

Preliminaries

Restrictions/Allowances

- You are allowed a calculator
- This test is closed book
- You may **not** code
- Access to any other Moodle course or page is forbidden and will be logged

Rounding

- All numerical answers must be rounded to 3 decimal places.
- Unless otherwise state, when your calculations involve multiple steps, round any intermediate results to 5 decimal places before using them in the next step.

Question 1

Correct

Mark 1.00 out of 1.00

Which of the following tasks is best described as a supervised learning problem?

- ☐ a. Finding the most common words in a collection of documents.
- ☒ b. Predicting house prices based on features like size and location. ✓
- ☐ c. Grouping customers based on purchasing habits.
- ☐ d. Teaching a robot to navigate a maze through trial and error.

The correct answer is: Predicting house prices based on features like size and location.

Question 2

Correct

Mark 1.00 out of 1.00

The "Curse of Dimensionality" primarily refers to the issue that:

Select one:

- ☐ a. Models become too simple as features increase.
- ☐ b. Validation sets become ineffective with many features.
- ☒ c. Required training data often grows exponentially with the number of features. ✓
- ☐ d. Extracting meaningful features becomes impossible with high dimensions.

The correct answer is: Required training data often grows exponentially with the number of features.

Question 3

Correct

Mark 1.00 out of 1.00

In machine learning, the 'validation set' is primarily used to report the final, unbiased performance of a fully trained model.

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 4

Correct

Mark 1.00 out of 1.00

A model with high bias typically performs very well on the training data but poorly on unseen test data.

Select one:

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 5

Correct

Mark 1.00 out of 1.00

A model's predictions change drastically when trained on slightly different subsets of the same overall training dataset. This instability is characteristic of:

Select one:

- ☒ a. High Variance ✓
- ☐ b. High Bias
- ☐ c. Low Variance
- ☐ d. Low Bias

The correct answer is: High Variance

Question 6

Correct

Mark 1.00 out of 1.00

The bias-variance trade-off implies that decreasing model complexity always reduces both bias and variance.

Select one:

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 7

Correct

Mark 1.00 out of 1.00

Which type of machine learning typically involves learning from (potentially delayed) rewards received through interaction with an environment?

Select one:

- ☐ a. Unsupervised Learning
- ☐ b. Supervised Learning
- ☐ c. Semi-supervised Learning
- ☒ d. Reinforcement Learning ✓

The correct answer is: Reinforcement Learning

Question 8

Correct

Mark 1.00 out of 1.00

The choice of features (representation) used to describe the data is fundamental in machine learning, but a sufficiently powerful learning algorithm can always find a good solution regardless of how the data is initially represented.

Select one:

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 9

Correct

Mark 2.00 out of 2.00

Match the following terms with their definitions:

Big Data

A broad term for data sets so large or complex that traditional data processing applications are inadequate.



Machine Learning

Explores the study and construction of algorithms that can learn from and make predictions on data.



Artificial Intelligence

The field which studies how to create computers and computer software capable of intelligent behaviour.



Data Science

An interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data.



The correct answer is: Big Data → A broad term for data sets so large or complex that traditional data processing applications are inadequate., Machine Learning → Explores the study and construction of algorithms that can learn from and make predictions on data., Artificial Intelligence → The field which studies how to create computers and computer software capable of intelligent behaviour., Data Science → An interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data.

Question 10

Correct

Mark 1.00 out of 1.00

The "naïve" assumption made by the Naïve Bayes classifier is that:

Select one:

- ☐ a. The dataset does not contain any noise.
- ☐ b. The prior probability of each class is equal.
- ☒ c. All features are conditionally independent given the class. ✓
- ☐ d. All features are continuous.

The correct answer is: All features are conditionally independent given the class.

Question 11

Correct

Mark 1.00 out of 1.00

What is the primary purpose of using Laplace smoothing in Naïve Bayes?

Select one:

- ☐ a. To simplify the model by removing features with low counts.
- ☐ b. To increase the probability estimates for frequently occurring features.
- ☒ c. To prevent zero probability estimates. ✓
- ☐ d. To ensure all features contribute equally to the final classification.

Your answer is correct.

The correct answer is: To prevent zero probability estimates.

Question 12

Correct

Mark 2.00 out of 2.00

Consider the standard formulation of Bayes' Theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Label each term:

- $P(X|Y)$ Likelihood ✓
- $P(Y|X)$ Posterior ✓
- $P(Y)$ Prior ✓
- $P(X)$ Normalisation Term ✓

The correct answer is: $P(X|Y) \rightarrow$ Likelihood, $P(Y|X) \rightarrow$ Posterior, $P(Y) \rightarrow$ Prior, $P(X) \rightarrow$ Normalisation Term

Information**Naive Bayes Formulae****Naive Bayes**

$$P(c|x) = \frac{\prod_{i=1}^n P(x_i|c)P(c)}{P(x)}$$

$$\text{where } P(x) = \sum_{y'} P(x|y')P(y')$$

Laplace Smoothing

$$P(x_i, c) = \frac{\text{Count}(x_i, c) + k}{\text{Count}(c) + k \times n_{x_i}}$$

Where:

- $\text{Count}(x_i, c)$ is the number of times x_i and c have jointly occurred
- $\text{Count}(c)$ is the number of times c has occurred
- n_{x_i} is the number of distinct possible values that x_i can take

Information

Table 1: Dataset for Q13 - Q16

For Q13 - Q16 consider the following dataset for predicting if a user will 'Like' ('Yes'/'No') a song based on its 'Genre' and 'Tempo'.)

Genre	Tempo	Like
Pop	Fast	Yes
Rock	Slow	No
Pop	Slow	Yes
Classical	Slow	No
Rock	Fast	Yes
Pop	Fast	Yes
Rock	Slow	No
Classical	Fast	No

Question 13

Correct

Mark 2.00 out of 2.00

Based only on the dataset provided in Table 1, what is $P(\text{Like} = \text{"Yes"})$?

Answer: 

The correct answer is: 0.5

Question 14

Correct

Mark 2.00 out of 2.00

Based only on the dataset provided in Table 1, what is $P(\text{Genre} = \text{"Pop"} | \text{Like} = \text{"Yes"})$?

Answer: 

Answer for without smoothing

The correct answer is: 0.75

Question 15.1

Correct

Mark 2.00 out of 2.00

Using only the dataset provided in Table 1, calculate $P(\text{Like} = \text{"Yes"} | \text{Genre} = \text{"Pop"}, \text{Tempo} = \text{"Fast"})$. Assume conditional independence.

Answer: 

The correct answer is: 1

Question 15.2

Correct

Mark 1.00 out of 1.00

Using only the dataset provided in Table 1, calculate $P(\text{Like}=\text{'No'} \mid \text{Genre}=\text{'Pop'}, \text{Tempo}=\text{'Fast'})$. Assume conditional independence.

Answer: 

The correct answer is: 0


Question 15.3

Correct

Mark 1.00 out of 1.00

Based on your calculations in 15.1 and 15.2, what is the most probable class (MAP prediction) for 'Like'?

Select one:

- ☐ a. No
- ☒ b. Yes 

The correct answer is: Yes

Question 16

Correct

Mark 3.00 out of 3.00

A new song has (Genre='Jazz', Tempo='Fast'). Using the dataset from Table 1 and applying Laplace smoothing with $k=1$ for all features and classes, calculate the smoothed probability $P(\text{Genre}=\text{'Jazz'} \mid \text{Like}=\text{'No'})$.

Answer: 

The correct answer is: 0.125

Information

A Gaussian Naïve Bayes classifier is used to predict patient outcome ('Recovered'/'Not Recovered') based only on 'Temperature'. For patients who 'Recovered', the recorded temperatures were {36.5, 37.0, 37.5, 36.0}. Calculate the sample mean ($\mu_{x,c}$) and sample variance ($\sigma_{x,c}^2$) for $P(\text{Temperature} \mid \text{Outcome} = \text{'Recovered'})$ based only on this data. x = "Temperature" and c = 'Recovered'.

Use the equations:

$$\mu = \frac{1}{n} \sum_{i=1}^n x^{(i)}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu)^2$$

Provide your answer for mean in Question 17.1 and your answer for variance in Question 17.2.

Question 17.1

Correct

Mark 2.00 out of 2.00

Provide the numerical value of $\mu_{x,c}$ (rounded to 3 decimals) as per Question 17.

Answer: 36.75



The correct answer is: 36.75

Question 17.2

Incorrect

Mark 0.00 out of 2.00

Provide the numerical value of $\sigma_{x,c}^2$ (rounded to 3 decimals) as per Question 17.

Answer: 0.3125



The correct answer is: 0.313

Question 18

Complete

Not graded

A Gaussian Naïve Bayes classifier is used to classify plants into class 'a' or class 'b' based on the features height (x_h) and weight (x_w). Assume height and weight are conditionally independent given the class and are modelled by Gaussian distributions. The following parameters were learned from training data:

- Class 'a'
 - $P(C = a) = 0.6$
 - $height \sim N(\mu_{h,a}, \sigma_{h,a}^2)$ where $\mu_{h,a} = 10, \sigma_{h,a}^2 = 4$
 - $width \sim N(\mu_{w,a}, \sigma_{w,a}^2)$ where $\mu_{w,a} = 3, \sigma_{w,a}^2 = 1$
- Class 'b'
 - $P(C = b) = 0.4$
 - $height \sim N(\mu_{h,b}, \sigma_{h,b}^2)$ where $\mu_{h,b} = 12, \sigma_{h,b}^2 = 4$
 - $width \sim N(\mu_{w,b}, \sigma_{w,b}^2)$ where $\mu_{w,b} = 5, \sigma_{w,b}^2 = 1$

A new plant is observed with $x_h = 11$ and $x_w = 4$. Calculate $P(C = a | x_h = 11, x_w = 4)$. Provide the numerical value (to 3 decimal places) and use 5 decimals during your working out.

Use the following in your calculations:

$$p(x_i | c) = \frac{1}{\sqrt{2\pi\sigma_{x_i,c}^2}} \exp\left\{-\frac{1}{2}\left(\frac{(x_i - \mu_{x_i,c})^2}{\sigma_{x_i,c}^2}\right)\right\}$$

Answer: 2.943×10^{-6}

The correct answer is: 0.6

Information

Decision Trees Formulae**Entropy**

$$H(p) = - \sum_{i=1}^n p_i \log_2 p_i \text{ where } p = \{p_1, \dots, p_n\}$$

Information Gain

$$\text{Gain}(D, F) = H(D) - \frac{1}{|D|} \sum_{f \in \text{values of } F} |D_f| H(D_f)$$

Question 19

Correct

Mark 1.00 out of 1.00

The primary goal of the ID3 algorithm when selecting the best feature to split on at a node is to:

Select one:

- ☐ a. Maximise the number of branches from the node.
- ☒ b. Maximise the information gain. ✓
- ☐ c. Ensure all leaf nodes have the same number of instances.
- ☐ d. Minimise the depth of the tree.

The correct answer is: Maximise the information gain.

Question 20

Correct

Mark 1.00 out of 1.00

Following a single path from the root to a leaf in a decision tree represents a sequence of logical OR conditions.

Select one:

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 21

Correct

Mark 1.00 out of 1.00

What is the main purpose of pruning a decision tree?

Select one:

- ☐ a. To make the tree easier to visualise.
- ☐ b. To handle missing values in the dataset.
- ☒ c. To reduce the complexity of the tree. ✓
- ☐ d. To increase the tree's performance on the training data.

The correct answer is: To reduce the complexity of the tree.

Question 22

Correct

Mark 1.00 out of 1.00

Maximum entropy for a subset of a dataset implies that all datapoints in that subset share the same class, i.e., the subset is perfectly classified.

Select one:

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 23

Correct

Mark 1.00 out of 1.00

How does the ID3 algorithm (or similar decision tree algorithms) typically handle continuous-valued features?

Select one:

- ☐ a. It uses the continuous value directly as a branch.
- ☐ b. It ignores them as it can only process discrete features.
- ☐ c. It calculates the average value for the feature and uses that.
- ☒ d. It converts them to discrete features by defining potential split points (thresholds). ✓

The correct answer is: It converts them to discrete features by defining potential split points (thresholds).

Question 24

Correct

Mark 1.00 out of 1.00

When building a regression tree (predicting continuous values), information gain based on entropy is the standard metric used to evaluate potential splits.

Select one:

- ☐ True
- ☒ False ✓

The correct answer is 'False'.

Question 25

Correct

Mark 3.00 out of 3.00

A dataset contains 10 instances belonging to two classes. 6 instances are Class 'Positive' and 4 instances are Class 'Negative'. Calculate the Entropy of this dataset (use log base 2).

Answer: ✓

The correct answer is: 0.971

Question 26

Correct

Mark 4.00 out of 4.00

A dataset D has an initial Entropy $H(D) = 0.990$. Splitting on Feature 'F' results in two subsets: Subset D_1 (8 instances, $H(D_1) = 0.811$) and Subset D_2 (12 instances, $H(D_2) = 0.650$). Calculate the Information Gain of splitting on Feature 'F'.

Answer: ✓

The correct answer is: 0.276

Question 27

Correct

Mark 3.00 out of 3.00

In a regression tree designed to predict house prices (in thousands of Rands), a specific leaf node contains training instances with the following target prices: {350, 400, 380, 420}. What price would this leaf node predict for a new instance that reaches it? (Do not give units)

Answer: ✓

The correct answer is: 387.5

Question 28

Correct

Mark 3.00 out of 3.00

Consider a dataset with a continuous feature 'Age' and a binary class ('Yes'/'No'). The initial Entropy is $H(D) = 0.971$ and the dataset is given below:

Age	Class
25	Yes
30	Yes
35	No
40	No

Evaluate a potential split point at Age < 32.5. Calculate the Information Gain for this specific split.

Answer: 

The correct answer is: 0.971


Question 32

Correct

Mark 1.00 out of 1.00

Which of the following could typically not be used directly as a basis function $\phi_i(x)$ within the linear regression model formulation $f(x, \theta) = \sum_{i=0}^k \theta_i \phi_i(x)$? (Assume x is one dimensional).

Select one:

- ☐ a. $\phi_i(x) = e^{-x^2}$
- ☐ b. $\phi_i(x) = x^3$
- ☒ c. $\phi_i(x) = \theta_0 + \theta_1 x$ 
- ☐ d. $\phi_i(x) = \sin(x)$

Your answer is correct.


The correct answer is: $\phi_i(x) = \theta_0 + \theta_1 x$

Question 30

Correct

Mark 1.00 out of 1.00

Adding the regularisation term $\lambda \sum_{j=1}^d \theta_j^2$ (for $j = 1 \dots d, \lambda > 0$) to the linear regression cost function primarily aims to reduce the model's variance, potentially at the cost of increased bias.

- ☒ True 
- ☐ False

The correct answer is 'True'.

Question 31

Correct

Mark 1.00 out of 1.00

In gradient descent for linear regression, if the learning rate (α) is set too small, what is the most likely outcome?

- ☐ a. The algorithm might overshoot the minimum and fail to converge.
- ☒ b. Convergence will likely be very slow. ✓
- ☐ c. The cost function $J(\theta)$ will increase rapidly.
- ☐ d. The model parameters (θ) will become zero.

The correct answer is: Convergence will likely be very slow.

Question 32

Correct

Mark 2.00 out of 2.00

Given a single data point ($x=3, y=8$) and a simple linear model $f(x, \theta) = \theta_0 + \theta_1 x$ with current parameters $\theta = (2, 1.5)$. Calculate the squared error $(f(x, \theta) - y)^2$ for this single point.

Answer: ✓

The correct answer is: 2.25

Question 33

Correct

Mark 4.00 out of 4.00

Perform one step of gradient descent for parameter θ_1 only, given the following:

- Data point $(x^{(i)} = 2, y^{(i)} = 9)$
 - Current model $f(x, \theta) = \theta_0 + \theta_1 x$ with $\theta_0 = 1, \theta_1 = 3$
 - Learning rate $\alpha = 0.1$
 - Use the update rule: $\theta_j \leftarrow \theta_j - \alpha(f(x^{(i)}, \theta) - y^{(i)})x^{(i)}$ (for $j = 1$)
- Calculate the updated value for θ_1 .

Answer: ✓

The correct answer is: 3.4

Information

Consider the dataset: Data point 1: ($x = 1, y = 2$), Data point 2: ($x = 2, y = 4$).

Calculate the optimal parameter θ using the closed-form solution with regularisation $\theta = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}')^{-1} \mathbf{X}^T \mathbf{y}$ for the linear model $f(x, \theta) = \theta_0 + \theta_1 x$, given $\lambda = 1$. (\mathbf{I}' is the identity matrix with 0 at the top-left for the bias term).

Note: If $\mathbf{A} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$, then $\mathbf{A}^{-1} = \frac{1}{a_1 a_4 - a_2 a_3} \begin{pmatrix} a_4 & -a_2 \\ -a_3 & a_1 \end{pmatrix}$

The answer for θ_0 must be provided in Question 34.1 and the answer for θ_1 must be provided in Question 34.2.

N.B. For this question do not round during your working, and rather use fractions. Only round the final answer (to 3 decimal places).

Question 34.1

Correct

Mark 3.00 out of 3.00

Provide the numerical answer for θ_0 as per Question 34.

Answer: 

The correct answer is: 2

Question 34.2

Correct

Mark 3.00 out of 3.00

Provide the numerical answer for θ_1 as per Question 34.

Answer: 

The correct answer is: 0.667

Question 35

Incorrect

Mark 0.00 out of 3.00

Given a linear regression model $f(x, \theta) = \theta_0 + \theta_1 x$ with parameters $\theta = [\theta_0, \theta_1]^T = [1.0, 2.0]^T$. Calculate the value of the regularised cost function

$$E(\theta) = \frac{1}{2} \sum_i (f(x^{(i)}, \theta) - y^{(i)})^2 + \lambda \sum_{j=1}^d \theta_j^2$$

for the dataset: Data point 1: ($x = 1, y = 4$), Data point 2: ($x = 2, y = 6$), using regularisation parameter $\lambda = 2$.

Answer: 

The correct answer is: 9

