Questions

1. Suppose that we are training a linear classifier using the perceptron learning rule and that the current linear classifier is given by the line $2 + 2x_1 - x_2 = 0$. The next feature point in our training set is given by $x = (2, 4)$. Assume that this feature point is misclassified, what will be the new value for the weight $\mathbf{w}$ after one update if one applies a learning rate of 0.4?

   (a) $(1.6, 1.2, -2.6)$
   (b) $(2.4, 2.8, 0.6)$
   (c) $(2, 1.2, -2.6)$
   (d) $(1.2, 1.2, -0.6)$

2. Once again consider the situation of the previous question. In addition to the perceptron learning rule with learning rate 0.4 one also applies L_1 regularization of the form $|w_1| + |w_2|$ with parameter 0.1. What will now be the new value of $\mathbf{w}$?

   (a) $(1.5, 1.1, -2.5)$
   (b) $(1.6, 1.1, -2.5)$
   (c) $(2.4, 2.7, 0.5)$
   (d) $(1.6, 1.3, -2.7)$

3. Consider the neural network (NN) for which the input is 3 dimensional and that there are 2 neurons in the hidden layer and that there are 2 output neurons. The activation for all neurons is the sigmoid function σ. The weights of the NN are as follows. Hidden layer:

$$w_{1,0}^{(1)} = -5,\ w_{1,1}^{(1)} = 1,\ w_{1,2}^{(1)} = 2,\ w_{1,3}^{(1)} = 0 \tag{1}$$
$$w_{2,0}^{(1)} = -3,\ w_{2,1}^{(1)} = 1,\ w_{2,2}^{(1)} = 1,\ w_{2,3}^{(1)} = 1. \tag{2}$$

   Output layer:

$$w_{1,0}^{(2)} = -3,\ w_{1,1}^{(2)} = 1,\ w_{1,2}^{(2)} = 2 \tag{3}$$
$$w_{2,0}^{(2)} = 1,\ w_{2,1}^{(2)} = -1,\ w_{2,2}^{(2)} = 1 \tag{4}$$
$$\tag{5}$$

   What will be the output for the NN on the input $(x_1, x_2, x_3) = (2, 1, 1)$? A table with values for $\sigma(x)$ can be found at the end of the exam. Select the alternative which is closest to your answer.

   (a) $(0.22, 0.54)$
   (b) $(-1.3, 0.81)$
   (c) $(-1.3, 1.5)$
   (d) $(0.22, 0.81)$

4. Once again consider the NN of the previous question. Assume that for a given input the output is (0.6,0.4) and the target output is (1,1). Moreover assume that one applies stochastic gradient descent and the error function is given by:

$$\frac{1}{2}[(y_1 - t_1)^2 + (y_2 - t_2)^2]$$

What will be the vector $(\delta_1^{(2)}, \delta_2^{(2)})$ for the output neurons?

(a) $(-0.4, -0.6)$
(b) $(0.4, 0.6)$
(c) $(-0.10, -0.14)$
(d) $(0.10, 0.14)$

5. Once again consider the above NN structure and weights. Assume that the δ vector $(\delta_1^{(2)}, \delta_2^{(2)})$ of the output layer is $(-0.4, -0.6)$ and that the output of the hidden neuron 1 is 0.5 and the output of the hidden neuron 2 is 0.3. What will be the delta $(\delta_2^{(1)})$ of the hidden neuron 2?

(a) -0.29
(b) 0.05
(c) -1.4
(d) 0.02

6. Once again consider the same NN structure and weights as in the question above. Moreover the error (δ) vector of the output layer is $(-0.4, -0.6)$ and that the output of the hidden neuron 1 is 0.5 and the output of the hidden neuron 2 is 0.3. But now we assume that the NN shares the following weights: $w_{1,1}^{(1)} = w_{2,3}^{(1)}$, meaning these two variables are identical. What will be the adaptation dw to the weight $w_{1,1}^{(1)}$ if we apply a learning rate of 1?
Use for the input x the values $x = (2, 1, 1)$.

(a) -0.19
(b) -0.10
(c) 0.10
(d) -0.29

7. Consider a two-layer feedforward Neural Network. The back-propagation algorithm for this Neural Network consists of several steps:

   1: Calculate the δ's for the output layer.

   2: Calculate the output given the input, using forward-propagation.

   3: Update the weights of the output layer.

   4: Calculate the δ's of the hidden layer using back-propagation.

   5: Update the weights of the hidden layer.

   What is the right order of steps?

   (a) $1, 2, 3, 4, 5$

   (b) $2, 1, 3, 4, 5$

   (c) $2, 1, 4, 3, 5$

   (d) $2, 4, 5, 1, 3$

8. In image classification the following notions are relevant:

   (1) local features

   (2) translation invariance

   (3) rotation invariance

   Which of the above notions are incorporated in convolutional neural networks?

   (a) Only (1)

   (b) Only (2)

   (c) Both (2) and (3) and not (1)

   (d) Both (1) and (2) and not (3)

9. For marketing purposes a retailer wants to distinguish between costumers younger than 35 (class Y) and customers older than 35 (class O). The following table summarizes the data set in the data base of the retailer in an abstract form. The relevant attributes, determined by domain knowledge, are for convenience denoted by A with values *a1*, *a2* and *a3*, B with values *b1* and *b2*, C with values *c1* and *c2* and D with values *d1* and *d2*

A	B	C	D	Number of Instances	
				Y	O
a1	b1	c1	d1	4	12
a2	b1	c1	d2	3	2
a3	b1	c1	d1	6	0
a1	b2	c1	d2	8	0
a2	b2	c1	d1	4	0
a3	b2	c1	d2	9	0
a1	b1	c2	d1	2	4
a2	b1	c2	d2	2	8
a3	b1	c2	d1	3	4
a1	b2	c2	d2	2	2
a2	b2	c2	d1	2	6
a3	b2	c2	d2	5	2

What is the entropy of the above dataset above with respect to the class labels Y and O? Choose the alternative which is closest to your answer.

(a) 0.99

(b) 0.01

(c) 0.48

(d) 0.52

10. What is the Gini-index with respect to the class labels Y and O of the dataset of the previous question?

(a) 1.00

(b) 0.82

(c) 0.49

(d) 0.33

11. Consider a dataset similar to the above dataset with initial entropy of 0.95. For attribute (feature) B we get the following partition of the dataset:

B	Y	O
b1	40	10
b2	2	18

What is the information gain for attribute B? Select the alternative which is closest to your answer.

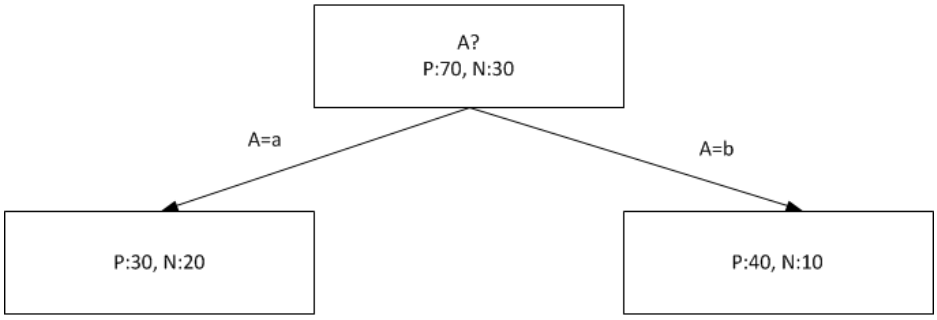
(a) 0.60

(b) 0.65

(c) 0.35

(d) 0.30

12. Consider the following part of the decision tree, with two leaf nodes and one parent node which splits on attribute A. The notation P:x N:y means that the node has x positive examples and y negative examples.



In order to apply χ^2 pruning one has to calculate, among others, the value of $\hat{p}_1$ and $\hat{n}_1$. What are the values of $\hat{p}_1$ and $\hat{n}_1$ in this case? Select the alternative closest to your answer.

(a) $\hat{p}_1 = 25$ and $\hat{n}_1 = 25$

(b) $\hat{p}_1 = 35$ and $\hat{n}_1 = 15$

(c) $\hat{p}_1 = 30$ and $\hat{n}_1 = 20$

(d) $\hat{p}_1 = 70$ and $\hat{n}_1 = 30$

13. Consider the following statements about Decision Trees:

    (I) An attribute with information gain equal to 0 is never added to the tree.

    (II) The tree induction algorithm based on the Gini-index will lead to an optimal tree.

    (III) If one uses information gain as heuristic then the attribute with the highest information gain will be the test attribute at the root of the tree.

    Which of the above statements are true?

    (a) Only (I) and (II)

    (b) Only (II)and (III)

    (c) Only (III)

    (d) None is true.

14. Consider the following confusion matrix

		Predicted class		
		C_1	C_2	C_3
Actual Class	C_1	110	8	7
	C_2	16	130	10
	C_3	26	5	120

    What is the accuracy of this classifier?

    (a) 110/125

    (b) 110/125+130/156+120/151

    (c) 360/432

    (d) 110/152+130/143+120/137

15. Once again consider the confusion matrix of the previous question. What is the recall for class C_1?

    (a) 110/125

    (b) 110/152

    (c) 110/172

    (d) 110/432

16. Consider a two class classification problem for which we apply a probabilistic approach. The loss matrix for this classification problem is given by:

$$\begin{pmatrix} 0 & 2 \\ 3 & 0 \end{pmatrix}$$

Assume that we apply a classification rule of the form: if $P(C_1|x) > \theta$ then x is classified as C_1. What is the optimal value for θ given the loss matrix above?

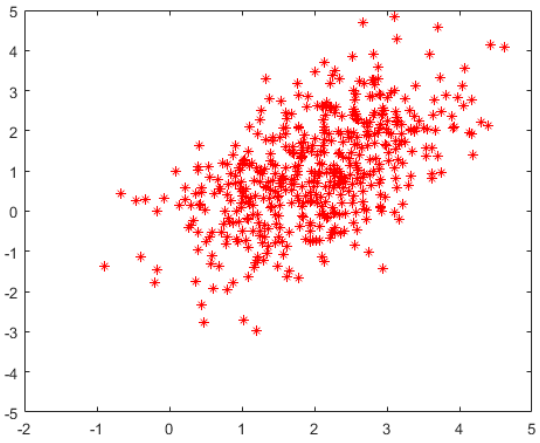
(a) 0.3

(b) 0.4

(c) 0.5

(d) 0.6

17. Consider the following visualization of a two dimensional dataset. The horizontal axis is x_1 and the vertical axis is x_2.



Which of the following directions is closest to the second principal component?

(a) $(0.5, 0.87)^T$

(b) $(-0.87, 0.5)^T$

(c) $(-0.87, -0.5)^T$

(d) $(-0.5, 0.87)^T$

18. 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 $11.0, 8.2, 6.4, 5.3, 4.1$. How much variance is explained by the first three PCAs?

(a) 31.4%

(b) 26.9%

(c) 94.4%

(d) 73.1%

19. Suppose we want to maximize the function $f(x_1, x_2) = 4 - 2x_1^2 - 3x_2^2$ subject to the constraint $2x_1 + x_2 = 3$. What are the coordinates for the maximum?

    (a) $(x_1, x_2) = (8/9, 11/9)$

    (b) $(x_1, x_2) = (1, 1)$

    (c) $(x_1, x_2) = (11/9, 5/9)$

    (d) None of the above.

20. Suppose we want to minimize the function $2x_1^2 + x_2^2 + 2x_1$ subject to the constraints:

    (1): $x_1 + x_2 \geq 0$ and

    (2): $-3x_1 + 2x_2 \leq 0$.

    Which constraints are active?

    (a) None

    (b) Only (1)

    (c) Only (2)

    (d) Both (1) and (2)

21. One problem with the probabilistic approach in which the class likelihoods are modelled by Normal (Gaussian) Distributions is the potential singularity of the estimated covariance matrix. Which of the following does **not** solve this singularity problem?

    (a) Assuming a shared covariance matrix.

    (b) Projecting the data on a lower dimensional space.

    (c) Assuming a diagonal shared covariance matrix.

    (d) Embedding the data in a higher dimensional space.

22. Assume that:

    1: We have a two class classification problem in a 4-dimensional space.

    2: We apply Bayes law to estimate $P(C_k|x)$, $k = 1, 2$.

    3: We assume that the likelihoods are modelled by normal (Gaussian) probability distributions.

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

    (a) 28

    (b) 30

    (c) 20

    (d) 24

23. Which of these statements about Batch Gradient Descent (BGD) and Stochastic Gradient Descent (SGD) is correct?

    (a) SGD optimises the function by cycling through the data once, modifying the parameters by computing the function gradient for each datapoint individually.

    (b) Each iteration of SGD computes the function gradient using a single datapoint (selected in some way, preferrably at random) and updates the model parameters.

    (c) Each iteration of SGD computes the function gradient using a single *misclassified* datapoint and updates the model parameters.

    (d) The only difference between SGD and BGD is the order in which datapoints are selected for gradient computation

24. Lagrange multipliers allow us to do "constrained function optimisation". Which of the following statements best describes what that means:

    (a) Lagrange multipliers allow us to find the maximum of a function that satisfies the constraints

    (b) Lagrange multipliers allow us to find *all* maxima of a function that satisfy the constraints

    (c) Lagrange multipliers allow us to find, of those locations in the input space for which the constraints are satisfied, those locations where the function is optimal

    (d) Lagrange multipliers allow us to find, of those locations in the input space for which the constraints are satisfied, those locations that are optima of the function

25. Mixture of Density networks output the parameters of a Probability Density Function (PDF). Which of the following statements is incorrect:

    (a) They can better deal with the high levels of noise in target values that are common in so-called "inverse problems"

    (b) They provide us with both a value for a best prediction given the input, and with a confidence estimate of the correctness of that value,

    (c) They can deal with training sets where identical inputs are associated with different targets, even when the differences in target are not due to noise

    (d) They are well-suited for problems that require the extra complexity and where we have enough training data to compensate for the extra model complexity, but will not perform as well as regular neural networks if this is not the case.

26. Naïve Bayes makes the simplifying assumption that all of the input dimensions are conditionally independent given the class. Which of the following statements is correct:

    (a) When the independence assumption is incorrect, Naïve Bayes will incorrectly classify the datapoint.

    (b) When the independence assumption is incorrect, Naïve Bayes will tend to overestimate the probability of the most likely class

    (c) The independence assumption makes the model faster to train, but does not reduce its performance

    (d) The naïve Bayes model is identical to a model with Gaussian class-conditional distributions with diagonal covariance matrices.

27. Which of the following statements is true? The "dual representation" of a Support Vector Machine (SVM)

    (a) Allows us to find a sparse solution for the problem

    (b) Expresses the weights of the SVM in terms of the training datapoints

    (c) Results in a model that uses the training data twice, and therefore requires less training data to obtain the same performance

    (d) Allows us to use a Lagrangian to indicate which training points are "support vectors", and is therefore much faster to compute than the "primal representation"

28. When making a classification decision based on Bayes' law, the evidence (denominator) does not need to be computed for each class, because:

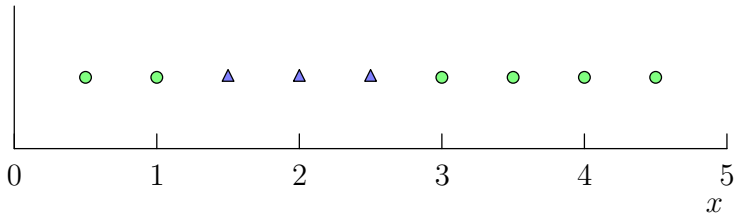
    (a) It does not affect the probability of the most likely class

    (b) It cannot change the probabilities so much that the classification decision would be changed

    (c) It is the same for all classes

    (d) You do need to compute it, actually, because otherwise you're not comparing probabilities

29. 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 classifier, using two basis functions to project the data into a higher-dimensional space. Which set of basis functions will allow correct classification of the training data?

(a) $\phi_1 = x^2, \phi_2 = x$

(b) $\phi_1 = x, \phi_2 = (x-2)^2$

(c) $\phi_1 = x^2, \phi_2 = (x-2)^2$

(d) All of the above

30. You are using a classification technique to a dataset, and you get a bad performance on the training set, but a good performance on the test set. How is this possible?

(a) The classifier is overfitting on the training set

(b) The training set is way too small for the model to train properly

(c) The classifier is not well suited for this problem

(d) The test set is way too small to be representative

31. In Bayesian learning, we compute the following probability of a prediction y given an observation $\mathbf{x}$ and a training set $\{\mathbf{x}, t\}$, using a set of model parameters $\boldsymbol{\theta}$:

(a) $p(y|\mathbf{x}, \{\mathbf{x}, t\})$, where parameters are explicitly marginalised out

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

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

(d) None of the above

32. "Leave-one-out" cross-validation is cross-validation with as many folds as we have datapoints in the training set. Which of the following statements is correct:

(a) Leave-one-out cross-validation minimises computational cost at the expense of precise evaluation of the classifier's performance

(b) Leave-one-out cross-validation optimises the evaluation of the classifier's performance at the expense of computational cost

(c) Leave-one-out cross-validation optimises both computational efficiency and the precise quantification of the classifier's performance

(d) Cross-validation is simply a shorter way of referring to leave-one-out cross-validation.

25223

NAAM / NAME Master

ADRES / ADDRESS

PLAATS / CITY

STUDENTNUMMER
STUDENT ID NUMBER

S 0 0 0 0 0 0 0

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

UNIVERSITEIT TWENTE.

NAAM VAK
COURSE NAME Basic ML

VAKCODE
COURSE CODE 201600070

VERSIE
VERSION

TENTAMENDATUM
EXAM DATE 9-11-2017

INSTRUCTIE / INSTRUCTIONS
ZO INVULLEN / FILL IN LIKE THIS: ●
NIET ZO / NOT LIKE THIS: ⊠ ✓

AANPASSEN / ADJUST
B= GOED / CORRECT

	A	B	C
01	○	●	⊠

MEERKEUZEVRAGEN / MULTIPLE-CHOICE QUESTIONS

	A	B	C	D	E	F	G
01	●	○	○	○	○	○	○
02	⊠	●	○	○	○	○	○
03	○	○	○	●	○	○	○
04	○	○	●	○	○	○	○
05	●	○	○	○	○	○	○
06	●	○	○	○	○	○	○
07	○	○	●	○	○	○	○
08	○	○	○	●	○	○	○
09	●	○	○	○	○	○	○
10	○	○	●	○	○	○	○
11	○	○	○	●	○	○	○
12	○	●	○	○	○	○	○
13	○	○	●	○	○	○	○
14	○	○	●	○	○	○	○
15	●	○	○	○	○	○	○
16	○	●	○	○	○	○	○
17	○	●	○	○	○	○	○
18	○	○	○	●	○	○	○
19	⊠	○	○	●	○	○	○
20	○	○	○	●	○	○	○
21	○	○	○	●	○	○	○
22	○	●	○	○	○	○	○
23	○	●	○	○	○	○	○
24	○	○	●	○	○	○	○
25	●	○	○	○	○	○	○

	A	B	C	D	E	F	G
26	○	●	○	○	○	○	○
27	○	●	○	○	○	○	○
28	○	○	●	○	○	○	○
29	○	○	○	●	○	○	○
30	○	○	○	●	○	○	○
31	●	○	○	○	○	○	○
32	○	●	○	○	○	○	○
33	○	○	○	○	○	○	○
34	○	○	○	○	○	○	○
35	○	○	○	○	○	○	○
36	○	○	○	○	○	○	○
37	○	○	○	○	○	○	○
38	○	○	○	○	○	○	○
39	○	○	○	○	○	○	○
40	○	○	○	○	○	○	○
41	○	○	○	○	○	○	○
42	○	○	○	○	○	○	○
43	○	○	○	○	○	○	○
44	○	○	○	○	○	○	○
45	○	○	○	○	○	○	○
46	○	○	○	○	○	○	○
47	○	○	○	○	○	○	○
48	○	○	○	○	○	○	○
49	○	○	○	○	○	○	○
50	○	○	○	○	○	○	○

True False VRAGEN/QUESTIONS

	TRUE	FALSE
01	○	○
02	○	○
03	○	○
04	○	○
05	○	○
06	○	○
07	○	○
08	○	○
09	○	○
10	○	○
11	○	○
12	○	○
13	○	○
14	○	○
15	○	○
16	○	○
17	○	○
18	○	○
19	○	○
20	○	○
21	○	○
22	○	○
23	○	○
24	○	○
25	○	○

	TRUE	FALSE
26	○	○
27	○	○
28	○	○
29	○	○
30	○	○
31	○	○
32	○	○
33	○	○
34	○	○
35	○	○
36	○	○
37	○	○
38	○	○
39	○	○
40	○	○
41	○	○
42	○	○
43	○	○
44	○	○
45	○	○
46	○	○
47	○	○
48	○	○
49	○	○
50	○	○

DOCENT / TEACHER

	0	1	2	3	4	5	6	7	8	9	10
01	○	○	○	○	○	○	○	○	○	○	○
02	○	○	○	○	○	○	○	○	○	○	○
03	○	○	○	○	○	○	○	○	○	○	○
04	○	○	○	○	○	○	○	○	○	○	○
05	○	○	○	○	○	○	○	○	○	○	○
06	○	○	○	○	○	○	○	○	○	○	○
07	○	○	○	○	○	○	○	○	○	○	○
08	○	○	○	○	○	○	○	○	○	○	○
09	○	○	○	○	○	○	○	○	○	○	○
10	○	○	○	○	○	○	○	○	○	○	○

Question 3. Correct answer is (1.6,-2.2,-0.4) - 0.2*(0,-1,1)=(1.6,-2.0,-0.6)

Question 9. Initial entropy is 0.97

Question 10: Gain is 0.24

Question 16: Direction of the first principal component is approximately the diagonal of the figure, so (-16,12) which is the same as the direction (-8/7,6/7) so B is the correct answer.

Question 17: Correct answer is 20.2/25.4=0.795