Machine Learning (COMP30027_2020_SM1)

This is a preview of the published version of the quiz.

Started: Jul 3 at 2:51

Quiz Instructions

Academic Integrity Declaration

By commencing and/or submitting this assessment I agree that I have read and understood the University's policy on academic integrity. (https://academicintegrity.unimelb.edu.au/)

I also agree that:

1. Unless paragraph 2 applies, the work I submit will be original and solely my own work (cheating);
2. I will not seek or receive any assistance from any other person (collusion) except where the work is for a designated collaborative task, in which case the individual contributions will be indicated; and,
3. I will not use any sources without proper acknowledgment or referencing (plagiarism).
4. Where the work I submit is a computer program or code, I will ensure that:
   a. any code I have copied is clearly noted by identifying the source of that code at the start of the program or in a header file or, that comments inline identify the start and end of the copied code; and
   b. any modifications to code sourced from elsewhere will be commented upon to show the nature of the modification.

This exam opens at 3.00 PM Australian Eastern Standard Time (AEST) on Monday 15/06/2020 in Canvas (lms.unimelb.edu.au). The exam must be completed by 3.15 PM AEST on Tuesday 16/06/2020. This exam has 15 minutes of reading time, and 120 minutes of writing time. You have a 24 hour window in which to complete and submit the exam.

The University of Melbourne
School of Computing and Information Systems

Final Examination, Semester 1, 2020
COMP30027 Machine Learning

Reading Time: 15 minutes. **Writing Time:** 2 hours

Total Time: 2.25 hours

Instructions to Students:

The exam includes 32 questions worth a total of 120 marks, making up 60% of the total assessment for the subject.

- This exam includes a combination of multiple-choice questions, short-answer questions, and longer-response questions. Please answer all questions in the fields provided.
- This is a timed quiz. The time remaining is shown in the quiz window and will continue to count down even if you leave the Canvas site.
- It is recommended that you do not close your browser while working on this quiz.
- At the end of the time limit, your answers will be submitted automatically.

Authorised Materials: This exam is open-book. While undertaking this assessment you ***are permitted to***:

- make use of textbooks and lecture slides (including electronic versions) and lecture recordings
- make use of your own personal notes and material provided as part of tutorials and practicals in this subject
- make use of code that has been provided as part of this subject, or that you have written yourself
- use calculators, code, or mathematical software to compute numeric answers

While you are undertaking this assessment you ***must not***:

- make use of any messaging or communications technology
- make use of any world-wide web or internet-based resources such as Wikipedia, Stack Overflow, or Google and other search services
- act in any manner that could be regarded as providing assistance to another student who is undertaking this assessment, or will in the future be undertaking this assessment.

The work you submit ***must be based on your own knowledge and skills***, without assistance from any other person.

Technical support

This exam is a Canvas Quiz. Technical support for this exam can be accessed at: **https://students.unimelb.edu.au/your-course/manage-your-course/exams-assessments-and-results/exams/technical-support** (https://students.unimelb.edu.au/your-course/manage-your-course/exams-assessments-and-results/exams/technical-support)

Additional information about Canvas Quizzes, including troubleshooting tips, can be found **here** (https://students.unimelb.edu.au/your-course/manage-your-course/exams-assessments-and-results/exams/exam-types) (scroll down to the Canvas Quiz section).

If you have questions about the exam content, please post in the LMS Discussion board.

Short response questions

This section asks you to demonstrate your conceptual understanding of various methods we have studied in this subject, your ability to apply them or evaluate them in the context of specific cases, and your ability to perform the numeric calculations involved.

Question 1 — 2 pts

Classify the following attributes as discrete, ordinal or continuous:

A. student ID number discrete

B. number of followers on Twitter continuous

Question 2 — 2 pts

Consider the following instances from the weather dataset:

outlook	temperature	humidity	windy
rainy	cool	high	yes
sunny	cool	medium	yes

What is the Hamming distance between these two instances?

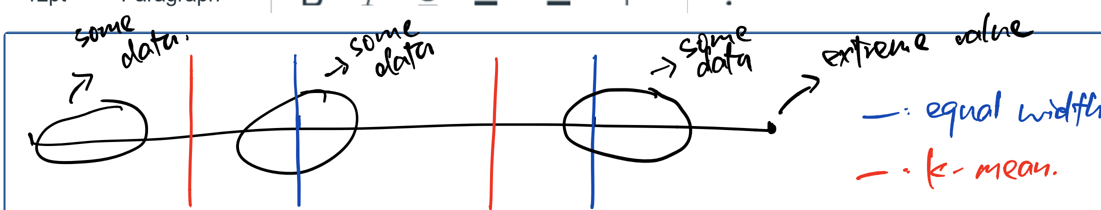
2

Question 3

2 pts

Describe a dataset for which a k-means discretisation strategy would be better than equal-width discretisation. (1-2 sentences)

12pt Paragraph



p

A dataset with extreme value

0 words </>

Question 4

2 pts

What are the requirements on the performance of the base classifiers when designing a classifier combination?

12pt Paragraph

There is no extremely strong one. Each one is reasonably good.

Errors should be uncorrelated, so they do not make same mistakes.

p

0 words </>

Question 5

2 pts

Given a labelled dataset with 500 instances, 80% of the instances (400) are used as a training set and the remaining 20% of the instances are used as test set. Two models A and B are built, and the accuracies of these two models are shown in the following table:

	Accuracy	
	Model A	**Model B**
Training	0.81	0.75
Test	0.65	0.70

Which model would you select to predict unseen instances? Briefly explain your reasoning. (1-2 sentences)

12pt Paragraph B I U A T²

B.
Less prone to over-fitting.
Higher testing accuracy.

p 0 words </>

Question 6

2 pts

What are the possible solutions to reduce evaluation variance? Select all that apply.

- [ ] Increase the holdout partition size for the training set.
- [x] Use majority voting to combine multiple classifiers, and evaluate the performance of ensemble classifiers on the test set.
- [x] Use repeated random subsampling and run multiple evaluations.
- [x] Use leave-one-out cross-validation.

Question 7

2 pts

Suppose you are using gradient descent to fit a linear regression model. After plotting the curve of the loss function of the gradient descent, you find that the value decreases, but only very slowly. What could be the reason and what should you do? (1-2 sentences)

12pt Paragraph B I U A T²

Learning rate is too small. => weights are updated in small amount, so is error. Increase learning rate.

p 0 words </>

Question 8 **1 pts**

Which of these design changes would make a neural network unable to learn arbitrary basis functions (no longer a universal approximator)?

- Do not include any regularisation
- Reduce the architecture to just a single hidden layer
- ✓ Replace sigmoid activation functions with identity (linear) activations
- Replace convolutional layers with fully-connected layers

Question 9 **2 pts**

For a binary classification problem, your task is to classify 10 instances using a combination of 3 base classifiers, each of which has an accuracy of 90%. You can assume the classifier performance is stable over different subsets of instances.

If the majority voting method is used, what is the maximum and minimum accuracy you can obtain on the 10 test instances?

12pt Paragraph B I U A T²

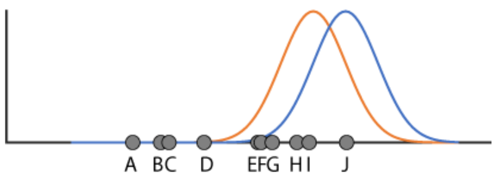
100% if three base classifiers make different errors.
90% otherwise.

p 0 words </>

Question 10

4 pts

The figure below shows an intermediate stage of an EM algorithm. The EM algorithm is fitting a Gaussian mixture model (k = 2) to a set of 10 instances, represented by the orange and blue curves) to a set of 10 instances, labelled A-J:



Suppose this is the state of the EM algorithm at the end of an expectation step. Explain what will happen on the next maximization step, making specific references to the figure. Do not report formulas or try to compute exact values -- qualitatively explain what is computed in this step, using the figure. (About 3 sentences)

12pt Paragraph

Both have lower mean => two curves move left. Orange curve moves further left.
Variance of two gaussian distribution also change. Both curves will fatten a bit.

p 0 words </>

Naïve Bayes

The follow questions relate to probability and naïve Bayes.

Question 11

1 pts

A researcher has collected the following dataset on two species of fish (the “?” symbol indicates missing data):

Fin_length 2	**Stripe** 2	**Tail_shape** 2	**Species**
Short	Dark	Wide	A
Short	?	Narrow	A
Short	Pale	Narrow	A
Long	Pale	Wide	A
Short	Dark	Narrow	B
Short	Dark	Wide	B

Long	Pale	?	B
Long	Pale	Narrow	B

What is the marginal probability P(Fin_length == Long)?

3/8

Question 12 **1 pts**

What is the joint probability P(Fin_length == Short, Species == A)?

3/8

Question 13 **8 pts**

Use this data to train a naïve Bayes classifier to distinguish the two species of fish. Use Laplacian (add-1) smoothing. Give the classification result for a new instance with the attributes:

Fin_length == Long, Stripe == Pale, Tail_shape == Narrow

Report the posteriors that the classifier computes for each class to support your answer.

12pt Paragraph

$$P(A|Long, Pale, Narrow) = P(Long|A)P(Pale|A)P(Narrow|A)P(A)$$
$$= \frac{1+1}{4+2} \cdot \frac{2+1}{3+2} \cdot \frac{2+1}{4+2} \cdot \frac{4}{8}$$

$$P(B|Long, Pale, Narrow) = P(Long|B)P(Pale|B)P(Narrow|B)P(B)$$
$$= \frac{2+1}{4+2} \cdot \frac{2+1}{4+2} \cdot \frac{2+1}{3+2} \cdot \frac{4}{8}$$

p 0 words

$\Rightarrow$ B is predicted class.

Decision Trees

The following questions are related to decision tree algorithms.

Question 14 **2 pts**

Given a dataset as follows:

A1	A2	A3	Class
False	Medium	True	+
False	Low	True	-
True	Medium	False	-
True	High	False	-
True	High	False	+
False	High	False	+
False	High	False	+
True	Medium	False	-
True	Medium	True	+
False	Low	True	-

What is the entropy of this dataset regarding the class label?

$-\left(\frac{5}{10}\log\frac{5}{10}+\frac{5}{10}\log\frac{5}{10}\right)=1$

Question 15

3 pts

What is the information gain of A1?

12pt Paragraph

$H(Class|A1=True)=-\left(\frac{2}{5}\log\frac{2}{5}+\frac{3}{5}\log\frac{3}{5}\right)=0.97095$

$H(Class|A1=False)=-\left(\frac{2}{5}\log\frac{2}{5}+\frac{3}{5}\log\frac{3}{5}\right)=0.97095$

$MI(A1)=\frac{5}{10}\times0.97095+\frac{5}{10}\times0.97095=0.97095$

p $\Rightarrow IG(A1)=1-0.97095=0.02905$.

0 words

Question 16

3 pts

What is the gain ratio of A1?

12pt Paragraph

$$SI(A1) = H(A1) = -\left(\frac{5}{10}\log\frac{5}{10} + \frac{5}{10}\log\frac{5}{10}\right) = 1$$

$$GR(A1) = \frac{IG}{SI} = \frac{0.02905}{1} = 0.02905.$$

p 0 words

Question 17

8 pts

What are the Chi-square values of A1 and A3? Which attribute, A1 or A3, is better at predicting the class based on Chi-square values? Show your calculation.

12pt Paragraph

For A1:

O	+	-	Total
True	2	3	5
False	3	2	5
Total	5	5	10

E	+	-	Total
True	2.5	2.5	5
False	2.5	2.5	5
Total	5	5	10

$$\chi^2(A1) = \frac{(2-2.5)^2}{2.5} + \frac{(3-2.5)^2}{2.5} + \frac{(3-2.5)^2}{2.5} + \frac{(2-2.5)^2}{2.5} = 0.4.$$

p 0 words

For A3:

O	+	-	Total
True	2	2	4
False	3	3	6
Total	5	5	10

E	+	-	Total
True	2	2	4
False	3	3	6
Total	5	5	10

$\chi^2 = 0.$

Evaluation

The following questions pertain to model evaluation.

Question 18

2 pts

A machine learning algorithm was trained to classify images of objects. It was tested on a set of 100 test images (20 images per class). The performance is summarized in the following confusion matrix. The number in each cell indicates the number of images of a particular class given a particular label:

		Classifier prediction					
		Bird	Car	Dog	Horse	Ship	
True class	Bird	9	2	3	3	3	20

	Car	1	9	2	4	4	20
	Dog	1	2	11	2	4	20
	Horse	4	3	1	10	2	20
	Ship	1	1	3	1	14	20
		16	17	20	20	27	

What is the accuracy of this classifier?

$$\frac{9+9+11+10+14}{100} = 0.53.$$

Question 19

3 pts

What is the one-vs.-rest precision of this classifier for the “Car” class? Briefly explain what this value represents (explain qualitatively, do not simply state the formula). (1-2 sentences)

12pt Paragraph

True \ Predict	Car	Not-car
Car	9	11
Not-car	8	

$$\text{Precision} = \frac{9}{9+8} = \frac{9}{17}.$$

Out of predicted cars, $\frac{9}{17}$ of them are actually cars.

p 0 words </>

Question 20

3 pts

What is the one-vs.-rest recall of this classifier for the “Car” class? Briefly explain what this value represents (explain qualitatively, do not simply state the formula). (1-2 sentences)

12pt Paragraph

$$\text{Recall} = \frac{9}{9+11} = \frac{9}{20}.$$

Out of all actual cars, $\frac{9}{20}$ of them can be identified correctly as cars.

p

$$F_1(\text{car}) = \frac{2 \cdot \frac{9}{17} \cdot \frac{9}{20}}{\frac{9}{17} + \frac{9}{20}} = 0.486$$

0 words </>

Question 21 — 4 pts

Is this classifier better at detecting cars or ships? Report the one-vs.-rest F1 score for each class to support your answer.

12pt Paragraph B I U A T²

True \ Predict	Ship	Not-ship
ships	14	6
Not-ships	13	

$$P(\text{ship}) = \frac{14}{14+13} = \frac{14}{27}$$

$$R(\text{ship}) = \frac{14}{14+6} = \frac{14}{20}.$$

$$F1(\text{ship}) = \frac{2 \cdot \frac{14}{27} \cdot \frac{14}{20}}{\frac{14}{27} + \frac{14}{20}} = 0.596.$$

p 0 words </>

⇒ Better at detecting ships.

Instance-based learning

The following questions are about instance-based learning.

Question 22 — 5 pts

The following dataset shows product information and customer purchase decision for 6 products:

Product_ID	Price ($) (1~8)	Health_star_rating	Brand (either 0 or 2)	Purchase_decision
A	5	3 6	Private (1, 0)	Yes
B	6	4 8	Private (1, 0)	Yes
C	3	4 8	National (0, 1)	Yes
D	8	5 10	National (0, 1)	No
E	1	2 4	Private (1, 0)	No
F	7	2 4	Private (1, 0)	No

Choose sensible distance metrics based on the data above, and use it to predict the Purchase_decision label of the test instance (Product_ID = G, Price = $4, Health_star_rating = 4 8, Brand = National) according to the method of 1-NN. Briefly describe your distance metric in 1-2 sentences, show your calculation and give the prediction label for this instance.

Double "Health_star_rating"

One-hot Brand

Use Manhattan distance

$d(A,G) = 1+2+2 = 5$

$d(B,G) = 2+0+2 = 4$

$d(C,G) = 1+0+0 = 1$

$d(D,G) = 4+2+0 = 6$

$d(E,G) = 3+4+2 = 9.$

$d(F,G) = 3+4+2 = 9.$

⇒ Nearest neighbor: C

Prediction: Yes.

12pt Paragraph

p 0 words </>

Question 23

2 pts

If 3-NN is used instead, what is the prediction for the instance (Product_ID = G, Price = $4, Health_star_rating = 4, Brand = National)? Use majority voting (no weighting). Give the Product_IDs of the neighbours to support your answer.

12pt Paragraph

Neighbors: A, B, C
Prediction: Yes.

p 0 words </>

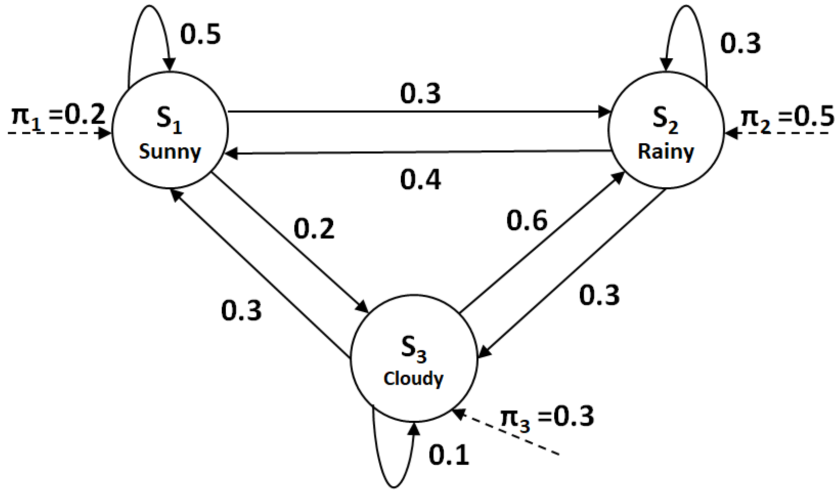
Hidden Markov Models

The next two questions relate to hidden Markov models (HMMs).

Question 24

1 pts

Suppose we have constructed a hidden Markov model (HMM) to represent the tasks completed on different days. The day's weather (Sunny, Rainy, Cloudy) is the hidden state, and the task completed (Task 1, Task 2, Task 3) is the observation. The transition probabilities and initial state distributions are illustrated in the figure below. The output probability matrix is given in the table below.



	Task 1	Task 2	Task 3
Sunny	0.6	0.3	0.1
Rainy	0.05	0.15	0.8
Cloudy	0.3	0.4	0.3

What are the transition probabilities (a_{21}, a_{22}, a_{23}) ?

12pt Paragraph

$a_{21} = 0.4$
$a_{22} = 0.3$
$a_{23} = 0.3$

p 0 words

Question 25 4 pts

What is the probability of observing Task 1 on the first day (t=1)? Report the forward probabilities $\alpha_1(S_i)$ used to compute the result.

12pt Paragraph

$\alpha_1(S_1) = 0.2 \times 0.6 = 0.12.$
$\alpha_1(S_2) = 0.5 \times 0.05 = 0.025.$
$\alpha_1(S_3) = 0.3 \times 0.3 = 0.09$

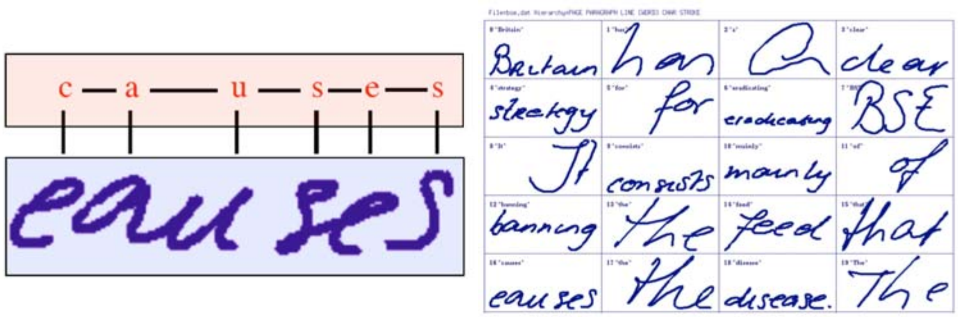
$P(O_1 = \text{Task } 1) = 0.12 + 0.025 + 0.09$
$= 0.235.$

p 0 words

Question 26

3 pts

Optical Character Recognition (as shown in the following figures) is a typical application of HMM. Briefly describe at conceptual level the states of the HMM model and the values stored in the output probability matrix B in the context of ab OCR application. (2-3 sentences)



12pt Paragraph

States: 26 alphabets.

Observations: images of characters.

Matrix B: Rows are states, columns are observations.
entries are probabilities of observing
a character image given state.

p 0 words </>

Neural networks

The following questions are related to neural networks.

Question 27

6 pts

Consider the following dataset with attributes [x1, x2] and class y:

Xo	x1	x2	y
1	0	2	1
1	0	-2	1
1	3	1	0
1	1	0	0

Suppose you wish to train a perceptron to predict y. The perceptron starts with bias 0 and weights [1, 1] and has a learning rate of 1. What will the perceptron's bias and weights be at the end of the first training epoch? (Assume the instances are input to the perceptron in the same order as shown in the table.)

i: ith instance

12pt Paragraph B I U A T²

$w^{(0)} = (b, w_1, w_2) = (0, 1, 1)$

$i=1$ $w^T x = (0,1,1)\cdot(1,0,2) = 2 \Rightarrow \hat{y}_i = 1 = y_i \Rightarrow$ no update.

$i=2$ $w^T x = (0,1,1)\cdot(1,0,-2) = -2 \Rightarrow \hat{y}_i = 0 \neq y_i \Rightarrow \lambda(y_i - \hat{y}_i) = 1 \Rightarrow w^{(1)} = (0,1,1) + (1)(1,0,-2)$
$= (1,1,-1)$

$i=3$ $w^T x = (1,1,-1)\cdot(1,3,1) = 3 \Rightarrow \hat{y}_i = 1 \neq y_i \Rightarrow \lambda(y_i - \hat{y}_i) = -1 \Rightarrow w^{(2)} = (1,1,-1) + (-1)(1,3,1)$

p 0 words </>

$= (0,-2,-2)$

$i=4$ $w^T x = (0,-2,-2)\cdot(1,1,0) = -2 \Rightarrow \hat{y}_i = 0 = y_i \Rightarrow$ no update

Question 28 $\Rightarrow b = 0$, weight $= [-2,-2]$.

3 pts

How do the assumptions of an SVM differ from the assumptions of a neural network? (2-3 sentences)

12pt Paragraph B I U A T²

SVM with kernel tricks assume data is linear separable when projected to higher dimensions.

Neural network assumes a series of linear combination with non-linear activations can map data to its label.

p 0 words </>

Question 29

3 pts

Describe two design choices that allow a neural network to recognize a set of features no matter where it appears in the input, with a brief explanation of each. (1-2 sentences.)

12pt Paragraph

Conv Net with parameter sharing: parameters in a filter represent a feature, and is shared when convolving the whole image

Pooling: Add translation invariance. Features in one region is aggregated. summarised.

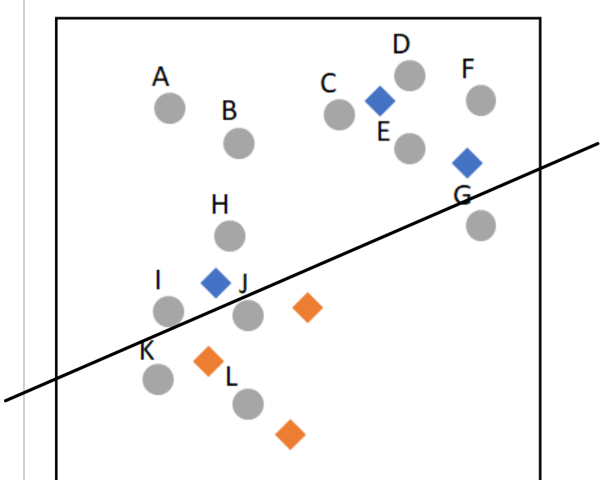
p 0 words

Un- and semi-supervised learning

The following questions relate to un- and semi-supervised learning.

Question 30 **4 pts**

Suppose you are training a perceptron on the dataset shown below. Instances have two attributes, which are continuous numeric values, represented by the x and y axes. The blue and orange diamonds represent labelled instances from two classes, and the grey circles (A-L) are unlabelled instances.



Describe the behaviour of a self-training system on this dataset. Which instance(s) would be labelled first and why? (Answer in about 3-4 sentences)

12pt Paragraph

It firstly tries to separate labelled data, which is possible.
Instances like A and C would be labelled first as they are distant from decision boundary.
It will then label other instances far away from decison boundary, but unsure about instances like I, J, K.

p 0 words </>

Question 31

5 pts

Suppose you have been given a labelled dataset for a classification problem. To learn more about the data, you first try an unsupervised clustering method (ignoring the labels) and evaluate the results. You find that the clusters have high cohesion and high purity.

What does it mean for a clustering result to have high cohesion and high purity? Will this dataset be easy or difficult for a nearest-neighbour algorithm to classify? Explain qualitatively, without using mathematical formulas. (Answer in about 4 sentences)

12pt Paragraph

High cohesion: instances in a cluster are close together, centred.
High purity: majority label of instances in a cluster has a very large proportion.

It would be an easy task for nearest neighbour, as a test instance should have its neighbours belonging to the same cluster since these neighbours are close together due to high cohesion. Due to high purity, these neighbours would have little disagreement, so label for test instance is quite certain.

p 0 words </>

Question 32

25 pts

Design and Application

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 to respond in a full paragraph for each of the questions below. These questions will require significantly more thought than those in the previous portion of the exam and it is recommended to attempt this section only after completing the earlier sections.

Growing a better tomato

A large farm has asked you to help them figure out optimal growing conditions for their tomato crop. They grow several different varieties of tomato, and they rate each tomato plant's produce on a three-point flavour quality scale ("poor," "acceptable," or "excellent"). However, assessing the quality of tomatoes is costly – it takes time, and the tomatoes sent out for taste-testing can't be sold afterwards – so the farm would like to develop a method to predict the quality label from the growing conditions of the tomato plant. (And, of course, they're also interested in learning what factors go into making the best possible tomatoes!)

The farm has provided you with a dataset that includes information about the growth conditions of each tomato plant and the ground-truth quality label. The data for each plant is a weekly summary of the growing conditions for that plant (e.g., the amount of sunlight, temperature, water, etc.) and some details of the plant's growth, such as its height, for every week from planting until harvest. The farm provides you this data for three of their eight tomato varieties, but the goal is to build a model that can predict quality in all varieties of tomato plant.

Since assessing the quality of tomatoes is costly, the amount of labelled data for each of the three tomato varieties is fairly limited. However, the farm can also provide a large amount of unlabelled data for each of the three varieties of tomato plant. This dataset includes all of the same details about the growth conditions of the plants, but does not include the quality of the tomatoes.

The objective is to build a prediction system with the data available. At the end of the year, your system will be tested on a dataset which includes all eight varieties of tomato. You will be sent the results of this evaluation, but you will only receive the true labels and your system's predicted labels for each plant, not the growth information used to make the predictions. You can use this information to update your algorithm, if you wish.

- What machine learning system would you use? Why? Explain the assumptions you are making.
- How would you train and evaluate this system?
- Would you use the additional datasets (the unlabelled data and end-of-year test results)? If so, how would you use them? Provide your justification for each choice.

12pt Paragraph B I U A T²

- Active learning, with SVM as classifier.

  why: active learning can make use of unlabelled data, gradually enlarge training set.

  SVM: I intend to treat growing conditions in different weeks as different features.

  => Many features, few instances, suitable for SVM.

  Assumptions: Linear separable.

  we have an oracle that can label quality correctly.

p 0 words </>

Not saved Submit Quiz

OR choose maybe logistic regression for its ability to predict probability

- Train: features: week1_sunlight, week1-temperature, ...

  week2_sunlight, week2-temperature, ...

  procedure: train SVM using labelled instances,

  predict on unlabelled instances,

  pick least confident instance to query,

  add this instance to training data.

  Query should give preference to other varieties.

  Evaluate: hold out part of labelled data for evaluation.

use accuracy and F1 for each class as key metrics.
Detailed evaluation can be done using confusion matrix,
too see how each class performs.
Also, can separately evaluate different varieties

- Unlabelled: used for active learning, to enlarge training set.

End-of-year test results: used to assess how metrics differ across:
- different classes
- different varieties.