The University of Melbourne

School of Computing and Information Systems

COMP90051
Statistical Machine Learning
2021 Semester 1

Identical examination papers: None

Exam duration: 120 minutes

Length: This paper has 4 pages including this cover page.

Authorised materials: Lecture slides, workshop materials and prescribed reading.

Instructions to students: The total marks for this paper is 120, corresponding to the number of minutes available. The mark will be scaled to compute your final exam grade.

This paper has two parts, A-B. You should attempt all 18 questions.

This is an open book exam. You should enter your answers in a Word document or PDF, which can include typed and/or hand-written answers. You should answer each question on a separate page, i.e., start a new page for each of Questions 1-18 – parts within questions do not need new pages. Write the question number clearly at the top of each page. You have unlimited attempts to submit your answer-file, but only your last submission is used for marking.

You must not use materials other than those authorised above. You are not permitted to communicate with others for the duration of the exam, other than to ask questions of the teaching staff via the LMS 'Exam Support' facility. Your computer, phone and/or tablet should only be used to access the authorised materials, enter or photograph your answers, and upload these files.

Library: This paper is to be lodged with the Baillieu Library.

Continued overleaf . . .

COMP90051 Statistical Machine Learning
Exam

Semester 1, 2021

Total marks: 120

Students must attempt all questions

Section A: Short Answer Questions [50 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each question in 1-3 lines, with longer responses expected for the questions with higher marks.

1. A *recurrent neural network* is as a type of *deep artificial neural network*. In what respect is it *deep*? [4 marks]

2. For the *hard-margin support vector machine* data points that are *support vectors* have a *margin* of 1 from the *decision boundary*. Explain how this can be the case even when support vectors are further than 1 unit away from the decision boundary in *Euclidean space*. [4 marks]

3. For a *support vector machine* with a quadratic *kernel*, in what situation, if any, would it be better to use the *primal program* instead of the *dual program* for training? [4 marks]

4. Explain how *artificial neural networks* can be considered to be a form of non-linear *basis function* when learning a linear model. [4 marks]

5. Consider a setting with high *uncertainty* over model parameters. Describe what effect, if any, this will have for the *maximum likelihood estimate*. [4 marks]

6. Explain why using the training likelihood, $p(\mathbf{y}|\mathbf{X}, \theta)$, for model selection can be problematic when choosing between models from different families. [4 marks]

7. Explain why the use of *momentum* in an optimiser can help avoid getting trapped at *local optima* or *saddle points*, with respect to the training of machine learning models with *non-convex training objectives*. [4 marks]

8. With respect to the training of machine learning models, describe a situation where it would be desirable to use an *estimator* with high *bias*. [4 marks]

9. With respect to learning an *autoencoder*, explain a failure case that can arise from the use of complex non-linear encoder and decoder components. [4 marks]

10. *Conjugacy* between the *likelihood* and *prior* is very important when using *Bayesian models*. However conjugacy is not critical when using the *maximum a posteriori (MAP)* estimator. Explain why this is the case. [4 marks]

11. For *Bayesian linear regression* (as presented in class) explain how the *posterior variance* can vary for different test points. [4 marks]

    The following formulae may be of use:

    > Bayesian regression with prior $\mathbf{w} \sim \text{Normal}\left(0, \gamma^2 \mathbf{I}_D\right)$ and likelihood $y \sim \text{Normal}\left(\mathbf{x}^\top \mathbf{w}, \sigma^2\right)$ leads to a posterior of the form $\mathbf{w} \sim \text{Normal}\left(\mathbf{w}_N, \mathbf{V}_N\right)$ and Gaussian predictive distribution $y_* \sim \text{Normal}\left(\mathbf{x}_*^\top \mathbf{w}_N, \sigma^2 + \mathbf{x}_*^\top \mathbf{V}_N \mathbf{x}_*\right)$, where $\mathbf{w}_N = \frac{1}{\sigma^2} \mathbf{V}_N \mathbf{X}^\top \mathbf{y}$ and $\mathbf{V}_N = \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\gamma^2} \mathbf{I}_D\right)^{-1}$.

12. *Gradient descent* is typically preferred over *coordinate descent*. Given this, why is *coordinate descent* used in the *Expectation Maximisation* algorithm? [6 marks]

Continued overleaf ...

Section B: Long Answer Questions [70 marks]

In this section you are asked to demonstrate your conceptual understanding of a subset of the methods that we have studied in this subject.

Question 13: Neural networks & Regularisation [10 marks]

a) Consider the training objective, $\mathcal{L}(\theta) = -\log P_\theta(\mathbf{y}|\mathbf{X}) + \lambda r(\theta)$, where the second term is a *regulariser*, $\lambda > 0$ and $r(\theta) \geq 0$ is a positive-valued function. For the minimiser of the regularised training objective to be a *maximum a posteriori* estimate, state the general form of the corresponding prior over θ. [4 marks]

b) Given that the output loss of a *feed-forward network*, L, is not a direct function of *hidden-layer weights*, $\mathbf{w}$, how is the derivative of loss with respect to hidden-layer weights, $\frac{\partial L}{\partial \mathbf{w}}$, calculated in *backpropagation*? As part of your answer include the formulation of $\frac{\partial L}{\partial w_j}$ where j is the index of a single weight parameter. [6 marks]

Question 14: Probabilistic models [14 marks]

Consider the conditional probability distribution

$$P(A, B, C, D, E, F) = P(A)P(B)P(C|A)P(D|A, B)P(E|D)P(F|C, E)\,,$$

where $A, B, \ldots G$ are binary valued random variables.

a) Draw the corresponding *directed probabilistic graphical model* (PGM). [4 marks]

b) When performing *Gibbs sampling*, state the distribution used to resample the value of E. Simplify as much as possible. [6 marks]

c) Instead of implementing $P(D|A, B)$ as a *large conditional probability table*, we decide to use a logistic regression model, with three parameters (one for each of the two conditioning variables, and a bias term). Besides reducing the number of parameters, state another important consequence of this change. [4 marks]

Question 15: Kernels [12 marks]

If $k_1(\mathbf{a}, \mathbf{a}')$ and $k_2(\mathbf{b}, \mathbf{b}')$ are both valid kernels over vector valued inputs of size d_1 and d_2 respectively, prove that

$$k_3\left(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\mathbf{a}'\\ \mathbf{b}'\end{bmatrix}\right) = k_1(\mathbf{a}, \mathbf{a}') + k_2(\mathbf{b}, \mathbf{b}')$$

is also a valid kernel over vectors of size $d_1 + d_2$. Note that the symbol $'$ means 'prime' not vector transpose, and $\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}$ denotes concatenation of the two vectors.

Question 16: Vapnik-Chervonenkis Dimension [12 marks]

Consider an *input domain* of six points $x_1, \ldots, x_6$, and *binary classifier family* $\mathcal{F}$ defined by the table of *dichotomies* given below.

x_1	x_2	x_3	x_4	x_5	x_6
1	0	1	1	1	0
0	0	0	0	1	1
0	1	0	1	1	0
0	0	1	0	1	1
0	0	1	1	1	0
1	0	0	0	0	1
1	0	0	1	0	0
1	1	1	0	1	1

a) Calculate VC($\mathcal{F}$) and show a corresponding shattered set of points. [5 marks]

b) List all shattered sets of points. How many are there in total? [7 marks]

Continued overleaf ...

Question 17: Multi-Armed Bandits [8 marks]

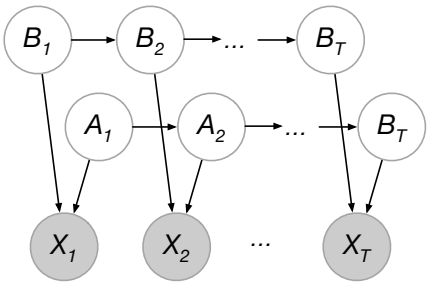
Describe two mechanisms used by the *Upper Confidence Bound* (UCB) *multi-armed bandit* (MAB) to explore arms that so far have no demonstrated high rewards.

The following formulae may be of use:

> UCB arm i value estimate after observing rewards: $Q_{t-1}(i) = \hat{\mu}_{t-1}(i) + \sqrt{\frac{\rho \log(t)}{N_{t-1}(i)}}$ where $N_{t-1}(i)$ is the number of observations of arm i, $\hat{\mu}_{t-1}(i)$ is the average of observed rewards on arm i so far; $\rho > 0$ a hyperparameter. Unobserved arms are initialised with value estimate Q_0 a hyperparameter.

Question 18: Hidden Markov Models [14 marks]

An alternative to the single chain *hidden Markov model* is the multiple chain variant, illustrated below:



Consider sequential data $x_1, x_2, \ldots, x_T$ of T symbols, and two hidden chains with hidden states $a_t, b_t, t \in \{1, \ldots, T\}$, respectively. All random variables are binary valued, i.e., $x_t, a_t, b_t \in \{0, 1\}, t \in \{1, \ldots, T\}$. Chains are *homogenous*, i.e., the probability of transitions and generations are independent of time step, t.

a) What random variables are *marginally independent* of a_1? As we are considering marginal independence, no values of x_t are observed. Explain your answer. [6 marks]

b) We wish to use the *elimination algorithm* to perform inference in this model given observed $\mathbf{x}$. Show the first two steps of the algorithm, for eliminating $X_1 = x_1$ then A_1. As part of your answer, state the formulation of the ψ and m functions and the linear algebraic operations used in each step. [8 marks]

End of Exam