derive to explain how adding the L_2 regulariser $\nu \|w\|_2^2$ improves the conditioning of the solution and is likely to improve generalisation. [5 marks]

$\overline{}$
5

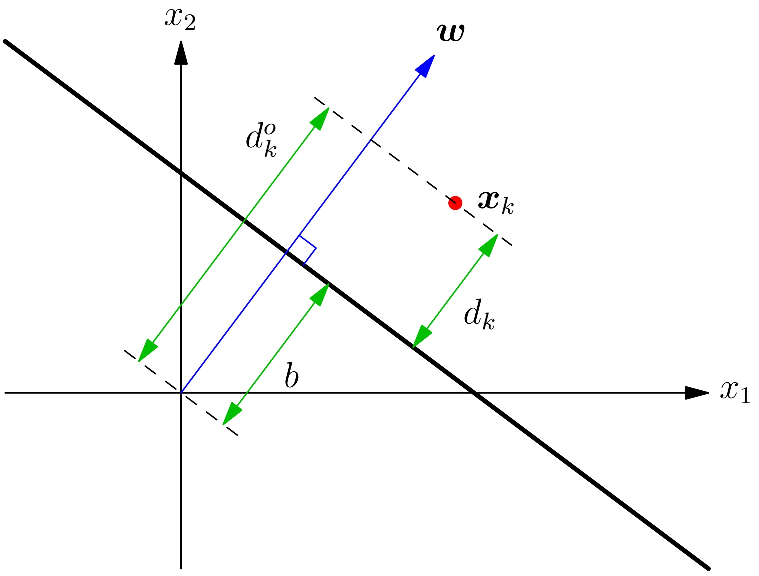
End of question B2

(a) $\frac{}{15}$ (b) $\frac{}{5}$ Total $\frac{}{20}$

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

B 3

(a) Write down a formula for the minimum distance, d_k^0, between $\boldsymbol{x}_k$ and a hyperplane through the origin perpendicular to $\boldsymbol{w}$, and the minimum distance d_k from $\boldsymbol{x}_k$ to the hyperplane perpendicular to $\boldsymbol{w}$ displaced by b. [5 marks]



$d_k^o =$ ______________________________

$d_k =$ ______________________________

$\overline{5}$

(b) Depending on the category $y_k \in \{-1, 1\}$, write down the condition for a data point to be at least a distance $\gamma > 0$ above (or below if $y_k = -1$) the hyperplane shown in part (a). [5 marks]

$\overline{5}$

(c) Define $\boldsymbol{w'} = \boldsymbol{w}/(\gamma \, \|\boldsymbol{w}\|)$ and $b' = b/\gamma$ to rewrite the condition from part (b) and explain why minimising $\|\boldsymbol{w}'\|^2$ is equivalent to maximising the margin γ.

[5 marks]

$\overline{}$
5

(d) Write down a Lagrangian for finding the maximal margin hyperplane for an SVM given data $(\boldsymbol{x}_k, y_k)$ for $k = 1,\, 2,\, \ldots,\, m$. [5 marks]

$\overline{}$
5

End of question B3

(a) $\frac{}{5}$ (b) $\frac{}{5}$ (c) $\frac{}{5}$ (d) $\frac{}{5}$ Total $\frac{}{20}$

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

B 4

(a) In a variational auto-encoder an input image, $\boldsymbol{x}$, is drawn from a dataset $\mathcal{D}$. The encoder generates a probability distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ defined in a latent space. A vector in the latent space, $\boldsymbol{z}$, is sampled from $q_{\phi}(\boldsymbol{z}|\boldsymbol{x})$ and sent to the decoder that then generates a probability distribution in the space of images $p_{\boldsymbol{\theta}}(\boldsymbol{x}'|\boldsymbol{z})$. The loss function is defined as

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}\big[\mathrm{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \big\| \mathcal{N}(\mathbf{0}, \mathbf{I})\big) - \log(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}))\big]$$

Provide an interpretation of the two terms in the loss function. [10 marks]

$\frac{\quad}{10}$

(b) By considering the second derivative show that $f(x) = -\log(x)$ is convex-up. [5 marks]

$\frac{\quad}{5}$

(c) Jensen's inequality for convex-up function states that

$$\mathbb{E}\left[f(X)\right] \geq f(\mathbb{E}\left[X\right]).$$

Use this to show that for any two categorical distributions $\mathbb{P}(F = i) = f_i$ and $\mathbb{P}(G = i) = g_i$, the KL-divergence

$$\mathrm{KL}(F\|G) = -\sum_i f_i \log\left(\frac{g_i}{f_i}\right)$$

is non-negative. [5 marks]

$\overline{5}$

End of question B4

(a) $\overline{10}$ (b) $\overline{5}$ (c) $\overline{5}$ Total $\overline{20}$

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