The University of Melbourne

Department of Computing and Information Systems

COMP90049

Introduction to Machine Learning

June 2022

Identical examination papers: None

Exam duration: 120 minutes

Reading time: Fifteen minutes

Length: This paper has 9 pages including this cover page.

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

Calculators: Permitted

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

You should enter your answers in a Word document or PDF, which can include typed and/or handwritten answers. You should answer each question on a separate page, i.e., start a new page for each of Questions 1–7 – 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 exam chat support. Your computer, phone and/or tablet should only be used to access the authorised materials, enter or photograph your answers, and upload these files. The work you submit **must be based on your own knowledge and skills**, without assistance from any person or unauthorized materials.

There is an **embargo on discussing the exam contents** for 48 hours after the end of the exam. You must not discuss the exam with anyone during this time (this includes both classmates and non-classmates.)

COMP90049 Introduction to Machine Learning
Final Exam

Semester 1, 2022

Total marks: 120

Students must attempt all questions

Section A: Short answer Questions [27 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: [27 marks]

(a) Both Naive Bayes and logistic regression are probabilistic machine learning models, and have a model of the underlying data. Explain the difference between the respective models in the context of the task of spam classification. [4 marks]

(b) Consider the mean squared error (MSE) as an alternative to the cross-entropy loss (CE),

$$L_{MSE} : \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2$$

$$L_{CE} : -\sum_{i} y_i \ \log \ \hat{y_i}$$

Here, y_i refers to the true label of the i^{th} instance, and $\hat{y_i}$ to its predicted label. Would you choose MSE or cross-engropy as a loss function when optimizing a supervised 5-way classification model? Justify your choice. [5 marks]

(c) In the following table, fill in each cell with **one** of : {increases, decreases, same} to indicate the most typical effect (column) of different strategies (rows). For example, cell (a) should indicate the effect that *increasing the training data set size* will have on *model bias.* [4.5 marks]

	model bias	model variance	model generalization
increasing the training data set size	(a)	(b)	(c)
increasing the model complexity	(d)	(e)	(f)
stacking three weak learners vs. one individual weak learner	(g)	(h)	(i)

(d) Anna and Bob have developed a classification model. Anna wants to test it using 20-fold cross-validation, but Bob would rather use 3-fold cross-validation. Explain one valid reason in favor of Anne's strategy and one reason in favor of Bob's strategy. [2 marks]

(e) (i) Explain in your own words the problem of *constrained optimization.* (ii) Explain in your own words how this concept relates to evaluating classifiers for fairness in the context of a concrete example. (*N.B. no formula or calculations are necessary, providing the intuitions is sufficient.*) [3 marks]

(f) Feng wants to build a machine learing model to predict how attractive a country is as a travel destination. For each country he has a large list of features, including 'average temperature', 'population (in million)', 'size (km^2)', 'location (longitude, latitude)', 'GDP', 'length of the national anthem (in seconds)', and many more. In the context of Feng's machine learning task, (i) explain both feature selection and feature normalizaton and (ii) explain one difference between the two. [3.5 marks]

(g) Consider a labelled dataset of 20 buildings, where for each building you want to predict a binary label: *keep* or *tear down*. For each building there are three categorical features: (1) its age (<20 years; between 20 and 50 years; >50 years); (2) its insulation quality (high, medium, low); (3) its location (in_nature, near_nature, downtown). (i) Describe how to build a Random Forest classifier, referring to the size and properties of the data set above (e.g., number of instances and features). (ii) Is a Random Forest a suitable model for the given data? Justify your answer by referring back to properties of Random Forests in (i). [5 marks]

Section B: Method Questions [68 marks]

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

Question 2: Probability [6 marks]

The headmaster of the arts department of a university is concerned about her students' health. The arts students share a cafeteria with the science students (but no others). On a typical day, 50% of arts students and 10% of science students want a burger for lunch. 30% of the cafeteria customers are science students. What is the percentage of arts students that eat burgers on a day? [6 marks]

Question 3: Fair classification and mutual information [15 marks]

The local high school is auditing its admission procedure for fairness. Parents have voiced a concern that admissions are impacted by the gender of the student: female students have a higher chance of being admitted than male students. Each application includes a whole range of information, including the student's height, gender, grades, postcode of home address, primary school they graduated from, and hobbies. In 2019, the admission statistics (by gender only) were as follows:

	admitted	not admitted
male	360	240
female	260	120

(a) You want to train a classifier that fairly predicts student admission, without discriminating and group. Describe the kind(s) of bias that you need to take care of in the context of this scenario. [2 marks]

(b) (i) In the context of the scenario above, explain the approach of *fairness through unawareness.* (ii) Is *fairness through unawareness* a valid approach to address the above problem? Justify your answer. [3 marks]

(c) *Data re-weighting* is one strategy for improving the fairness of a ML model. Explain in your own words the intuition behind data re-weighting. Refer to the problem given above, and draw connections to the concepts of statistical association measures typically used for feature selections. [3 marks]

(d) Apply data-reweighing to the data set in the table above. Explain the resulting weights in your own words. *(N.B. Show your mathematical working. Use precision of two or three decimal points.)* [7 marks]

Question 4: Decision Trees and Ensembling [13 marks]

Consider the following data set of seven train instances (1–7) and one test instance (8) for a binary classification problem of predicting whether a student was happy with their exam grade. Each student is characterized by three features: whether they cheated ('cheated'), whether they slept the night before the exam ('slept'), and the number of hours they studied for the exam.

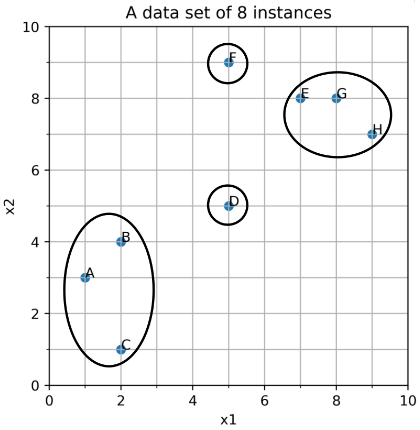
	cheated	slept	#hours studied	happy with exam grade
1	F	T	10	yes
2	F	T	2	yes
3	T	F	5	no
4	F	T	7	no
5	F	F	2	no
6	T	T	10	yes
7	F	F	7	yes
8	T	T	3	no

(a) The feature '#hours studied' is numeric, however, numeric features need some extra attention in order to be used in decision trees. We want to compare the feature when represented in two different ways: (i) represent its values as 4 discrete values (2, 5, 7, and 10); (ii) treat the values as numerical, and discretize them into two equal frequency bins. For both representations, compute the *Information Gain* compared to the root node entropy $H(R)$. Summarize your conclusions in terms of the utility of these two discretization methods 1-2 sentences. *(N.B. Show your mathematical working. Use precision of two or three decimal points and logarithm of base 2.)* [9 marks]

(b) Classify the test instance (8) with each of the decision stumps you built in part (a). Justify your approach. [4 marks]

Question 5: Unsupervised Learning and Anomaly Detection [15 marks]

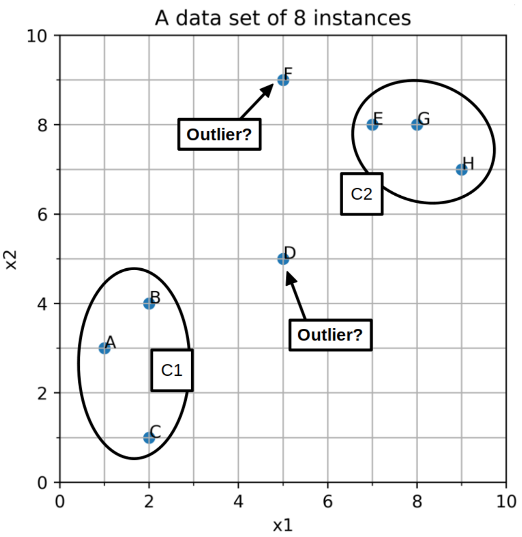
Consider the following data set of eight instances (labelled A...H), each characterized through two features (x1 and x2), and clusters are indicated as black circles. The points are shown visually on the left and coordinates are given for your convenience in the table on the right



Point	(x1, x2)	Point	(x1, x2)
A	(1, 3)	E	(7, 8)
B	(2, 4)	F	(5, 9)
C	(2, 1)	G	(8, 8)
D	(5, 5)	H	(9, 7)

(a) (i) Perform Agglomerative Clustering on the given data, using (1) one step of single link, and (2) one step of complete link clustering (both (1) and (2) should take the clusters in the plot above as their starting point). Use Manhattan distance. (ii) Which method do you find more reliable, and why? *(N.B. Either show your mathematical working, or clearly describe how you arrived at your comclusion in a few sentences. If you show your working, use a precision of two or three decimal points.)* [10 marks]

(b) You realize that points D and F are far away from the remaining points, and you want to test the hypothesis that one or both points are outliers. *[Question continues on next page.]*



(i) Compute the outlier score of F and D wrt. the *inverse of the density.* Use Manhattan Distance, and the k=2 nearest neighbors inside any clusters (i.e., consider only points A, B, C, E, G, H as nearest neighbors). (ii) What do you find? Is your result reliable? Discuss one strategy to improve it. *(N.B. Show your mathematical working. Use precision of two or three decimal points.)* [5 marks]

Question 6: Bias and Variance [19 marks]

This question explores the concepts of bias, variance and their trade-off. First, let us define some notation for relevant concepts:

c1. $x = \{x_1, \ldots, x_n\}$ refers to the observed data points.

c2. $f(x)$ refers to the true function which generated the data.

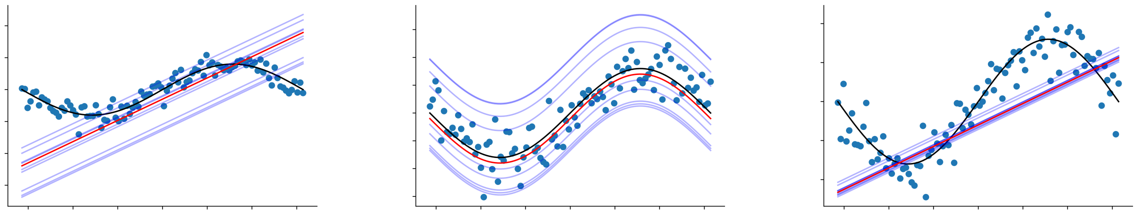
c3. $g(x)$ refers to a single estimate of $f(x)$ by the model.

c4. $E[g(x)]$ refers to the average estimate of from many modelling runs.

c5. $\sigma(x)$ refers to the noise in the data.

(a) In your own words, describe the intuition behind the bias-variance decomposition using the notation given in c1.–c5. above. *(N.B writing only the formula is not enough, you should comment on the individual factors and their significance)* [5 marks]

(b) The following three plots show the result of fitting many functions to a given data set.



(i) Label the x- and y-axes of the plots. (ii) Each plot reflects the concepts defined in c1.–c5. above. For each plot, indicate where each concept is depicted. In addition, indicate how the plots reflect the *bias* and *variance* of the underlying model. *(N.B. In total you should have 7 labels explaining different elements/characteristic of the plot. You may either annotate the plots, or answer the question as written text.)* [8 marks]

Continued overleaf . . .

(c) For each of the three figures, indicate whether the (i) model has high, low bias, (ii) high or low variance and (iii) whether the model is underfitting, overfitting, or appropriately fit. Provide reasons for the model behavior referring to properties of the model and/or data. [6 marks]

Continued overleaf ...

Section C: Design and Application Questions [25 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. Expect your answer to each question to be from one third of a page to one full page in length. These questions will require significantly more thought than those in Sections A–B, and should be attempted only after having completed the earlier sections.

Question 7: Identifying Bots on social media platforms [25 marks]

Professor Bird is a computational linguist who wants to develop a machine learning model that can detect messages on social media platforms that were generated by bots rather than human users. She has collected a large data base of messages, and would like to automatically classify each message into one of two types: 'bot' or 'human'.

Professor Bird has a data set of 1,000,000 messages, 5,000 labelled and 995,000 unlabeled. For **each** message she has available (a) the content mapped to a 56-dimensional embedding; (2) the name of the author; (3) the author's number of followers on the platform; (4) the average # of posts of the author per day. For **some** authors she also has the following features: (5) author location; (6) author nationality. To summarize the data set

ID	Emb	Name	Follower	#posts/day	location	nationality	Bot?
1	$[0.7, \ldots, 8.95]$	Paul	56	0.3	India	?	No
2	$[5.9, \ldots, -34.95]$	OFH_08	2	25	?	?	Yes
3	$[8.7, \ldots, 0.95]$	Dr_Clever	200	2	New York	American	No
...			...				...
5000	$[0.7, \ldots, 8.95]$	Jane1975	1250	3.4	Germany	Spanish	Yes
5001	$[0.7, \ldots, 8.95]$	Zhang	78	5.8	?	Chinese	?
...			...				...
1,000,000	$[0.7, \ldots, 8.95]$	Antonio	0	0.003	?	Spanish	?

In the labelled sample of 5000 instances, 4900 messages are labeled as 'human' and the remaining as 'bot'. Professor Bird is a little rusty in machine learning, and would appreciate your input to help her succeed in her classification task. Please answer the following questions.

(a) First, consider a classification problem. (i) Select a set of four features that you think are helpful, and describe a representation (discrete, continuous, ordinal, ...). (ii) Given your decision in (i), for each of the following algorithms, (a) indicate whether it is appropriate to use and (b) justify your decision. [4.5 marks]

- Gaussian Naive Bayes
- Linear Regression
- Multi-layer perceptron
- Decision Tree
- 30-nearest neighbor
- K-means (K=8)

(b) Design a neural network that takes as input *only* the embedding features to solve the given task in a supervised way. Include the number of input units, depth and width of hidden layers, number of output units, loss function and final layer activation. (*N.B. you may either draw or describe the network.*) [3.5 marks]

Continued overleaf ...

(c) You evaluate your model using accuracy on a stratified test sample of 1000 instances. Your accuracy value seems high, however, when you compare it to a majority class baseline based on the 5,000 labeled instances, you find that both models perform the same. (i) What is the performance of your model? (ii) Describe a better evaluation metric, and justify your choice. (iii) Explain one strategy to improve your model's performance, assuming that you may only modify the data or your current model (i.e., ensembling is *not* an option). [5 marks]

(d) Still failing to get your neural network to work, you want to try an ensembling method, which at the same time can *make use of the unlabeled and the labelled* data. Develop an adequate *ensembling* method which can work in a *semi-supervised* scenario. Justify (i) your choice of method, (ii) the type(s) and number of model; (iii) the features and their representations; (iv) describe in detail the semi-supervised learning process and how your ensemble makes predictions at test time. [12 marks]

— *End of Exam* —