Task 1

Multiple Topic Questions

Reply to each of the following questions.

a) List two categories to group a classifier.

gene, des
scor, prob.
cont, dis.

- Option A: Continuous vs discrete output.
- Option B: Scoring or probabilistic classifiers.
- Option C: Generative or discriminative.

b) What is a major disadvantage of the empirical risk minimization (ERM)?

- Over-fitting to the training set.

c) What is the difference of the lasso regression from the standard linear regression?

- It is a variation of linear regression with L1-regularization.

d) What is the Normal Equation approach for linear regression? A single sentence explanation is sufficient.

- It is the analytical approach / closed-form solution to linear regression.

e) Consider the target value y and the linear model $f(x)$ with x as input. Given m training samples, define the mean squared error loss function.

- Option A: $\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$.
- Option B: $\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2$.

f) Ensemble learning aims to modify the model variance and bias to improve the overall performance. Does it aim to increase or reduce the variance? Does it aim to increase or reduce the bias?

2 P.

- Reduce variance.
- Increase bias.

g) What type of ensemble learning approach is a random forest?

1 P.

- A bagging approach are the random forests.

h) What do you call the approach of ensemble learning, where a number of weak classifiers are combined to produce a powerful model?

1 P.

- Boosting.

i) What type of error does the generalisation error of a machine learning algorithm refer to?

1 P.

- Test error.

j) We can represent a neural network as a computational graph. What do the graph nodes and edges express?

2 P.

- The nodes correspond to operations.
- The edges correspond to the data flow.

k) We define the forward pass and backward pass for a computational graph. What do they express?

2 P.

- Forward pass: It refers to evaluating the function for a set of inputs, where the function describes the neural network. It is the process of making a prediction with a neural network.
- Backward pass: It calculates the gradients of all variables.

Actual

		T	F
pred	P	TP	FP
	N	FN	TN.

Task 2

/ 20 P.

Evaluation Metrics Classification

Consider a binary classifier that makes the following probabilistic predictions: 0.86, 0.65, 0.52, 0.28, 0.06, 0.95, 0.69, 0.59, 0.39, 0.15. The ground-truth values are: 1, 0, 0, 0, 0, 1, 1, 1, 0, 1. The positive class is noted with 1. Reply to the following questions:

a) Define what is a true positive (TP), a true negative (TN), a false positive (FP) and false negative (FN).

- True positive (TP): the number of correctly predicted samples.
- True negative (TN): the number of correctly predicted samples, as not the considered class.
- False positive (FP): the number of samples incorrectly predicted as the considered class.
- False negative (FN): the number of samples incorrectly not predicted as the considered class.

b) We have learned that the precision tells us how the classifier manages to avoid predicting a negative sample as positive. Also, the recall tell us whether the classifier can find all positive samples. Define the precision in terms of true positive (TP), true negative (TN), false positive (FP) and/or false negative (FN) predictions. Then, define the recall in the same way.

- Precision is the ratio tp / (tp + fp).
- Recall is the ratio tp / (tp + fn).

c) To convert the output of the classifier to binary, consider a threshold function with 0.5 as the threshold value. Then, convert the output of the binary classifier to 0-1 form and report the prediction 10 values.

5 P.

1, 1, 1, 0, 0, 1, 1, 1, 0, 0.

	0.86,	0.65,	0.52,	0.28,	0.06,	0.95,	0.69,	0.59,	0.39,	0.15.
~~pred~~	1 TP	1 FP	1 FP	0 TN	0 TN	1 TP	1 TP	1 TP	0 TN	0 FN.
act :	1	0	0	0	0	1	1	1	0	1

TP - 4
TN - 3

FP - 2
FN - 1.

		Act	
		T	F
Pred	P	TP	FP.
	N	FN	TN.

Continuation Task 2

d) Based on the binary output and the ground-truth values, compute the number of true positives, true negative, false positives, and false negatives. Also, compute the precision and recall.

6 P.

- tn,fp,fn,tp: 3, 2, 1, 4.
- precision: 0.667
- recall: 0.800

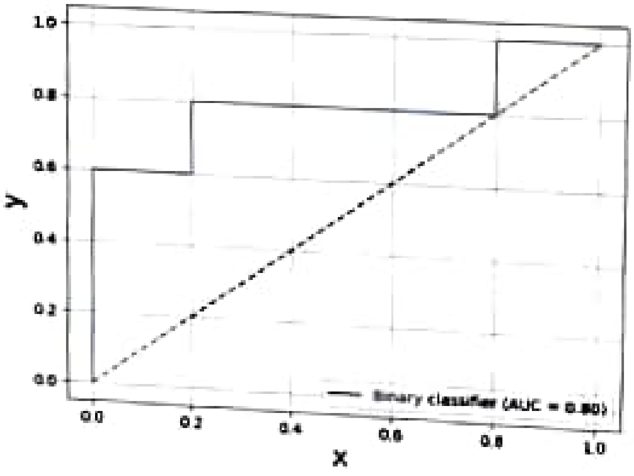
TN, FP, FN, TP.
3, 2, 1, 4.

$$pre = \frac{TP}{TP+FP} = \frac{4}{4+2} = \frac{4}{6} = \frac{2}{3} = 0.6$$

$$Rec = \frac{TP}{TP+FN} = \frac{4}{4+1} = \frac{4 \times 20}{5 \times 20} = \frac{80}{100} = 0.8$$

e) Consider the figure below showing the receiver operating characteristic (ROC) curve. What is does the x and y axis represent. What is the diagonal red dashed line?

3 P.



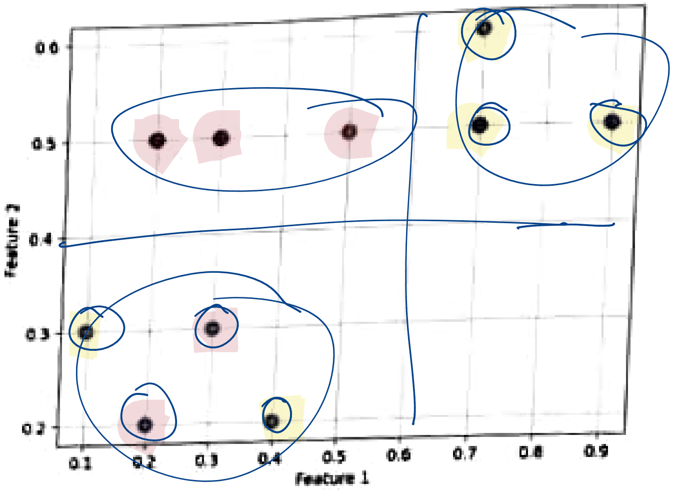
- x: False positive rate
- y: True positive rate
- diagonal: random classifier

Task 3

/ 20 P.

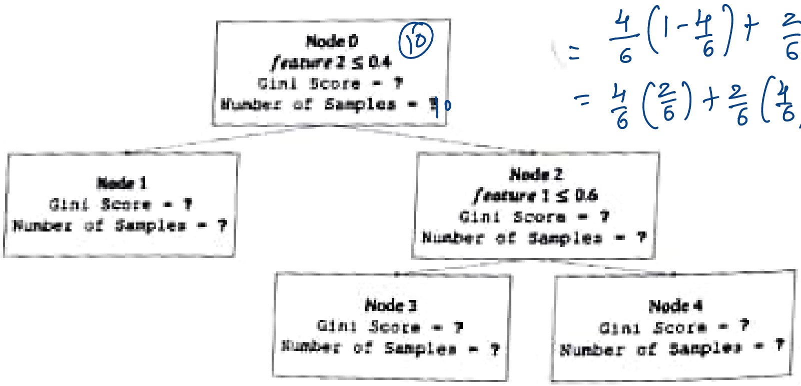
Decision Trees

In decision trees for classification, evaluating the possible split is based on the purity of a node. One purity measure is the Gini score that is defined as: $g(\mathbf{p}) = 1 - \sum_{i=1}^{k}(p_i)^2$ with $\mathbf{p} = \{p_1, \ldots, p_k\}$ and p_i the fraction of samples that correspond to class i of total k classes. The score is 0 if all samples are from the same class. Consider the following training set in the figure. There are 10 samples and $k = 2$ classes (red and blue) in total. Reply the following questions using the empty boxes.



a) The figure below shows a decision tree with two split functions. Given the training set of 10 samples, compute for each node the Gini score and the number of samples. You need to use the split functions that are given in the figure.

10 P.

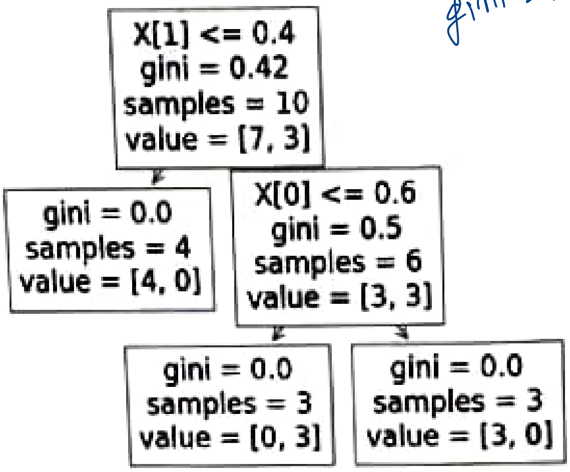


$$= \frac{4}{6}\left(1-\frac{4}{6}\right) + \frac{2}{6}\left(1-\frac{2}{6}\right)$$

$$= \frac{4}{6}\left(\frac{2}{6}\right) + \frac{2}{6}\left(\frac{4}{6}\right) = \frac{8}{36} + \frac{8}{36} = \frac{\overset{4\times4}{\cancel{16}}}{\underset{4\times9}{\cancel{36}}}$$

$$= \frac{4}{9}.$$

gini = Pk(1-Pk)



b) The leaf nodes of the tree can be used to perform classification of test samples. Report what is the most voted class (blue or red) for each leaf node of the above tree. You should use the node index for specifying each leaf node.

_____ 5 P.

Node 1: blue, Node 3: red, Node 4: blue.

c) Based on the constructed tree and samples at end nodes (leaf nodes) classify the following samples into blue and red class: (0.3, 0.2), (0.6, 0.3), (0.4, 0.6), (0.7, 0.4), (0.9, 0.6).

_____ 5 P.

- Class: blue, blue, red, blue, blue.

Task 4

/ 2

Neural Networks

Answer the following questions on neural networks.

a) We have three main gradient descent variants: batch gradient descent, stochast gradient descent, and mini-batch gradient descent. Define the parameter $\mathbf{w}$ updat equations given the learning rate η.

- Batch GD → It is defined as: $\mathbf{w} = \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}(f(\mathbf{x}; \mathbf{w}), \mathbf{y})$.
- Stochastic GD → It is defined as: $\mathbf{w} = \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}(f(x_i, \mathbf{w}), y_i))$.
- Mini-batch GD → It is defined as: $\mathbf{w} = \mathbf{w} - \eta \nabla_{\mathbf{w}} \mathcal{L}(f(\mathbf{x}_{(i:i+n)}, \mathbf{w}), \mathbf{y}_{(i:i+n)}))$.

b) Consider the input tensor of dimensions 10 × 10 × 10, where the last dimension is the channel dimension. The following operations are applied: ① 3 × 3 convolution (40 channels) with stride 1 and padding 1 for each dimension, ② ReLU activation, 3 × 3 max pooling with stride 1 and padding 1 for each dimension, ③ 3 × 3 convolution (20 channels) with stride 1 and padding 1 for each dimension, ④ ReLU activation, and ⑤ 2 × 2 max pooling with stride 2 and padding 1 for each dimension. What are the dimensions of the output tensor after every single operation? Report the name of the operation and the output dimensions as width x height x channels. There are 6 operations in total.

Input	10 × 10 × 10
After 3 × 3 conv.	10 × 10 × 40 ✓
After ReLU 1	10 × 10 × 40 ✓
After Max Pooling	10 × 10 × 40 ✓
After 3 × 3 conv.	10 × 10 × 20 ✓
After ReLU 2	10 × 10 × 20 ✓
After Max Pooling 2	6 × 6 × 20

?

3x3 conv.
40 chance.
1 stride
1 paddi

10 x10 x40.
10 x10 x40

Continuation Task 4

c) List the training steps of a stacked denoising autoencoder with three hidden layers.

6 P.

- Train an autoencoder with the single hidden layer.
- To train the autoencoder, corrupt the input with noise.
- Add a second hidden layer.
- Train the second hidden layer only by corrupting the first hidden layer's output only. This means that we pass a clean signal from the input and corrupt it after passing it through the first hidden layer.
- Add a third hidden layer.
- Train the third hidden layer only by corrupting the second hidden layer's output only.

d) Consider the reconstruction objective function of an encoder-decoder network that is combined with an L2-regularization:

5 P.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(f(\tilde{x}^i;\mathbf{w}),x^i) + \frac{\lambda}{2}\|\mathbf{w}\|^2 . \tag{1}$$

Compute the gradient and the update step of the gradient descent. In the solution, indicate the weight decay term.

$$\mathbf{w} = \mathbf{w} - \eta\nabla_{\mathbf{w}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(f(x^i;\mathbf{w}),y^i) + \frac{\lambda}{2}\|\mathbf{w}\|^2\right) \Rightarrow \tag{2}$$

$$\mathbf{w} = \mathbf{w} - \eta\nabla_{\mathbf{w}}\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(f(x^i;\mathbf{w}),y^i) - \eta\lambda\mathbf{w} \Rightarrow \tag{3}$$

$$\mathbf{w} = (1-\eta\lambda)\mathbf{w} - \eta\nabla_{\mathbf{w}}\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(f(x^i;\mathbf{w}),y^i). \tag{4}$$

where $(1-\eta\lambda)$ is the weight decay.

Task 5

/ 20 P.

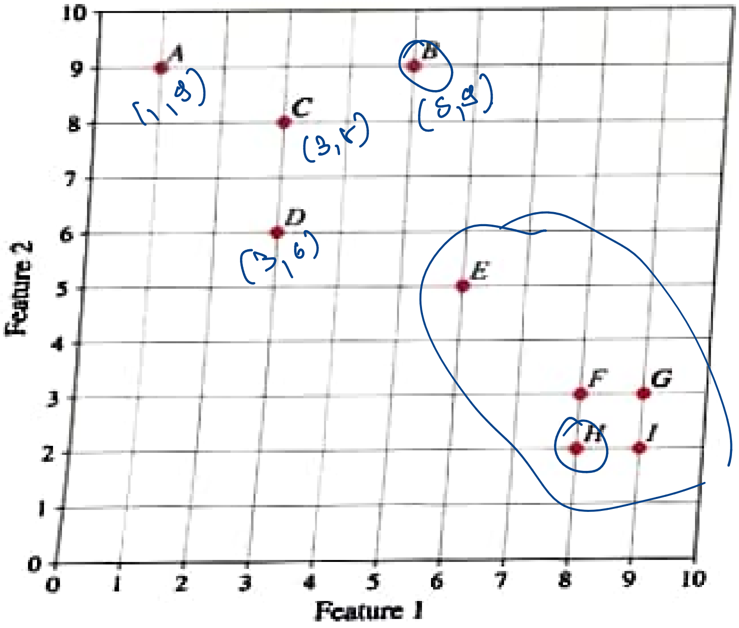
Clustering

Consider the K-means clustering task with the squared Euclidean distance as the distance metric to divide the dataset into 2 clusters (Cluster 1 and 2). The squared Euclidean distance d between points $\mathbf{p}$ with coordinates (p_1, p_2), and $\mathbf{q}$ with coordinates (q_1, q_2), is given by:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \|\mathbf{p} - \mathbf{q}\|^2 = (p_1 - q_1)^2 + (p_2 - q_2)^2. \tag{5}$$

The scatter plot of a dataset containing 9 samples is given in figure(a). The samples are annotated alphabetically. To help with your calculations, the squared Euclidean distance matrix containing the pairwise squared Euclidean distance between the data samples is provided in figure(b). For example, as denoted by the cells which are underlined, the squared Euclidean distance between points A and I is 113 units.

BA - 16 BF - 45 ~~HF - 1~~ HD - 41
BC - 5 BG - 52 ~~HG - 2~~ HC - 61
BD - 13 BH - 85 ~~HI - 1~~ HB - 58
BE - 17 BI - 65 HE - 13 HA - 28



(a) Scatter plot

8, 8

	A	B	C	D	E	F	G	H	I
A	0	16	5	13	41	85	100	98	113
B	16	0	5	13	17	45	52	58	65
C	5	5	0	4	18	50	61	61	72
D	13	13	4	0	10	34	45	41	52
E	41	17	18	10	0	8	13	13	18
F	85	45	50	34	8	0	1	1	2
G	100	52	61	45	13	1	0	2	1
H	98	58	61	41	13	1	2	0	1
I	113	65	72	52	18	2	1	1	0

(b) Distance matrix

Given that points B (cluster 1) and H (cluster 2) were randomly selected as cluster centres in the initialisation step, perform the first iteration of the K-means algorithm by answering the following questions.

Continuation Task 5

a) Find the nearest cluster (1 or 2) for each point using the cluster centres from the above initialisation step.

4 P.

By visual inspection we can assign clusters to each data sample as follows:

- Cluster 1: Points A, B, C, and D.
- Cluster 2: Points E, F, G, H and I.

b) What is value of the objective function calculated using the initial cluster centres? *Hint:* Recall that the objective function for a clustering problem with a dataset containing N samples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$, and K cluster centres $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K$, is given by:

2 P.

$$\mathcal{L} = \sum_{k=1}^{K}\sum_{i=1}^{N} l_{i,k}\,\|\boldsymbol{\mu}_k - \mathbf{x}_i\|^2, \qquad \text{where } l_{i,k} = \begin{cases} 1, & \text{if } x_i \in \text{cluster } k \\ 0, & \text{if } x_i \notin \text{cluster } k \end{cases} . \tag{6}$$

$$\mathcal{L} = (16 + 0 + 5 + 13) + (13 + 1 + 2 + 0 + 1) = 51$$

c) Calculate the new cluster centres after the first iteration of the algorithm.

4 P.

- Cluster 1:

$$A = \begin{bmatrix}1\\9\end{bmatrix}, \quad B = \begin{bmatrix}5\\9\end{bmatrix}, \quad C = \begin{bmatrix}3\\8\end{bmatrix}, \quad D = \begin{bmatrix}3\\6\end{bmatrix}$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix}\frac{1+5+3+3}{4}\\ \frac{9+9+8+6}{4}\end{bmatrix} = \begin{bmatrix}3\\8\end{bmatrix} = C$$

- Cluster 2:

$$E = \begin{bmatrix}6\\5\end{bmatrix}, \quad F = \begin{bmatrix}8\\3\end{bmatrix}, \quad G = \begin{bmatrix}9\\3\end{bmatrix}, \quad H = \begin{bmatrix}8\\2\end{bmatrix}, \quad I = \begin{bmatrix}9\\2\end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix}\frac{8+9+8+9+6}{5}\\ \frac{3+3+2+2+5}{5}\end{bmatrix} = \begin{bmatrix}8\\3\end{bmatrix} = F$$

Continuation Task 5

4 P.

d) **Find the nearest cluster (1 or 2) for each point using the new cluster centres after the first iteration.**

- Cluster 1: Points A, B, C, and D.
- Cluster 2: Points E, F, G, H and I.

2 P.

e) **What is value of the objective function calculated using the new cluster centres?**

$\mathcal{L} = (5 + 5 + 0 + 4) + (8 + 0 + 1 + 1 + 2) = 26$

2 P.

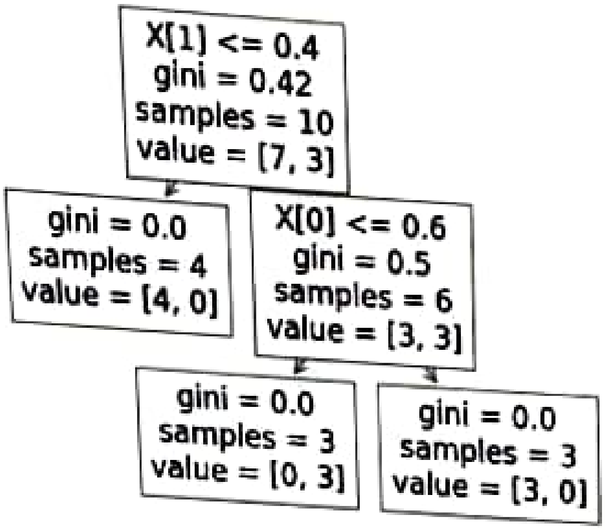
f) **Give an example each for a generative clustering model and a discriminative clustering model.**

- Generative Model: Gaussian Mixture Model
- Discriminative Model: K-means clustering
  (Alternate solution: Spectral Clustering or Agglomerative Clustering)

2 P.

g) Give an example each for a top-down clustering approach and a bottom-up clustering approach.

- Top-down approach: Spectral Clustering
- Bottom-up approach: Agglomerative Clustering



b) The leaf nodes of the tree can be used to perform classification of test samples. Report what is the most voted class (blue or red) for each leaf node of the above tree. You should use the node index for specifying each leaf node.

Node 1: blue, Node 3: red, Node 4: blue.

5 P.

c) Based on the constructed tree and samples at end nodes (leaf nodes) classify the following samples into blue and red class: (0.3, 0.2), (0.6, 0.3), (0.4, 0.6), (0.7, 0.4), (0.9, 0.6).

5 P.

- Class: blue, blue, red, blue, blue.