The University of Melbourne

School of Computing and Information Systems

COMP90051
Statistical Machine Learning
2021 Semester 2 – Practice Exam

Identical examination papers: None

Exam duration: 120 minutes

Reading time: 15 minutes

Upload time: 30 minutes additional to exam + reading; upload via Canvas

Late submissions: -1 against final subject mark per minute late, starting 120+15+30 minutes after exam start, up to 30 minutes late maximum. Late submissions permitted by provided OneDrive upload link.

Length: This paper has 7 pages (real exam is longer) including this cover page.

Authorised materials: Lecture slides, workshop materials, prescribed reading, your own projects.

Calculators: permitted

Instructions to students: The total marks for the real exam is 120 (practice exam has fewer questions and marks), corresponding to the number of minutes available. The mark will be scaled to compute your final exam grade.

This paper has three parts, A-C. You should attempt all the questions.

This is an open book exam (see authorised materials above). 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 1–6 – 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 during the course of the exam, but only your last submission is used for marking.

You must not use materials other than those authorised above. You should not use private tutor notes, nor use materials off the Internet. 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 Exam chat tool in Canvas (BigBlueButton). 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.

COMP90051 Statistical Machine Learning
Practice Exam

Semester 2, 2021

Total marks: 120 in real exam based on more questions; 90 in this practice exam

Students must attempt all questions

Section A: Short Answer Questions [25 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.

Question 1: [25 marks]

(a) In words or a mathematical expression, what quantity is minimised by *linear regression*? [5 marks]

Acceptable: The residual sum of errors
Acceptable: The mean-squared error
Acceptable: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (terms are true and estimated labels) or this times a constant

(b) In words or a mathematical expression, what is the *marginal likelihood* for a *Bayesian probabilistic model*? [5 marks]

Acceptable: the joint likelihood of the data and prior, after marginalising out the model parameters
Acceptable: $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$ where $\mathbf{x}$ is the data, θ the model parameter(s), and $p(\mathbf{x}|\theta)$ the likelihood and $p(\theta)$ the prior
Acceptable: the expected likelihood of the data, under the prior

(c) In words, what does $\Pr(A, B \mid C) = \Pr(A \mid C)\Pr(B \mid C)$ say about the *dependence* of A, B, C? [5 marks]

Acceptable: A and B are conditionally independent given C.

(d) What are the *free parameters* of a *Gaussian mixture model*? What algorithm is used to fit them for *maximum likelihood estimation*? [10 marks]

Acceptable 5 mark: For a Gaussian mixture with k components the parameters are probabilities for $(k-1)$ components, a mean vectors for each of the k components, and a symmetric positive-definite covariance matrix for each of the k components.
Acceptable 5 mark: The EM algorithm is appropriate for maximum likelihood estimates.

Section B: Method & Calculation Questions [45 marks]

In this section you are asked to demonstrate your conceptual understanding of methods that we have studied in this subject, and your ability to perform numeric and mathematical calculations. NOTE: in the real exam, a small number of questions from this section will be a bit harder/longer than others.

Question 2: [10 marks]

(a) Consider a 2-dimensional *dataset*, where each point is represented by two *features* and the *label* (x_1, x_2, y). The features are binary, the label is the result of XOR function, and so the data consists of four points $(0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$ and $(1, 1, 0)$. Design a *feature space transformation* that would make the data *linearly separable*. [5 marks]

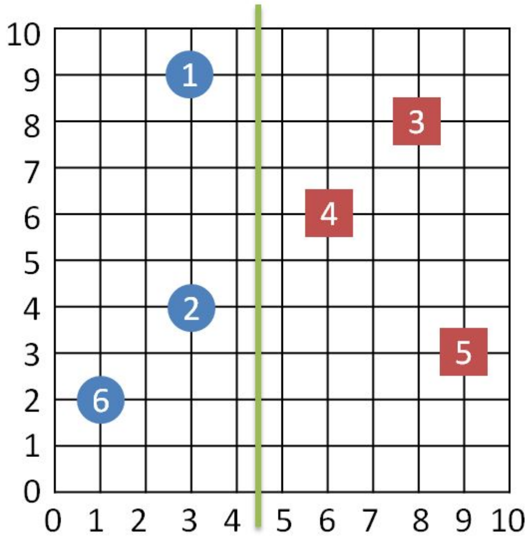
Acceptable: new feature space (x_3), where $x_3 = (x_1 - x_2)^2$

(b) How does SVM handle data that is not linearly separable? List two possible strategies [5 marks]

Acceptable 2.5 marks: using kernels to transform the data.
Acceptable 2.5 marks: using soft-margin SVM to relax the constraints.
Acceptable 2.5 marks: using both soft-margin SVM and kernels.

Question 3: [10 marks]

Consider the data shown below with *hard-margin linear SVM decision boundary* shown between the classes. The right half is classified as red squares and the left half is classified as blue circles. Answer the following questions and explain your answers.



(a) Which points (by index 1–6) would be the *support vectors* of the *SVM*? [5 marks]

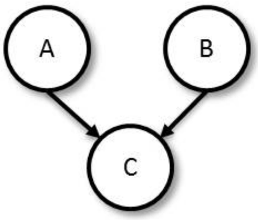
Acceptable: Points 1, 2, 4 would all be support vectors as they all lie on the margin.

(b) What is the value of the *hard margin SVM loss* for point 3? [5 marks]

Acceptable: zero, since the point is on the right side of the boundary and is outside the margin.

Question 4: [15 marks]

Consider the following directed PGM



where each random variable is Boolean-valued (True or False).

(a) Write the format (with empty values) of the *conditional probability tables* for this graph. [5 marks]

```
------------
Pr(A=True)
------------
?
------------

------------
Pr(B=True)
------------
?
------------

------------------
A B Pr(C=True|A,B)
------------------
T T ?
T F ?
F T ?
F F ?
------------------
```

(b) Suppose we observe n sets of values of A, B, C (complete observations). The *maximum-likelihood principle* is a popular approach to training a model such as above. What does it say to do? [5 marks]

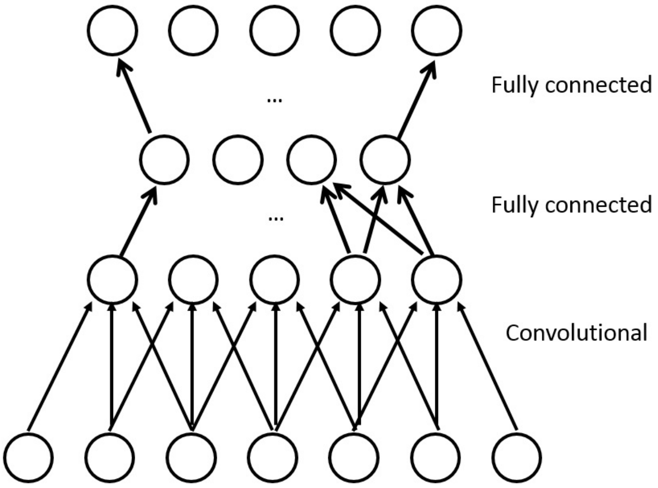
Acceptable: It says to choose values in the tables that maximise the likelihood of the data.
Acceptable: $\arg\max_{tables} \prod_{i=1}^{n} \Pr(A = a_i) \Pr(B = b_i) \Pr(C = c_i \mid A = a_i, B = b_i)$

(c) Suppose we observe 5 training examples: for (A, B, C) — $(F, F, F); (F, F, T); (F, T, F); (T, F, T); (T, T; T)$. Determine *maximum-likelihood estimates* for your tables. [5 marks]

Acceptable: The MLE decouples when we have fully-observed data, and for discrete data as in this case — where the variables are all Boolean — we just count.
The $Pr(A = True)$ is 2/5 since we observe A as true out of five observations. Similarly for B we have the probability of True being 2/5. Finally for each configuration TT, TF, FT, FF of AB we can count the times we see C as True as a fraction of total times we observe the configuration. So we get for these probability of $C = True$ as 1.0, 1.0, 0.0, 0.5 respectively.

Question 5: [10 marks]

How many *parameters* does the following *convolutional neural network* have (exclude the *bias*)? Show your working.



Acceptable: The convolutional network has 3 matrix-valued parameters, adding up the sizes gives $3+5\cdot 4+4\cdot 5 = 43$ parameters in total (assuming that biases are not included).

Section C: Design and Application Questions [20 marks]

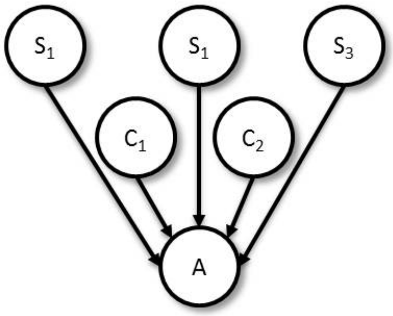
In this section you are asked to demonstrate that you have gained a high-level understanding of the methods and algorithms covered in this subject, and can apply that understanding. Answers should be about 1 page in length for each question.

Question 6: [20 marks]

Your task is to design a system for alerting residents of the Dandenongs that they should evacuate for an impending bushfire. The Dandenongs is an area of Victoria that suffers from regular bushfires in the summer when temperatures are high, and humidity low. When fires are close to a fictional town called Bayesville, you must alert residents that they should evacuate. If fires are too close, then you should not advise evacuation as residents are safer if they stay where they are (at home).

The Country Fire Association has deployed sensors around the area that monitor whether a fire is in progress at each sensor's location; in particular if any of three sensors S_1, S_2, S_3 are 'on' then residents should evacuate. However if either of the closer sensors C_1, C_2 are 'on' then residents should stay put.

(a) Model the above problem as a *directed probabilistic graphical model* (PGM). In particular, you need not provide any probabilty tables, just the graph relating random variables S_1, S_2, S_3, C_1, C_2 and an additional r.v. A for alerting residents to evacuate. [5 marks]



(b) How many *conditional probability tables (CPTs)* should be specified for your model, and what should these tables' dimensions be? [4 marks]

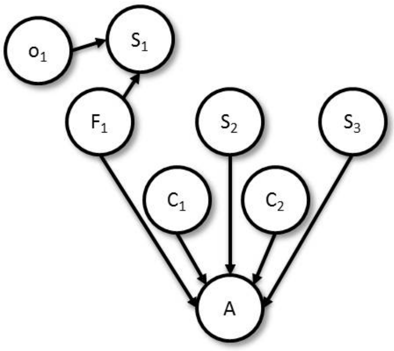
Acceptable: One table per r.v. so 6 tables. All 5 sensor r.v.'s should have tables with one column — a value for the probability that entry is True (probability of False is one minus this). The CPT for A should have 6 columns, one per parent and one for A.

(c) Suppose you are told by the Fire Commissioner that the sensors are always accurate. What could you say about your CPTs? [4 marks]

Acceptable: This doesn't tell you how often the sensors are 'on' so nothing about the sensors' tables. However it tells us that you can trust the sensors and you can determine the CPT for A exactly as $(S_1 \vee S_2 \vee S_3) \wedge \neg(C_1 \vee C_2)$

(d) How would you change your model if the Commissioner then tells you that the sensor S_1 is not perfectly accurate? [4 marks]

Acceptable: We should add a r.v. F_1 as to whether there's really a fire at S_1, and another r.v. o_1 as to whether S_1 is operational. Now we have that the alarm should depend on the unobserved F_1.



(e) Given this final model, assuming you have trained it and completed all the necessarily CPTs, how would you use it to drive the alarm to evacuate? [3 marks]

Acceptable: We can use probabilistic inference — the elimination algorithm — to determine $\Pr(A = True)$ from observations of the five sensors. When doing elimination, the five sensors will be observed (no real summing there) but we will be summing over the unobserved o_1 and F_1.

— End of Exam —