Exam Master Course Basic Machine Learning
Course code: 201600070,
Thursday February 1 2018

Name and student number

Name: _____

Student number: _____

Introduction

This exam is open book and consists of multiple-choice questions. You are allowed to use a simple calculator, but not your mobile phone, tablet or laptop or any other electronic means of computation or communication. Fill in your answers on the multiple choice answer form. Tips:

- Read each question carefully keeping the possible answers covered.
- Try to answer the question yourself, before you look at the answers you are given to choose from. Make a note of your first thoughts and calculations on a scribbling-paper. For this you can use the blank pages in the exam leaflet.
- Beware of double negations (negatives) as these can be confusing.
- Do not stay on any one question too long. If you do not know the answer and have spent more than 10 minutes on the question, move on to the next question and come back to this one later.
- Fill in your answers on the answer form and hand it in with your name and student number on it. Also hand in the exam.
- If there is some time left at the end, check your answers before you hand in exam and the answer form. Did you write your name and student id on it?

Good luck!

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) $(1.6, 2.8, -2.6)$

→ (b) $(2.4, 1.2, 0.6)$

   (c) $(2.4, -1.2, -0.6)$

   (d) $(2.4, 2.8, -2.6)$

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) $(2.4, 1.1, 0.7)$

   (b) $(1.6, 2.9, -2.7)$

   (c) $(2.4, 1.3, 0.5)$

   (d) $(1.6, 2.7, -2.5)$

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

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

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

   Output layer:

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

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

$$\tag{5}$$

   What will be the output of the hidden layer of the NN on the input $(x_1, x_2) = (2, -1)$? 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) $(-0.73, -0.88)$

   (c) $(1.00, 2.00)$

→ (d) $(0.73, 0.88)$

4. Once again consider the NN of the previous question. Assume that for a certain given input the output is (0.6,0.4) and the target output is (1,0). 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) $(-0.40, 0.40)$
(b) $(0.60, 0.40)$
(c) $(0.40, -0.40)$
(d) $(-0.01, 0.01)$

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 $(0.6, -0.4)$ and that the output of the hidden neuron 1 is 0.3 and the output of the hidden neuron 2 is 0.2. What will be the delta $(\delta_1^{(1)})$ of the hidden neuron 1?

(a) 0.40
→ (b) 0.08
(c) -2.4
(d) -0.38

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 $(0.6, -0.4)$ and that the output of the hidden neuron 1 is 0.3 and the output of the hidden neuron 2 is 0.2. But now we assume that the NN shares the following weights: $w_{1,1}^{(1)} = 2 \times w_{2,1}^{(1)}$, (**be aware of the factor 2!**). What will be the formula for the adaptation of the weight $w_{1,1}^{(1)}$ if we apply a learning rate of 1?

(a) $-(\delta_1^{(1)} \times x_1 + 2 \times \delta_2^{(1)} \times x_1)$.
→ (b) $-(\delta_1^{(1)} \times x_1 + 1/2 \times \delta_2^{(1)} \times x_1)$.
(c) $-(\delta_1^{(1)} \times x_1 + 2 \times \delta_2^{(1)} \times x_2)$.
(d) None of the above.

7. 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)

8. A data analyst has collected data (see table below) about customer loans. The goals is to predict, based on the customer profile, if a loan for a customer has a high risk or not.

payment history	debt	guarantee	income	risk
average	low	no	15-35 KEuro	low
average	low	no	0-15 KEuro	high
average	low	no	$>$ 35 KEuro	low
average	low	sufficient	$>$ 35 KEuro	low
bad	low	no	0-15 KEuro	high
bad	low	sufficient	$>$ 35 KEuro	low
good	high	sufficient	$>$ 35 KEuro	low
good	high	no	0-15 KEuro	high
good	high	no	15-35 KEuro	low
good	high	no	$>$ 35 KEuro	low
bad	high	no	15-35 KEuro	high

   What is the information gain of the attribute *payment history*?

→ (a) 0.11

   (b) 0.89

   (c) 0.85

   (d) 0.15

9. The analyst wants to learn the above classification problem using decision trees. If he uses "information gain" as selection criteria what will be the the attribute at the root of the decision tree (top node)?

   (a) payment history

   (b) debt

   (c) guarantee

→ (d) income

10. 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

11. 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 precision for C_1 for this classifier?

→ (a) 110/152

(b) 110/125

(c) 110/169

(d) None of the above.

12. 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 recall for class C_2?

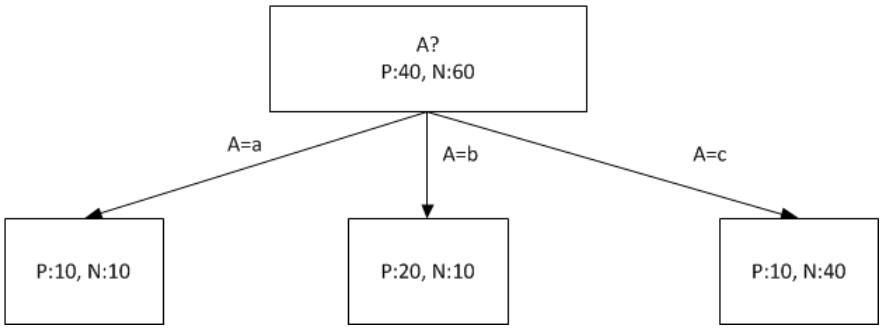
→ (a) 130/156

(b) 130/143

(c) 130/169

(d) 130/432

13. 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 $\hat{p}_i$ and $\hat{n}_i$, $i = 1, 2, 3$. Branches are numbered from left to right. What are the values of $\hat{p}_2$ and $\hat{n}_3$ in this case? Select the alternative closest to your answer.

(a) $\hat{p}_2 = 20$ and $\hat{n}_3 = 40$

→ (b) $\hat{p}_2 = 12$ and $\hat{n}_3 = 30$

(c) $\hat{p}_2 = 20$ and $\hat{n}_3 = 20$

(d) $\hat{p}_2 = 20$ and $\hat{n}_3 = 10$

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

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

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

15. Assume that:

1: We have a two class classification problem in a 3-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 with shared covariance matrices.

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

→ (a) 16 MP: Should that not be 6+6+1=13? Covariance matrix is symmetric

(b) 25

(c) 26

(d) 29

16. Which of these statements about Batch Gradient Descent (BGD) and Stochastic Gradient Descent (SGD) is not correct?

(a) SGD is more likely to find a good solution than BGD, because it is less likely to get stuck in a local optimum

→ (b) BGD is likely to require more iterations to find an optimum, because the gradient computed in SGD is not exact

(c) SGD is deals with redundant datapoints in the training set more efficiently than BGD

(d) The stopping criterion is harder to implement with SGD, because the error is not guaranteed to decrease at every iteration

17. 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 the optima of the function, those optima for which the constraints are satisfied

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

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

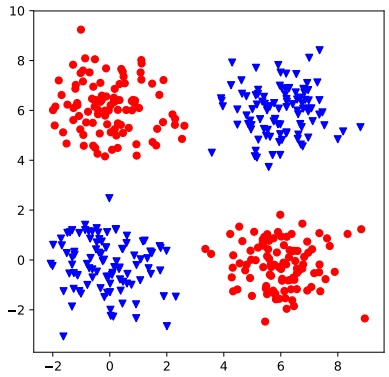
    (a) They can better deal with the high levels of noise in target values that are common in so-called "inverse problems"

    (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.

19. Consider the dataset depicted below.



    Which of the following statements is true?

    (a) A properly trained Naïve Bayes classifier with uniform distributions as conditional distributions will perform close to perfectly on this problem

    (b) A properly trained Naïve Bayes classifier with Gaussian distributions as conditional distributions will perform close to perfectly on this problem

    (c) A properly trained Naïve Bayes classifier with Gaussian Mixture Models as conditional distributions will perform close to perfectly on this problem

    → (d) A Naïve Bayes classifier cannot perform better than random on this problem

20. Which of the following statements is true?

    (a) Tangent propagation is formally equivalent to implementing code that randomly modifies using the same function

    (b) Tangent propagation is more efficient than implementing code that randomly modifies using the same function, but requires us to know the analytical derivative of that function

→ (c) We can represent the modifications that we do not want our model to be sensitive to as a mathematical function, and add a penalty term to our objective function that contains the first derivative of our output w.r.t. the parameters of that function

    (d) The only way in which we can improve generalisation is by making the model simpler or adding more training data

21. 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"

22. Consider a sequence of throws from a fair coin. Which of the following sequences is most likely (has highest probability of being observed?)

→ (a) T T T

    (b) H T T T

    (c) T H T H T

    (d) H T T H T

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

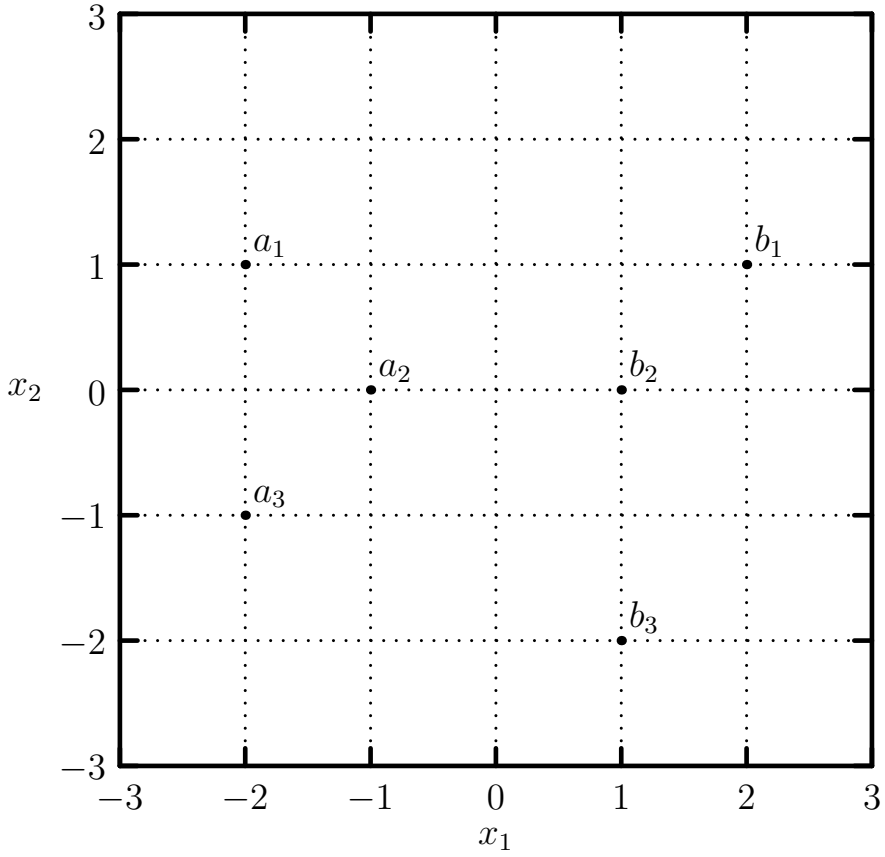
    (a) It does not affect the probability of the most likely class

    (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

24. Consider the dataset depicted below:



We want to classify this dataset with a support vector machine without kernel function. What are the support vectors in this case?

(a) a_2, a_3, b_2 and a_2, a_3, b_1 give rise to the same margin and are both correct solutions

→ (b) a_2, b_2, b_3

(c) a_2, b_2

(d) All points in this simple case are support vectors

25. 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) You need cross-validation to identify this condition

(b) The error on the training set is low, but the error on the test set is high

(c) The errors on both training and test set are low

→ (d) The error on both training and test set are high

26. You are trying to learn a very complex function using Neural Networks on a training set. Knowing that the training is sensitive to the initial parameter values, you use multiple random initialisations, train the models and keep the best performing model as evaluated on a validation set. When testing this model on a test set, your results are very bad. What is happening?

    (a) By training multiple models and evaluating them on a single validation set, the model is overfitting on the validation set.

    (b) The model is overfitting on the training set

→ (c) All of the above

    (d) None of the above

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

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

    (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 $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}, \{\mathbf{x}, t\})$

→ (d) Answers (b) and (c) are the same and correct

28. Which statement about K-fold cross validation is true?

    (a) Every data element is used K-1 times for testing and 1 time for training.

    (b) Every data element is used K times for testing and K times for training.

    (c) Every data element is used 1 time for testing and K times for training.

→ (d) Every data element is used 1 time for testing and K-1 times for training.

29. What is the purpose of regularization?

    (a) Training the model on k different folds of the data to reduce overfitting.

    (b) Reducing the number of weights to reduce overfitting.

→ (c) Preventing large weight values to reduce overfitting.

    (d) Maximising the performance of the model on unseen data to reduce overfitting.

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

    (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 model the joint probability of the observation and target, while discriminative models model the probability of the target given the observation

    (d) 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

31. Find the optimum of $x_1^2 + x_2$, subject to the constraint that $x_2 - x_1 = 5$. Which of the following statements is correct?

(a) $x_1 = \frac{1}{2}, x_2 = \frac{9}{2}$, this is a maximum.

→ (b) $x_1 = \frac{1}{2}, x_2 = \frac{9}{2}$, this is a minimum.

(c) $x_1 = 0, x_2 = 5$, this is a maximum.

(d) $x_1 = 0, x_2 = 5$, this is a minimum.

Table for $-p\log_2(p)$

p	$-p\log_2(p)$	p	$-p\log_2(p)$	p	$-p\log_2(p)$
0	0	1/8	0.38	1/10	0.33
1	0	2/8	0.50	2/10	0.46
1/2	0.50	3/8	0.53	3/10	0.52
1/3	0.53	4/8	0.50	4/10	0.53
2/3	0.39	5/8	0.42	5/10	0.50
1/4	0.50	6/8	0.31	6/10	0.44
2/4	0.50	7/8	0.17	7/10	0.36
3/4	0.31	1/9	0.35	8/10	0.26
1/5	0.46	2/9	0.48	9/10	0.14
2/5	0.53	3/9	0.53	1/11	0.31
3/5	0.44	4/9	0.52	2/11	0.45
4/5	0.26	5/9	0.47	3/11	0.51
1/6	0.43	6/9	0.39	4/11	0.53
2/6	0.53	7/9	0.28	5/11	0.52
3/6	0.50	8/9	0.15	6/11	0.48
4/6	0.39			7/11	0.42
5/6	0.22			8/11	0.33
1/7	0.40			9/11	0.24
2/7	0.51			10/11	0.13
3/7	0.52				
4/7	0.46				
5/7	0.35				
6/7	0.19				

Table for $\sigma(x)$

x	$\sigma(x)$
-4	0.02
-3	0.05
-2	0.12
-1	0.27
0	0.50
1	0.73
2	0.88
3	0.95
4	0.98