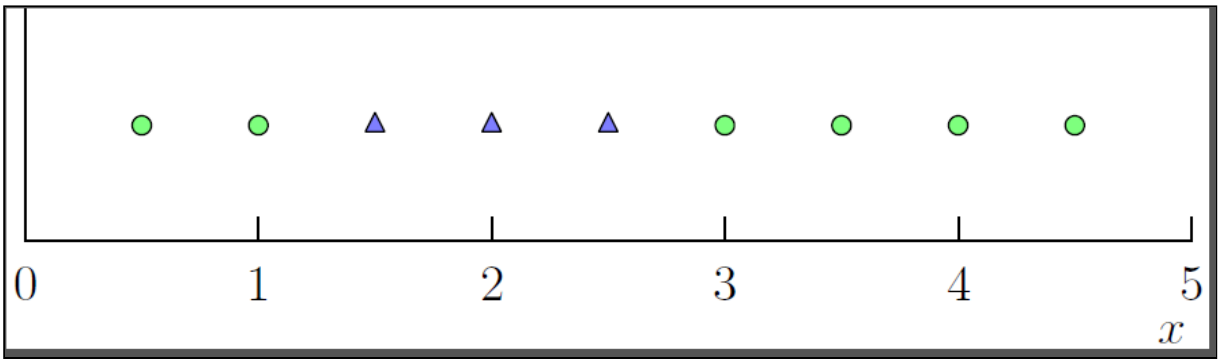
1 1 pt. Consider the following one-dimensional, two-class dataset, where each datapoint consists of a single measurement, x :



We want to classify this dataset with a linear classier, using two basis functions to project the data into a higher-dimensional space. Which set of basis functions will **not** allow a correct classication of the training data?

a. $\phi_1(x) = x \text{ and } \phi_2(x) = (x-2)^2$

b. $\phi_1(x) = x^2 \text{ and } \phi_2(x) = (x-2)^2$

c. $\phi_1(x) = x^2 \text{ and } \phi_2(x) = x$

d. Any of these basis functions allow linear separation of the classes

2 1 pt. The role of dropout in articial neural networks training is:

a. to prevent underfitting with regularization.

b. to model non-linear data.

c. to prevent overfitting using randomization.

d. to solve the curse of dimensionality problem

3 1 pt. Which of the below statements are true about kernel in SVM?

(1) Kernel function map low dimensional data to high dimensional space

(2) It's a similarity function

a. Only (2)

b. Only (1)

c. None

d. Both (1) and (2)

4
1 pt.

An articial neuron has three inputs x1, x2, and x3. Each input is connected to the neuron with a weight, w1, w2 and w3, respectively. Also, the neuron has a bias b and a ReLU (Rectified Linear Unit) activation function. Assuming that x1 = 1, x2 = 2, x3 = 0.5, w1 = 0.5, w2 = -1, w3 = -2, and b = 1 which is the neuron output value rounded to three decimals?

a. 0

b. -1.5

c. 1.5

d. -2.5

5
1 pt.

Consider the training example shown in the table below for a binary classification problem.

$Instance$	$a1$	$a2$	$Target$
1	Y	Y	$+$
2	Y	Y	$+$
3	Y	N	$-$
4	N	N	$+$
5	N	Y	$-$
6	N	Y	$-$
7	N	N	$-$
8	Y	N	$+$
9	N	Y	$-$

What is the classification error rate for a1? What is the best split (between a1 and a2) according to the Gini index?

a. classification error rate for a1 is 0.22; a2 is a better split

b. classification error rate for a1 is 0.44; a2 is a better split

c. classification error rate for a1 is 0.22; a1 is a better split

d. classification error rate for a1 is 0.49; a1 is a better split

6
1 pt.

You try to apply some classification technique to a dataset, and the classification technique is not capable of performing this classification well because the best discriminant is too complex to be learnt by the technique you're using. What symptoms do you observe:

a. The error on the training set is low, but the error on the test set is high

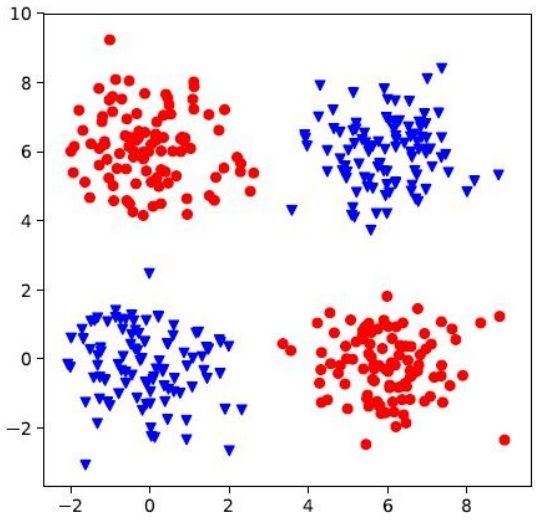
b. The error on both training and test set are high

c. The errors on both training and test set are low

d. You need cross-validation to identify this condition

7 Consider the dataset depicted below.

1 pt.



Which of the following statements is true?

a. A Naive Bayes classifier cannot perform better than random on this problem

b. A properly trained Naive Bayes classifier with Gaussian Mixture Models as conditional distributions will perform close to perfectly on this problem

c. A properly trained Naive Bayes classifier with uniform distributions as conditional distributions will perform close to perfectly on this problem

d. A properly trained Naive Bayes classifier with Gaussian distributions as conditional distributions will perform close to perfectly on this problem

8 Which of the following statements is true? Generative and discriminative probabilisitic models are probabilistic models, where:

1 pt.

a. Generative models use sampling to estimate target values for an observation, while discriminative models compute the target probabilities exactly

b. Generative models model the probability of the observation given the target, while discriminative models model the probability of the target given the observation

c. Generative models do not model the probability of the targets, but only the probability of the observations, while discriminative models do model the probability of the targets

d. Generative models model the joint probability of the observation and target, while discriminative models model the probability of the target given the observation

9
1 pt.

Node impurity can differ between 0 and 1 for a specific node. What does it mean when the node impurity is 1/2 at a binary node?

a. When the node impurity is 1/2, both classes at the node are equally likely, as a purely random selection.

b. When the node impurity is 1/2, all patterns at the node are part of the first class.

c. No answer is correct.

d. When the node impurity is 1/2, all patterns at the node are of the same class.

10
1 pt.

Suppose we want to minimize the function f(x; y) = x^2+y^2 subject to the constraint x+y=1. What are the coordinates for the minimum?

a. (x,y)=(-1/2,-1/2)

b. (x,y)=(1,0)

c. (x,y)=(0,1)

d. (x,y)=(1/2,1/2)

11
1 pt.

If the the loss-function of a neural network is non-convex then any optimization method derived from gradient descent used to train this neural network will:

Choose the correct answer.

a. We cannot decide from the given problem details if the network will diverge or converge.

b. Diverge.

c. Converge to the global optimum.

d. Converge to a local optimum.

12
1 pt.

In order for a Multilayer Perceptron to be able to model non-linear relations, the activation function of the hidden neurons must be:

Choose the correct answer.

a. two options, "linear" and "non-linear", are correct

b. linear

c. always zero

d. non-linear

13 1 pt.

In a driverless car system, we are creating a classifier to distinguish pedestrian from other obstacles on the road so that, in a case of an emergency situation, the car can make ethical decisions and avoid hitting pedestrians even at the cost of hitting other obstacles. In this situation, we assign a classification loss to the problem, as follows:

		Classification	
		Pedestrian	Obstacle
Real class	Pedestrian	0	100
	Obstacle	1	0

Given the following confusion matrices, which classifier is preferred with this loss model?

a.

		Classification	
		Pedestrian	Obstacle
Real class	Pedestrian	4025	124
	Obstacle	670	3985

b.

		Classification	
		Pedestrian	Obstacle
Real class	Pedestrian	3075	575
	Obstacle	125	4025

c.

		Classification	
		Pedestrian	Obstacle
Real class	Pedestrian	4025	575
	Obstacle	125	3075

d.

		Classification	
		Pedestrian	Obstacle
Real class	Pedestrian	4025	125
	Obstacle	575	3075

14
1 pt.

When predicting y given an observation $\mathbf{x}$ using Maximum a Posteriori learning, we compute the following probability of the prediction y, the observation x and a training set $\{\mathbf{x}, t\}$, using a set of model parameters $\boldsymbol{\theta}$:

a. $p(y|x, \hat{\theta})$ where we optimised $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}, \{\mathbf{x}\})$

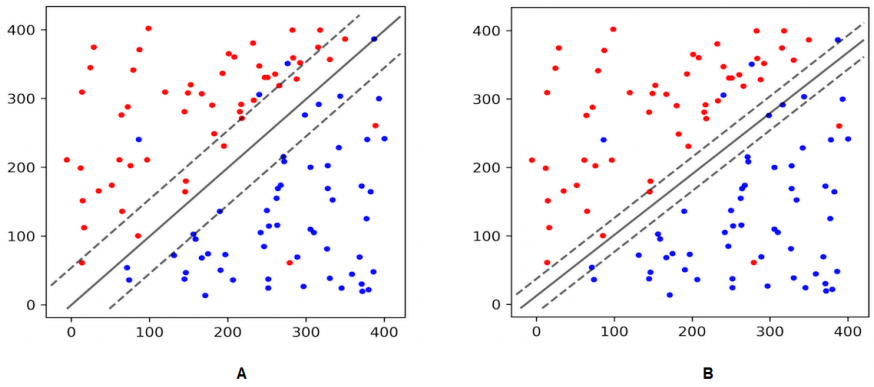
b. $p(y|\mathbf{x}, \hat{\boldsymbol{\theta}})$ where we optimised $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\{\mathbf{x}, t\})$

c. $p(y|\mathbf{x}, \hat{\boldsymbol{\theta}})$ where we optimised $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\ p(\{\mathbf{x}, t\}|\boldsymbol{\theta})$

d. None of the above.

15
1 pt.

Consider a dataset on which one applies SVM. Let C be the penalty parameter of the SVM used to control the margin. In which of the two cases, C has a larger value?



a. A and B

b. None

c. A

d. B

16
1 pt.

The *chain rule* is a key element in:

a. loss function

b. dropout

c. backprogation

d. data preprocessing

17 Given the following details of an articial neuron:

1 pt.

• $\mathbf{x}^T = [x_0, x_1, x_2]$, a vector collecting all inputs to the neuron, and in which x_0 is the bias.

• $\mathbf{w}^T = [w_0, w_1, w_2]$, a vector collecting all weights, and in which w_0 corresponds to the bias.

• y, a scalar value representing the true output for an input **x**

• f(z) = z, the activation function

• $o = f(\mathbf{x}^T\mathbf{w})$, the output value of the neuron

• Sum of Squared Error (SSE) as the loss function L

If we consider the following data point ($\mathbf{x}^T = [1,2,4]$, y = 2), the weights set to $\mathbf{w}^T = [2,3,-1]$, and a learning rate of 0.02, which is the new value of w_2 if we perform one weight update using the gradient descent update rule?

a. 1.16

b. -1.18

c. -1.16

d. -1.32

18 How does the regularization parameter λ effect the complexity of the model when L_1 or L_2 regularization is used?

1 pt.

a. A too large λ causes overfitting for L_1 and overfitting for L_2 regularization

b. A too large λ causes underfitting for L_1 and overfitting for L_2 regularization

c. A too large λ causes overfitting for L_1 and overfitting for L_2 regularization

d. A too large λ causes underfitting for L_1 and underfitting for L_2 regularization

19 Consider a 5 dimensional dataset on which one applies PCA. The covariance matrix corresponding to the PCA's has the following elements on the diagonal 12.1; 11.0; 4.1; 2.2; 1.3. How many principal components we need to use to capture at least 75% of the total variance in the dataset?

1 pt.

a. 5

b. 2

c. 1

d. 3

20
1 pt.

If PC_1 and PC_2 are both principal component vectors, what statement is correct about them?

a. variance along PC_2 is bigger than variance along PC_1

b. variance along PC_1 is bigger than variance along PC_2

c. No answer is correct.

d. PC_1 is parallel to PC_2

21
1 pt.

Assume that:
1: We have a two-class classication problem in a 3-dimensional space.
2: We apply Bayes' law to estimate $p(C_k|\mathbf{x})$, $k = 1, 2$.
3: We assume that the likelihoods $p(\mathbf{x}|C_k)$, $k = 1, 2$ are modelled by normal (Gaussian) probability density functions with full-rank covariance matrix (that is, unconstrained covariance matrices).

How many parameters does one need to estimate or learn from the data?

a. 28

b. 27

c. 19

d. 25

22
1 pt.

When using stochastic gradient descent to optimize the parameters of a linear regression or a multilayer perceptron model, what happens if we use a fixed learning rate which is too small?

a. The model will converge in normal time.

b. The model may converge in a very long time.

c. The model will surely converge very fast.

d. There are no problems, as gradient descent is robust to the choice of the learning rate.

23
1 pt.

Which of the following affirmations is false with respect to the relation between autoencoders and PCA.

a. They are both reinforcement learning methods.

b. PCA is only a linear transformation to the subspace while autoencoder is a nonlinear transformation to the hidden units.

c. They are both unsupervised learning methods.

d. They are both feature representation learning methods.

24
1 pt.

Complete the above proposition by selecting the right continuation.

For a multilayer perceptron model, a loss (or cost) function J(θ)...

a. ... is minimized using an optimization method in order to find the optimal model parameters θ.

b. ... is used to measure the memory requirements of the model.

c. ... computes analytically (directly from the data) the model parameters θ.

d. ... is used to measure the computational time necessary to train the model.

Correctiemodel

1. D
1 pt.

2. C
1 pt.

3. D
1 pt.

4. A
1 pt.

5. C
1 pt.

6. B
1 pt.

7. A
1 pt.

8. D
1 pt.

9. A
1 pt.

10. D
1 pt.

11. A
1 pt.

12. D
1 pt.

13. A
1 pt.

14. B
1 pt.

15. D
1 pt.

16. C
1 pt.

17. D
1 pt.

18. D
1 pt.

19. B
1 pt.

20. B
1 pt.

21. C
1 pt.

22. B
1 pt.

23. A
1 pt.

24. A
1 pt.