The University of Melbourne

Department of Computing and Information Systems

COMP90049
Introduction to Machine Learning
June 2021

Identical examination papers: None

Exam duration: 120 minutes

Reading time: Fifteen minutes

Length: This paper has 10 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.

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 1–8 – 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 discussion board. 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, 2021

Total marks: 120

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) Name one classifier which can achieve perfect test performance on *any* linearly separable data set; and one classifier which cannot. [2 marks]

(b) You trained a random forest for the task of geolocation classification. Your classifier achieves very high performance on the training set, but low performance on the test set. (i) What is the problem? (ii) Name two possible reasons for this behavior. [3 marks]

(c) For each of the three feature selection methods Wrapper, Filter and Embedded Methods: (i) describe in your own words how it measures the "usefulness" of a feature; (ii) describe in your own words a scenario where it would be more appropriate than the other two methods. [6 marks]

(d) Consider a multi-class classification problem over K classes, where for each instance we observe a label y as a K-dimensional 1-hot vector. We also assume a classifier which predicts $\hat{y}$: a K-dimensional distribution over the same set of labels. y^{max} is the true label of the instance and $\hat{y}^{max}$ is the predicted label (i.e., the class with highest probability assigned by the classifier).

Consider the following loss functions L_a, L_b and L_c, defined for a single input instance,

$$L_a = \sum_{k=1}^{K} y_k \log \hat{y}_k$$

$$L_b = \sum_{k=1}^{K} (y_k - \hat{y}_k)^2$$

$$L_c = \begin{cases} 1, & \text{if } \hat{y}^{max} == y^{max} \\ 0, & \text{otherwise} \end{cases}$$

(i) In your own words, describe how each of the loss function measures the quality of a model prediction. In other words: describe the intuition behind each loss function. [3 marks]

(ii) Can all three loss functions be used to optimise a Multi-layer perceptron? Why (not)? [1 mark]

(iii) Which of the three loss functions is the most appropriate for a classification task? Why? [1 mark]

(e) You are applying leave-one-out cross validation to evaluate your latest machine learning model. Your boss is concerned that your approach leads to high evaluation variance because of the very small test sets (just one instance). Is your boss right? Justify your answer. [2 marks]

(f) Connect the machine learning algorithms on the left with *all concepts* on the right that apply. (*You may copy the answers onto your answer sheet. You do not need to justify your answer.*) [3 marks]

1-Nearest Neighbor
3-Nearest Neighbor
Naive Bayes
Multi-layer perceptron
Decision stump
Decision tree (depth: 3)

Parametric model
Non-parametric model
Probabilistic model
Instance-based model
Linear decision boundary
Non-linear decision boundary
Generative model

(g) You are developing a model for diagnosing a highly contagious disease from a blood sample. Which of the following metrics is the most important to optimize: (a) precision; (b) recall; (c) accuracy; (d) F-1 measure; (e) None of them. Justify your choice. [1 mark]

(h) (i) Explain in your own words the problem of *constrained optimization.* (ii) Explain in your own words how this concepts relates to evaluating classifiers for fairness, naming both the target and the constraint(s). (*N.B. no formula or calculations are necessary, providing the intuitions is sufficient.*) [3 marks]

Section B: Method Questions [75 marks]

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

Question 2: Feature Selection [17 marks]

You want to explore a data set of Nutrition information, where each instance is a fruit or vegetable characterized by three features: `shape`, `color` and `sweetness`. The target class is VITAMIN-C level.

ID	`shape`	`color`	`sweetness`	VITAMIN-C
1	oval	green	5.2	HIGH
2	round	red	5.0	HIGH
3	pointy	orange	1.0	LOW
4	round	red	4.8	HIGH
5	round	purple	4.3	LOW
6	oval	brown	0.3	LOW
7	square	blue	0.2	LOW
8	oval	green	0.8	HIGH
9	round	red	2.1	HIGH

Your favorite classifier accepts discrete features only. You want to compare three methods of feature discretization, and ultimately select the best one.

(N.B. Show your mathematical working for each sub-question.)

(a) Discretize the `Sweetness` feature into three equal-width bins [2 marks]

(b) Discretize the `Sweetness` feature into three equal-frequency bins [2 marks]

(c) Discretize the `Sweetness` using K-means clustering, with K=3 and L1 (Manhattan) distance. Your initial centroids are $c_1 = 0.5, c_2 = 1.0, c_3 = 2.0$, where c_i refers to cluster i. Compute two rounds of updates. [6 marks]

(d) (i) Compute the *Mutual Information* (MI) of the `Sweetness` feature after discretization by K-means (part (c)) with the class label. *(N.B. as defined in the lectures, logarithms should be base 2.)* [6 marks]

(ii) The MI of `Sweetness` after equal-width discretization (part (a)) with the class label is 1.11, and the MI of `Sweetness` after equal-frequency discretization (part (b)) is 0.85. Which of the three discretization methods would you choose based on MI? [1 mark]

Question 3: Classification with Missing Features [9 marks]

Real world data sets very often have *missing features*, i.e., some instances do not have a value for one or more features. More formally, assume that for each data instance i, we observe a label y_i, a set of features $\mathbf{x_i}$ consisting of *observed* features $\mathbf{o_i} = \{o_i^1 \ldots, o_i^m\}$, and *missing* features $\mathbf{m_i} = \{m_i^1, ..., m_i^k\}$ with no associated value.

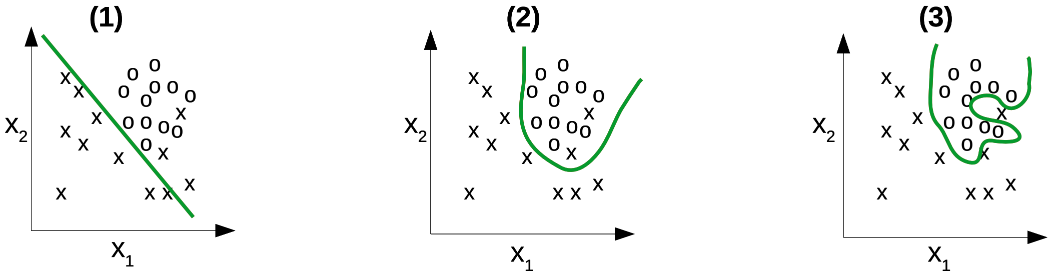
Assume a trained, probabilistic classifier which predicts labels y_i from features $\mathbf{x_i}$: $P(y_i|\mathbf{x_i})$

(a) Using the statistical concept of *marginalisation*, and the notation introduced above, derive mathematically (that is: write equations) a classifier which predicts y_i using both observed features $\mathbf{o}_i$ and missing features $\mathbf{m}_i$. *(N.B. Show your mathematical working.)* [6 marks]

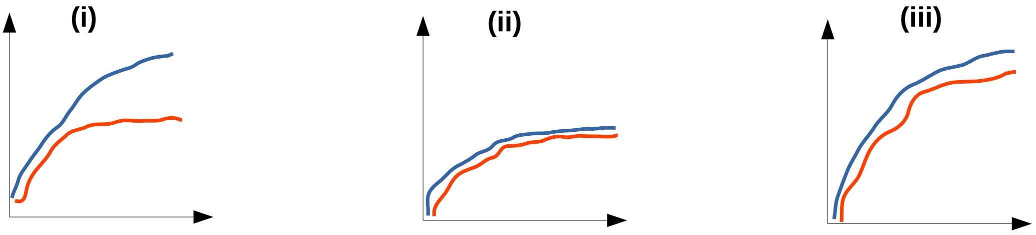
(b) Is your classifier discriminative, generative, neither or both? Justify your answer by referring to your derivation in the first part of the question. [3 marks]

Question 4: Evaluation [11 marks]

Consider the following two sets of plots. Plots (1)–(3) depict three decision boundaries (green), learnt by three different models over the same data set. The data set consists of instances, each described by two features (x_1, x_2) and a class label (x or o):



Plots (i)–(iii) depict learning curves.



(a) Provide a plausible label for the x-axis, y-axis, and the two lines (red and blue) in plots (i)–(iii). [3 marks]

(b) Find the most plausible 1:1 alignment of plots (1)–(3) with plots (i)–(iii). Justify your choice, referring to the concepts of *bias* and *variance*, and *model complexity*. [6 marks]

(c) Out of models (1)–(3), which one would you choose? Justify your choice. [2 marks]

Question 5: Fair Classification [16 marks]

Consider the following data set consisting of 8 training instances, where each instance corresponds to an applicant for a job. Each instance has four features: `work` experience (in years), `education` (in years), LinkedIn page `views`, and gender encoded as binary `female` (value=1 if female, 0 if male). For the purpose of this question, we consider the `female` feature as a protected attribute. Each training instance has a true binary label y which denotes whether the applicant received a high (1) or low (-1) suitability score. We also have access to predicted labels from some classifier , $\hat{y}_{full}$, which was trained to automatically predict the label from *all* available features.

ID	`work`	`edu`	`views`	`female`	y	$\hat{y}_{full}$
1	15	6	1300	0	1	1
2	22	10	1700	1	1	-1
3	44	6	150	1	-1	1
4	33	0	470	1	1	1
5	50	7	700	0	1	1
6	14	3	6	0	-1	1
7	10	4	300	0	-1	-1
8	4	5	130	1	-1	-1

(a) Define in your own words the fairness criterion of *equal opportunity* in the context of the above scenario. [2 marks]

(b) Is the *full* model (column $\hat{y}_{full}$) fair with respect to the concept of *equal opportunity*? *(N.B. Show your mathematical working.)* [3 marks]

(c) (i) Define in your own words the concept of *fairness through unawareness* in the context of the above scenario. (ii) Would the resulting model be a truly fair classifier? Justify your answer [2 marks]

(d) Train a Perceptron implementing *fairness by unawareness* , using the data set given above as training examples. Perform two training steps, i.e., process *only* the first two instances (`ID` 1 and 2) in the data set. Assume the following:

- learning rate $\eta = 0.1$
- bias=1
- all parameters in $\boldsymbol{\theta}^0$ are initialized as 0;
- step function

$$f(z) = \begin{cases} 1 & \text{if } z > 0 \\ -1 & \text{otherwise} \end{cases}$$

(N.B. Show your mathematical working.) [9 marks]

Question 6: Probability [6 marks]

You developed a classifier which predicts whether a German movie will be successful in Australia, or not. "Successful" here is defined as $> 10,000$ viewers in Australia within the first 4 weeks after release. From historical data it is known that 0.5% of all German movies turn out successful in Australia. After some development and evaluation, you find that your classifier has a false positive rate of 4%, and a false negative rate of 1%.

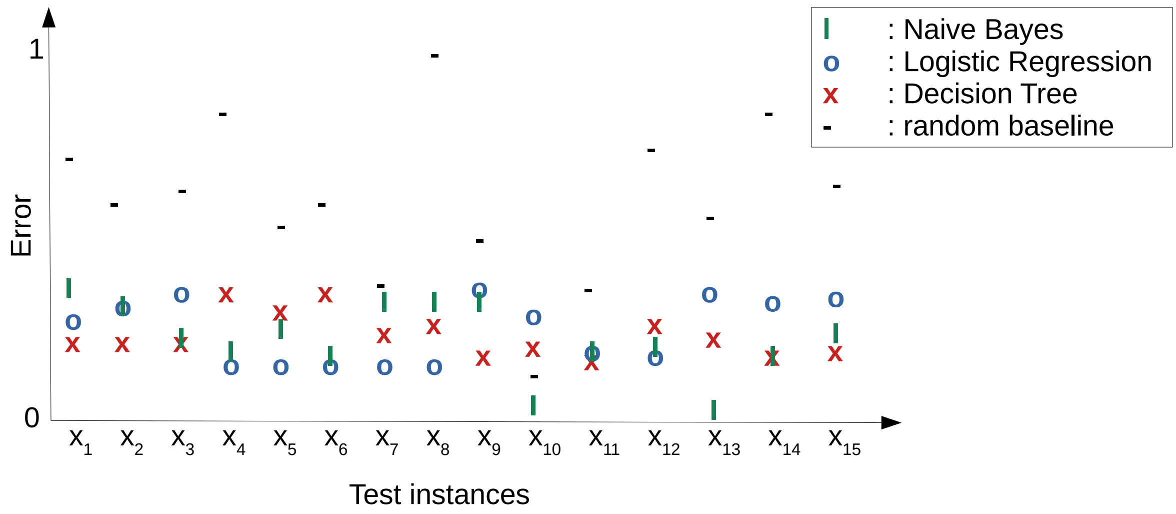
(N.B. Show your mathematical working for each sub-question.)

(a) What are the odds that a random novel German movie will be unsuccessful in Australia? [2 marks]

(b) Your classifier predicts “successful” for a new German movie. What is the probability that the movie will indeed be successful? [4 marks]

Question 7: Ensembling [16 marks]

The following graph shows the error of three classification models and a random baseline on 15 individual test instances $(x_1, \ldots, x_15)$. The error for each test instance is a continuous number between 0 and 1, where 0 is best.



(a) Explain the general concept of *ensembling* in the context of the above scenario. (*N.B. you do not need to provide formulas or perform computations.*) [2 marks]

(b) Would you expect ensembling to improve performance? Why (not)? [3 marks]

(c) Name an appropriate ensembling technique for the above scenario. Justify your choice. [1 mark]

For the remainder of the question, consider a slighlty different scenario:

A team of data scientists has access to a labelled training data set and found that several different classifiers lead to severe {`overfitting, underfitting`}. They consequently decided to apply *Boosting* using AdaBoost algorithm with Decision Stumps as their base classifiers. In doing so, they expects to obtain more {`complex decision boundaries`, `stable predictions`}.

(d) Select the most appropriate term from the {`options`} (`underfitting/overfitting` and `complex decision boundaries/stable predictions`) in the text above. [1 mark]

After $t-1$ iterations, the data scientists obtained the following results for a set of $N = 5$ test instances (only the last base classifier shown):

instances x_i	true labels y	instances weights w^t	predicted labels $\hat{y}^t$
x_1	0	0.1	0
x_2	0	0.5	0
x_3	0	0.05	0
x_4	1	0.05	0
x_5	1	0.3	0

(e) Compute the error rate ϵ_t, classifier weight α_t, and the new instance weights w_{t+1}. Use the following definitions of AdaBoost update formulas:

$$\epsilon_t = \sum_{j=1}^{N} w_t^j \delta(\hat{y}_t \neq y_t)$$

$$\alpha_t = \frac{1}{2} \log_e \frac{1-\epsilon_t}{\epsilon_t}$$

$$w_{t+1}^j = w_t^j \times \begin{cases} e^{-\alpha_t} & \text{if } \hat{y}_t = y_t \\ e^{\alpha_t} & \text{if } \hat{y}_t \neq y_t \end{cases}$$

(N.B. Show your mathematical working.) [5 marks]

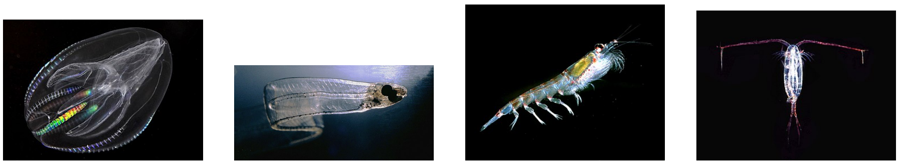
(f) (i) Compare the new weights w_{t+1} to the weights w_t in the table above, and explain your observations in the context of the ideas underlying the AdaBoost algorithm. (ii) What are the two functions of the classifier weight α_t? [4 marks]

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. 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 8: Plankton Classification [20 marks]

Professor Shell is a marine biologist specialising in plankton. She has collected a large data base of plankton images, and would like to automatically classify the depicted instance into one of four plankton types (see Figure 1). For each image, she is able to obtain the following measurements which she wants to use as features: length, number of legs, number of eyes, circumference, and color intensity. She has labelled a small data set of 100 plankton images with the correct plankton type, and has an additional 2000 unlabeled images of varying quality (resolution), brightness, zoom and angle.



Ctenophore Eel larva Antarctic krill Copepod

Figure 1: Four types of plankton. (Image source: Wikipedia)

Professor Shell is an expert on plankton, but does not know much about machine learning. She requires input from a machine learning scientist to help her succeed in her classification task. Please answer the following questions.

(a) For each of the following algorithms, (a) indicate whether it is appropriate to use and (b) justify your decision. [2.5 marks]

- Multinomial Naive Bayes
- Decision Tree
- 30-nearest neighbor
- K-means (K=8)
- K-means (K=4)

You ultimately decide to design a neural network (NN).

(b) How many input units and output units would your NN have? [1 mark]

(c) Would your NN have hidden units? Justify your answer. [1 mark]

(d) What would be the activation function of the final layer? [1 mark]

(e) What learning algorithm would you use? Justify your choice. [1.5 marks]

(f) Considering your training set size($n = 100$) how would you evaluate your model, making sure that you obtain a reliable estimate of its generalization performance? Describe all steps of your chosen evaluation strategy. [4 marks]

(g) After evaluation, performance is not quite satisfactory. You want to improve model performance using all the available resources mentioned above. (i) Select an appropriate machine learning algorithm and justify your choice. (ii) Explain the algorithm in the context of this data set. (iii) Justify any settings of the algorithm you may need to decide on. [5 marks]

(h) Your classifier is working well, and you look forward to a break. However, all of a sudden Professor Shell presents you with a picture of a plankton unlike anything she has seen before! (i) Describe two reasons for why Professor Shell may be encountering a highly unusual data instance. (ii) Using all your machine learning knowledge, how would you help Professor Shell to make sense of the new data instance? [4 marks]

— *End of Exam* —