SEMESTER 2 EXAMINATION 2021/22

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:			

University approved calculators MAY be used.

17 page examination paper

Section A

A 1

(a) In the bias variance dilemma explain (1) what does the variance measure and (2) why adding a regularisation term might reduce the variance. [5 marks]

1

2

5

(b) Explain how *gradient boosting* works? [5 marks]

5

(c) (1) Give a definition of a positive definite kernel and (2) explain why kernels for Gaussian Processes must be positive definite. [5 marks]

1

2

$\overline{5}$

(d) 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]

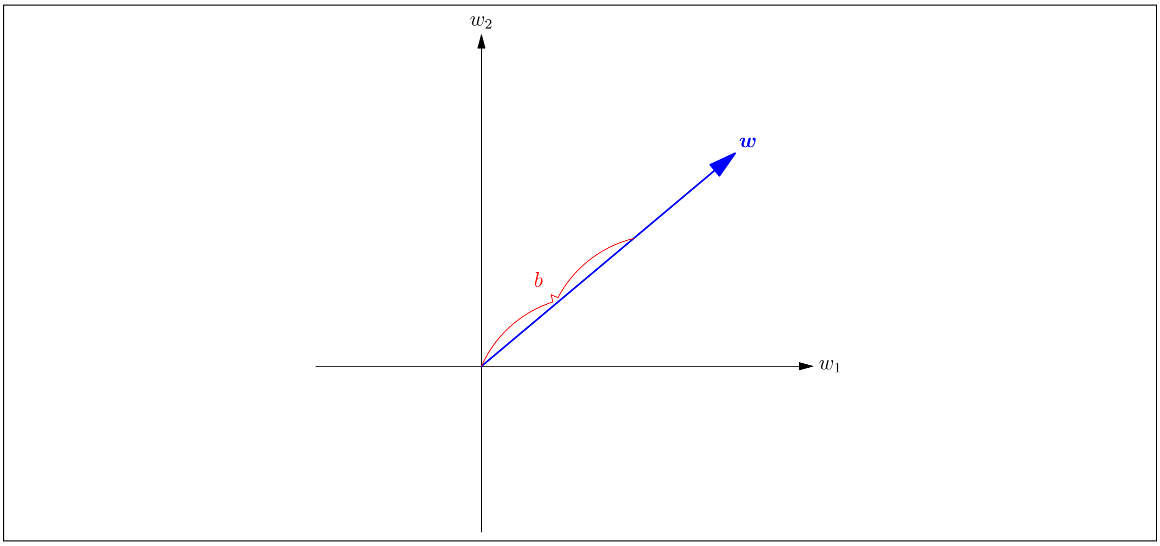
$\overline{5}$

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

(e) Sketch the set of points $\{\boldsymbol{x} \,|\, \boldsymbol{x}^\mathsf{T}\boldsymbol{w} = b\,\|\boldsymbol{w}\|\}$, which generate a separating plane and explain why

$$y_k\left(\frac{\boldsymbol{x}_k^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|} - b\right) \geq \gamma$$

implies that all points $(\boldsymbol{x}_k, y_k)$ with $y_k \in \{-1, 1\}$ will lie a distance of γ or more from the separating plane. [5 marks]



$\overline{5}$

(f) Explain how to do model selection in Bayesian inference. [5 marks]

$\overline{5}$

(g) 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]

$\overline{5}$

(h) Show that positive semi-definite kernels form a convex set. You may assume elementary properties of positive semi-definite kernels. [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 consider a regression problem where the data $(\boldsymbol{x}, y)$ is distributed according to $\gamma(\boldsymbol{x}, y)$. We consider a learning machine that makes a prediction $\hat{f}(\boldsymbol{x}|\boldsymbol{\theta})$, where the parameters, $\boldsymbol{\theta}$ are trained using a stochastic algorithm that returns parameters distributed according to a probability density $\rho(\boldsymbol{\theta})$. We can define the mean machine as $\hat{m}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\hat{f}(\boldsymbol{x}|\boldsymbol{\theta})\right]$. We assume that

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\gamma}\left[(\hat{m}(\boldsymbol{x}) - y)^2\right] = B, \qquad \mathbb{E}_{(\boldsymbol{x},y)\sim\gamma}\left[\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\left[\left(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}) - \hat{m}(\boldsymbol{x})\right)^2\right]\right] = V.$$

That is, we can define a bias B and variance V. We now consider ensembling n machines

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ are drawn independently from $\rho(\boldsymbol{\theta})$. Compute the expected generalisation error of $\hat{f}_n(\boldsymbol{x})$.

[10 marks]

$\overline{10}$

(b) Consider a classification problem where $\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})$ is the probability that a learning machine with parameters $\boldsymbol{\theta}$ predicts that input $\boldsymbol{x}$ belongs to class $c \in \mathcal{C}$. Assume the training is stochastic so the probability of obtaining parameters $\boldsymbol{\theta}$ is $\rho(\theta)$. Let $\hat{m}_c(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\right]$ be the output of the mean machine for class c. Assuming that for a data point $(\boldsymbol{x}, y)$, where y is a class label, we use a cross entropy loss

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

show that the expected loss over inputs and parameters can be written as the expected loss of the mean machine plus a second loss. Use Jensen's inequality ($\mathbb{E}\left[\log(X)\right] \leq \log\big(\mathbb{E}\left[X\right]\big)$) to show the second term is positive. [10 marks]

$\overline{10}$

End of question B2

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

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

B 3 The Lagrangian for an SVM with constraints $y_k\left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_k - b\right) \geq 1$ is given by

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{k=1}^{m} \alpha_k \left(y_k\left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_k - b\right) - 1\right).$$

(a) Sketch how slack variables, ξ_k, are introduced to allow some data points to lie within the margins. [5 marks]

$\overline{5}$

(b) Write down how adding slack variables modifies the constraint on the $\boldsymbol{w}$ and b for an SVM. Also write down the penalty term on the slack variables and the constraint on ξ_k. [5 marks]

$\overline{5}$

(c) Write down the modified Lagrangian and by computing the saddle-point equation with respect to the slack variables obtain an additional constraint on each Lagrange multiplier, α_k. [10 marks]

$\overline{}$
10

End of question B3

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

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

B 4 We have some counts, $\mathcal{D} = (k_1, k_2, \ldots, k_m)$, of a defective protein taken from patients with lung cancer—these measurements are assumed to be independent. We hypothesise that there are in fact two variants of the disease $d \in \{0, 1\}$. We model the counts, k_i, by a Poisson distribution $\mathbb{P}(k_i \mid \mu_d) = \mu_d^{k_i}\, \mathrm{e}^{-\mu_d}/k_i!$ where μ_d takes different values for the two variants of the disease. As we do not know which variant the different patients have we introduce a latent variable $z_i \in \{0, 1\}$ to signify if count i was from patient with disease $d = z_i$.

(a) Write down the likelihood $\mathbb{P}(k_i \,|\, \mu_0, \mu_1, z_i)$. [3 marks]

$\overline{3}$

(b) Let p_d be the prior probability of a patient with lung cancer to have variant $d \in \{0, 1\}$ (we are assuming $p_0 + p_1 = 1$). Write down the joint probability $\mathbb{P}(k_i, z_i \,|\, \mu_0, \mu_1)$. [2 marks]

$\overline{2}$

(c) Use Bayes rule to compute $\mathbb{P}(z_i = d \,|\, \mu_0, \mu_1, p_d, k_i)$. [5 marks]

$\overline{5}$

(d) We wish to estimate the set of parameters $\boldsymbol{\theta} = (\mu_0, \mu_1, p_1)$ by maximising the likelihood. Direct maximisation is difficult because the latent parameters will depend on the parameters $\boldsymbol{\theta}$, but their maximum likelihood values will depend on the latent variables. To solve this we use the EM algorithm where we iteratively maximise

$$Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{m} \sum_{z_i \in \{0,1\}} \mathbb{P}\Big(z_i|\boldsymbol{\theta}^{(t)}\Big) \log(\mathbb{P}(k_i, z_i|\mu_0, \mu_1))\,.$$

Let $P_i^{(t)} = \mathbb{P}(z_i|\boldsymbol{\theta}^{(t)})$ write down $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ explicitly and compute the new set parameters

$$\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(t)})$$

(that is, compute $\mu_0^{(t+1)}$, $\mu_1^{(t+1)}$ and $p_1^{(t+1)}$). [10 marks]

$\overline{}$ / 10

End of question B4

(a) $\frac{}{3}$ (b) $\frac{}{2}$ (c) $\frac{}{5}$ (d) $\frac{}{10}$ Total $\frac{}{20}$

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