Exercises

1	2	3	4	5	6	7	8

Surname, First name

____

Introduction Machine Learning (8BB020)
Exam 8BB020 Introduction Machine Learning

○1	○1	○1	○1	○1	○1	○1
○2	○2	○2	○2	○2	○2	○2
○3	○3	○3	○3	○3	○3	○3
○4	○4	○4	○4	○4	○4	○4
○5	○5	○5	○5	○5	○5	○5
○6	○6	○6	○6	○6	○6	○6
○7	○7	○7	○7	○7	○7	○7
○8	○8	○8	○8	○8	○8	○8
○9	○9	○9	○9	○9	○9	○9
○0	○0	○0	○0	○0	○0	○0

○a ⊗b ○c ○d ○e ○f → b
○a ○b ⊗c ○d ○e ○f → c
● ○b ○c ⊗d ○e ○f → a

□a ☒b □c ☒d □e □f → b, d
□a □b ☒c □d □e □f → c
■ □b □c ■ ☒e □f → a, d

Answer multiple-choice questions as shown in the example. Circular checkboxes have only one correct answer. Square checkboxes may have multiple correct answers.

Particular Ans on paper exam instructions

- Write in a black or blue pen.
- You answer open-ended questions by using the text box. Provide your answers on the papers inside the answer box underneath a question. ***If you need more space for your answers, use the extra space at the end of the exam, and clearly indicate there which question you continue answering. In the text box of the particular question, clearly state that you proceed with your answer on a different page.***
- Hand in all pages. Do not remove the staple. If you remove it anyhow, check that you hand in all pages.

Dear student,

You're about to take an exam. Write down your name and your student ID at the appropriate places above. Make sure that you enter your student ID by fully coloring the appropriate boxes. On the examination attendance card, you fill in the PDF number. You can find the correct number on the top of the first page of your exam (e.g. 1234.pdf).

Please read the following information carefully:

Date exam:
Start time 09.00
End time: 12.00 (+30 minutes for time extension students)

Number of questions:
Maximum number of points/distribution of points over questions: 32 multiple choice questions (total of

35 points) + 7 open questions (total 35 points) = 70 points.
Method of determining the final grade: number of points / 70 * 10.
Answering style: formulation, foundation of arguments, multiple choice.

Permitted examination aids

- Scrap paper (fully blank)

Important:

- You are only permitted to visit the toilets under supervision
- Examination scripts (fully completed examination paper, stating name, student number, etc.) must always be handed in
- The house rules must be observed during the examination
- The instructions of subject experts and invigilators must be followed
- Keep your work place as clean as possible: put pencil case and breadbox away, limit snacks and drinks
- You are not permitted to share examination aids or lend them to each other
- Do not communicate with any other person by any means

During written examinations, the following actions will in any case be deemed to constitute fraud or attempted fraud:

- using another person’s proof of identity/campus card (student identity card)
- having a mobile telephone or any other type of media-carrying device on your desk or in your clothes
- using, or attempting to use, unauthorized resources and aids, such as the internet, a mobile telephone, smartwatch, smart glasses etc.
- having any paper at hand other than that provided by TU/e, unless stated otherwise
- copying (in any form)
- visiting the toilet (or going outside) without permission or supervision

First-year bachelor students: The final grade for this exam will be announced no later than fifteen working days after the date of this exam, unless this exam takes place in Q4 or the interim period. For Q4 final exams, grades will be announced within five working days after the end of the Q4 final test period. For interim period final exams, grades will be announced no later than five working days before September 1.
All other students: Generally, the final grade for this exam will be announced no later than fifteen working days after the date of this examination. Specifically for bachelor exams administered in the interim period, exam grades will be announced no later than five working days before September 1.

You can start the exam now, good luck!

Multiple choice questions

1p **1a** Which of the following are examples of nonparametric machine learning algorithms?

□ Logistic regression

□ Random forest

□ Linear regression

□ k-Nearest Neighbours

1p **1b** Which of the following are examples of unsupervised machine learning tasks?

□ Clustering

□ Regression

□ Dimensionality reduction

□ Classification

2p **1c** In a study, researchers aim to assess the extent to which blood pressure reduction in response to a certain drug can be predicted based on clinical data. For this purpose, they collected data for 150 patients with measurements for:

- Age (years)
- Weight (kg)
- Initial Blood Pressure (mmHg)
- Cholesterol Level (mg/dL)
- Heart Rate (bpm)
- Blood Pressure Reduction (mmHg)

They used 70% of the data for model training and keep 30% for model testing. What are the number of observations (n) and the number of input features (p) in the training dataset?

ⓐ $n = 105, p = 5$

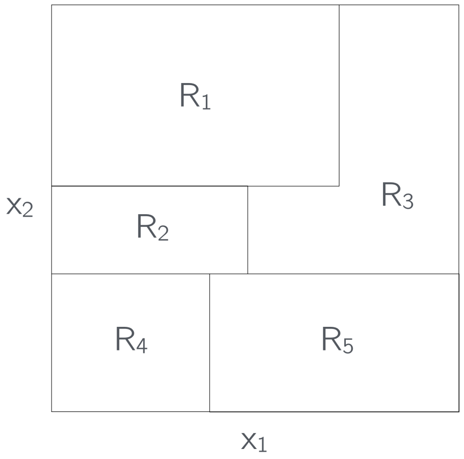
ⓑ $n = 150, p = 6$

ⓒ $n = 105, p = 6$

ⓓ $n = 150, p = 5$

1p **1d**

Consider the following partition of a two-dimensional feature space in five non-overlapping regions ($R_1, \ldots, R_5$).



Could this result from a recursive binary split using a decision tree?

ⓐ Yes ⓑ No

1p **1e** What is the vanishing gradient problem, and why does it occur in deep networks?

ⓐ It results from using too few hidden layers in a network

ⓑ It occurs when the gradients become too large, leading to unstable training

ⓒ It happens when the gradients become too small, making it difficult for the network to update weights, especially in early layers

ⓓ It happens when all neurons in a layer are dropped during dropout

1p **1f** Which activation function is commonly used to avoid the vanishing gradient problem?

ⓐ Softmax

ⓑ ReLU

ⓒ Sigmoid

ⓓ Tanh

1p **1g** What is the main purpose of model averaging in neural networks?

ⓐ To reduce the risk of overfitting by combining multiple models

ⓑ To decrease training time by using fewer parameters

ⓒ To increase the depth of the network

ⓓ To improve the network’s learning rate

1p **1h** Consider the following statement: "I already know the neural network architecture I want to use, I do not not need a validation set". Is the following statement:

ⓐ Cannot be determined if True or False

ⓑ True

ⓒ False

1p **1i** Which regularization technique adds noise to the network during training but is removed at test time?

ⓐ L2 regularization

ⓑ Dropout

ⓒ L1 regularization

ⓓ Weight decay

1p **1j** Why is Stochastic Gradient Descent (SGD) preferred for training large datasets?

ⓐ It requires fewer iterations to converge compared to other gradient descent methods

ⓑ It uses smaller batches or individual data points for gradient estimation, making it computationally feasible for large datasets

ⓒ It uses the entire dataset at each iteration to compute precise gradients

ⓓ It prevents the vanishing gradient problem by using larger batch sizes

1p **1k** What is the primary purpose of adding momentum to the gradient descent optimization process?

- ⓐ To slow down the updates when the loss is decreasing too quickly
- ⓑ To prevent overfitting by adding noise to the updates
- ⓒ To make updates based on a moving average of past gradients, helping to accelerate convergence in the relevant direction
- ⓓ To increase the learning rate dynamically during training

1p **1l** What is the main role of backpropagation in training a neural network?

- ⓐ To compute the gradients of the loss with respect to the model parameters, enabling weight updates
- ⓑ To apply regularization during training
- ⓒ To adjust the learning rate at each layer
- ⓓ To compute the forward pass of the network

1p **1m** Why is the chain rule of differentiation essential for backpropagation in neural networks?

- ⓐ It allows the computation of the output of the network
- ⓑ It reduces the risk of overfitting
- ⓒ It normalizes the inputs to each layer
- ⓓ It enables the computation of gradients layer-by-layer in a computationally efficient manner

1p **1n** What is a common sign of overfitting in neural networks during training?

- ⓐ Both the training and validation losses increase simultaneously
- ⓑ The training loss increases while the validation loss decreases
- ⓒ The validation loss decreases
- ⓓ The training loss decreases but the validation loss increases

1p **1o** In a support vector classifier (SVC), what is the effect of increasing the parameter C, defined as the total amount of slack (i.e., the budget allowed for misclassification), on the number of support vectors?

ⓐ Tuning C does not affect the number of support vectors.

ⓑ The number of support vectors decreases with increasing C.

ⓒ The number of support vectors increases with increasing C.

1p **1p** Imagine you have a dataset of 60 patients, where you measure four variables: cholesterol, BMI, blood pressure, and heart rate. You apply principal component analysis (PCA) to this dataset. What is the length of the principal component scores vector and the loading vectors for each principal component?

ⓐ The principal component scores vector has length 60, and the loading vectors have length 4.

ⓑ Both the principal component scores vector and the loading vectors have length 60.

ⓒ Both the principal component scores vector and the loading vectors have length 4.

ⓓ The principal component scores vector has length 4, and the loading vectors have length 60.

1p **1q** In Ridge regression and the Lasso applied to linear models, is the intercept regularized?

ⓐ No, in neither of them.

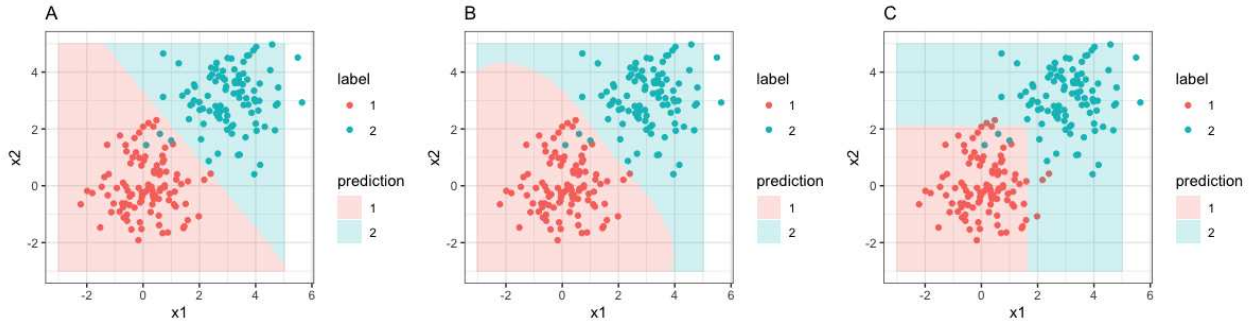
ⓑ Yes, but only in Ridge regression.

ⓒ Yes, but only in the Lasso

ⓓ Yes in both.

1p **1r**

Below are three plots showing decision boundaries for a binary classification problem. For which plot(s) could the decision boundary be created using support vector machines (SVM)?



□ Panel A

□ Panel C

□ Panel B

1p **1s**

Consider the same three plots as in question 1r, showing decision boundaries for a binary classification problem. For which plot(s) could the decision boundary be created using a decision tree?

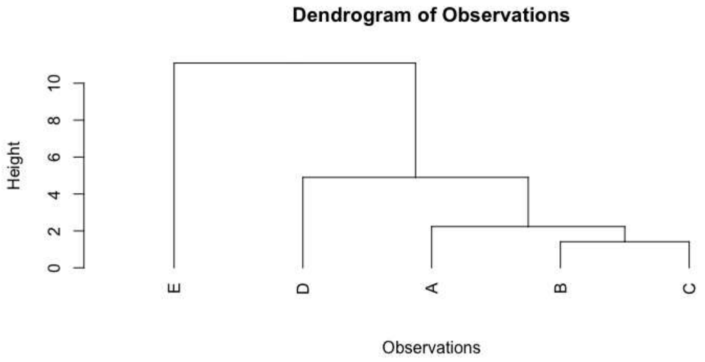
□ Panel A

□ Panel C

□ Panel B

1p **1t**

Below is a dendrogram created from a dataset containing five observations (A, B, C, D, E). Examine the dendrogram and determine which of the following pairs of observations is the most similar based on the clustering.



ⓐ Observations A and D

ⓑ Observations E and D

ⓒ Observations A and C

ⓓ Observations C and D

1p **1u** Consider the same dendogram as in question 1t, if we cut the dendogram in 3 clusters, which observations would fall in the same cluster?

□ Observations A and C

□ Observations A and D

□ Observations E and D

□ Observations A and B

1p **1v** Which of the following methods we have seen can be used for regression?

□ Linear regression

□ Clustering

□ Logistic regression

□ Random forest

□ Neural networks

□ Boosting

□ Support vector machines

1p **1w** Which of the following methods we have seen can be used for classification?

□ Clustering

□ Support vector machines

□ Neural networks

□ Logistic regression

□ Tree-based methods

□ Linear regression

1p **1x** In hierarchical clustering, which of the following statements correctly describes the single linkage method?

ⓐ The distance between two observations within a cluster is used to measure the overall dissimilarity of the cluster.

ⓑ The distance between two clusters is defined as the minimum distance between the two centroids of the clusters.

ⓒ The distance between two clusters is defined as the distance between the two furthest observations, one from each cluster.

ⓓ The distance between two clusters is defined as the minimum distance between any two observations, one from each cluster.

1p **1y** In K-means clustering, which of the following statements is false?

ⓐ K-means clustering partitions the data into K clusters, each represented by the centroid of the cluster.

ⓑ The algorithm minimises the sum of squared distances between each observation and the nearest cluster centroid.

ⓒ The number of clusters K must be pre-specified before applying the K-means algorithm.

ⓓ K-means clustering always guarantees finding the globally optimal solution.

1p **1z** We are using a linear regression model to predict lung capacity (in liters) based on two input features: height and weight. The linear model is given by the following equation:

Lung capacity = $\beta_0 + \beta_1 \times$Height + $\beta_2 \times$Weight

Which of the following statements about the linear model are correct?

□ The model can capture interactions between height and weight without additional terms.

□ The coefficient β_2 represents the change in predicted lung capacity for each unit increase in weight, holding height constant.

□ The model assumes that the relationship between height and lung capacity is linear.

□ The model assumes that the relationship between height and weight is linear.

2p **1aa** In the context of training a linear regression model, which of the following statements about gradient descent are true?

□ The learning rate determines the size of the steps taken towards the minimum of the loss function.

□ Gradient descent may overshoot the minimum if the learning rate is too high.

□ Gradient descent guarantees convergence to the global minimum for any loss function.

□ The gradient of the loss function indicates the direction to adjust the parameters to minimise the loss.

□ Gradient descent can only be applied to convex loss functions.

□ Gradient descent iteratively updates the model parameters to minimise the loss function.

1p **1ab** Which of the following biomedical engineering problems are classification tasks?

□ Predicting whether a patient will develop heart disease based on cholesterol levels, age, and blood pressure.

□ Identifying whether a patient has diabetes based on fasting glucose levels and BMI.

□ Diagnosing whether a tumour is malignant or benign based on imaging data.

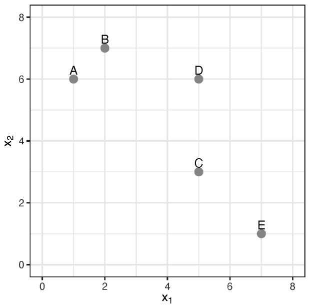
□ Estimating a patient's blood glucose level based on dietary habits and insulin dosage.

□ Predicting a patient's risk category (low, medium, high) for stroke based on family history and lifestyle factors.

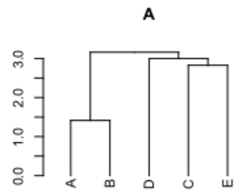
□ Predicting the percentage reduction in tumour size after chemotherapy based on treatment dosage and patient characteristics.

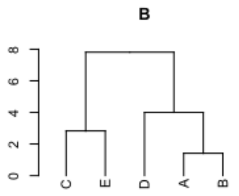
2p **1ac**

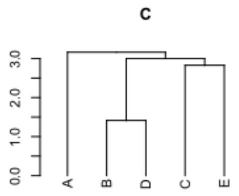
You are given the following small dataset with two variables (x_1 and x_2) for 5 observations:

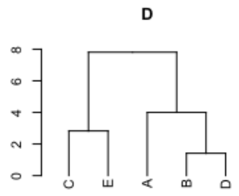


Which of the following dendograms would result by using Euclidean distance as a measure of dissimilarity between observations and applying complete linkage?









ⓐ Plot C

ⓑ Plot A

ⓒ Plot B

ⓓ Plot D

1p **1ad** What is the role of the slack variables ϵ_i in a soft-margin support vector classifier (SVC)?

ⓐ It allows some misclassifications by penalising points on the wrong side of the margin.

ⓑ It reduces the margin width for better classification accuracy.

ⓒ It ensures that all points lie exactly on the margin.

ⓓ It penalises points that are on the correct side of the margin.

1p **1ae** Which of the following statements are true about the support vectors in a support vector machine (SVM)?

□ Support vectors are the points that do not influence the position of the decision boundary.

□ The support vectors are the points that lie directly on the margins or on the wrong side of the margin.

□ The support vectors are the only points that determine the orientation of the hyperplane

□ The support vectors are all the points that lie outside of the margins.

1p **1af** Which of the following statements best describes Elastic Net regularisation?

ⓐ It combines L1 regularisation and L2 regularisation to handle multicollinearity and select variables.

ⓑ It combines L1 and L2 regularisation but does not help with feature selection or multicollinearity issues.

ⓒ It uses only L1 regularisation to perform feature selection by shrinking some coefficients to exactly zero.

ⓓ It applies only L2 regularisation to prevent overfitting by shrinking all coefficients towards zero.

k-Nearest Neighbours (k-NN))

2.5p **2a** What is the k-Nearest Neighbors (k-NN) algorithm, and how does it work? Describe the steps involved in classifying a new data point using k-NN.

2.5p **2b** Explain the impact of the value of ’k’ in k-NN. What happens when you choose very small or very large values of ’k’? How do you select an appropriate value of ’k’?

Neural networks

1.5p **3a** Describe and sketch how logistic regression can be expanded into a neural network for solving the XOR classification problem.

1.5p **3b** Describe the concept of "learned features" in deep learning. How does this concept differ from traditional feature engineering?

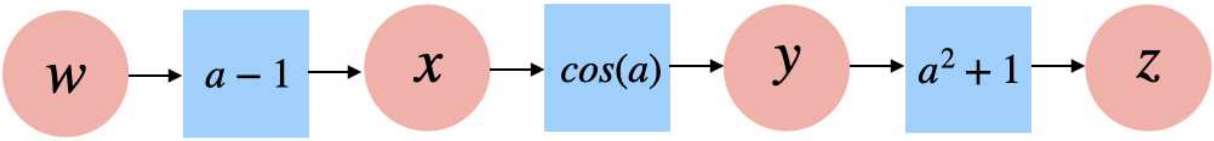
⤷

1.5p **3c** What would happen if all non-linearities (except for the output) were removed from a neural network? Explain the impact on the network's learning capabilities.

1.5p **3d** Explain how dropout functions as a regularization technique. Why does dropout reduce overfitting in neural networks?

4p **3e**

Consider the computational graph:



The flow of operations is defined as follows:

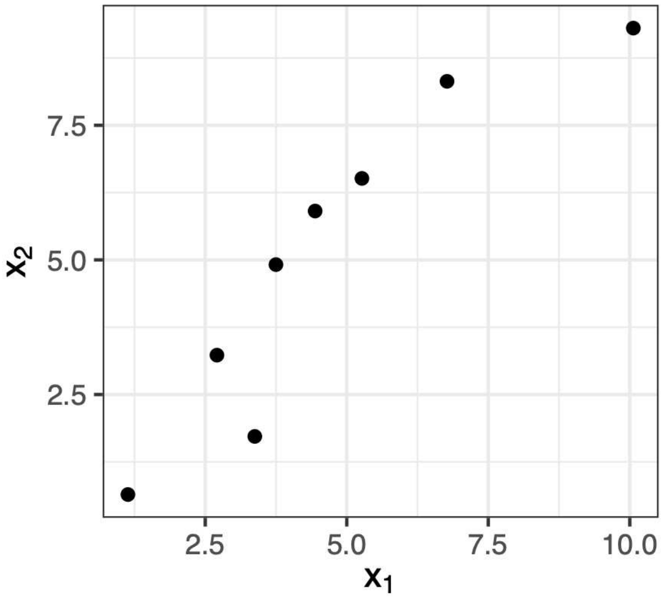
$$x = w - 1$$
$$y = \cos(x)$$
$$z = y^2 + 1$$

For the input $w = 1$, compute the value of z in the forward pass and then use backpropagation to compute the value of the derivative $\frac{dz}{dw}$.

⤵

Principal component analysis

Consider the following 2D scatter plot of 8 observations, showing two variables (x_1 and x_2).



2p **4a** Sketch the first and second principal components on the plot as lines. Clearly label each line. You can do it directly on the plot above, use the space below only if needed.

1p **4b** Mark the principal component scores, which are the projections of all points onto the first principal component, by drawing dashed lines from each original point to its corresponding projection. You can do this directly on the plot above; use the space below only if needed.

2p **4c** Do you expect the first and second principal components to explain a similar amount of variance in this example? Explain your reasoning, considering the meaning of the principal components and their relationship to the original variables.

Logistic regression

You are given the task to develop a model to support the diagnosis of cardiovascular disease using a biomarker measured in blood. It is expected that a value of the biomarker below 100 is linked to healthy condition and above 100 to cardiovascular disease.

3p **5a** You have two logistic regression models M1 and M2:
M1 with parameters ß$_0$ = -100, ß$_1$= 1
M2 with parameters ß$_0$ = -90, ß$_1$= 1

you want to validate the model on a test set composed of three patients with the value of the biomarker measured equal to 85, 95 and 120 respectively. Compute the predictive outcome of the two models for each patient using the logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

and determine which model is most suitable for the described classification task shortly motivating your answer.

↳

2p **5b** Briefly explain why a linear regression model with the same parameters ß$_0$ and ß$_1$ would not work for this task. You can make the drawing of the resulting model to support your explanation.

Regularised linear models

Imagine you have a clinical dataset of patients suspected of lung cancer. For each patient, the following variables were recoded during a check-up visit:

x_1 Number of cigarettes smoked per day
x_2 Age of the patient
x_3 Insulin level
x_4 Number of DNA mutations

For the same patients, you also have the final diagnosis: lung cancer or no lung cancer.

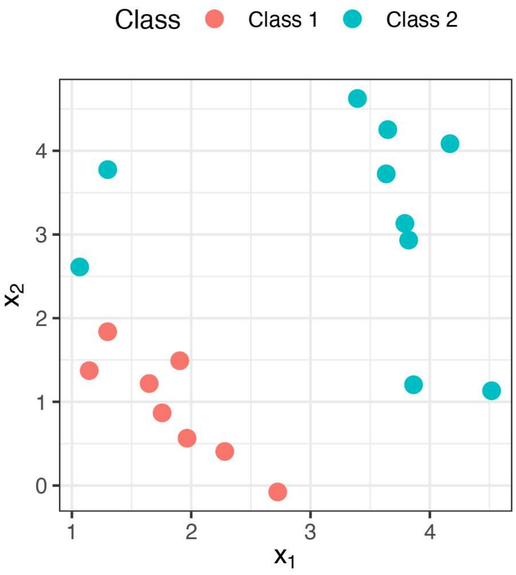
You are tasked with developing a model to support the diagnosis of lung cancer. A clinician tells you that they expect the most predictive feature to be the number of DNA mutations, followed by the number of cigarettes smoked per day and the age of the patient. Insulin level is not expected to be predictive.

2p **6a** Draw the expected profile of the Lasso coefficients (β_1 for x_1, β_2 for x_2, β_3 for x_3 and β_4 for x_4) as the regularisation parameter λ increases. The plot should have λ on the x-axis and the coefficient values on the y-axes. The y-axes does not need to have a specific scale, it is just a sketch to show the trend. Briefly explain your plot reflecting on the clinical expectations about the importance of each feature.

3p **6b** You are training your model using cross-validation. What do you expect the prediction error to be for increasing values of the regularisation parameter λ for both the training set and the validation set? Plot a sketch of the expected trends with λ on the x-axis and the prediction error on the y-axis and briefly explain why the error behaves this way. In your explanation refer to model complexity and bias and variance trade-off.

Decision tree

Consider the dataset depicted in the following figure.



1p **7a** Briefly explain why a decision tree is better suited than a support vector classifier for discriminating the two classes in this example. Your answer should discuss aspects such as the nature of the data distribution, model complexity, and interpretability.

2p **7b** Build a decision tree that uses the misclassification error the measure of node impurity for the cost function. Build the full tree without considering any termination criteria. Draw the resulting decision tree here below, including the approximate splitting conditions and the final classification for each leaf.

2p **7c** Draw the partition of the feature space corresponding to the decision tree you have build in the previous point directly in the figure above (use the space below only if needed).

Extra space

8 This is just for some extra empty space if needed.

↳

↷