Exam Machine Learning Master Course
Course code: 192166420 & 201300239,
Friday June 24, 2016.

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!

1. Which of the following can cross-validation *not* help us obtain?

   (a) A more accurate estimate of the final ("true") performance of a model

   (b) Values for so-called "hyper-parameters" of a model, such as the number of clusters in a clustering problem, or the amount of regularisation in a prediction problem

   (c) For a given model, a better estimation of the parameter values

   (d) An quantification of the uncertainty over the estimated true performance of the model

   (e) None of the above

2. Suppose that we are training a linear classifier using the perceptron learning rule and that the current linear classifier is given by the line $2 - x_1 + 2x_2 = 0$. The next feature point in our training set is given by $x = (3, 6)$. Assume that this feature point is misclassified, what will be the new value for the weight $\mathbf{w}$ if one applies a learning rate of 0.4?

   (a) $(1.0, 3.0, 6.0)$

   (b) $(1.6, -2.2, -0.4)$

   (c) $(2.4, 0.2, 4.4)$

   (d) $(2, -2.2, -0.4)$

3. Once again consider the situation of the previous question. In addition to the perceptron learning rule with learning rate 0.4 one also applies regularization of the form $|w_1| + |w_2|$ (L_1 regularization) with parameter 0.2. What will now be the new value of $\mathbf{w}$?

   (a) $(1.6, -2.0, -0.8)$

   (b) $(1.6, -2.0, -0.2)$

   (c) $(1.4, -2.0, -0.2)$

   (d) None of the above

4. Consider the neural network (NN) for which the input is 2 dimensional and that there are 3 neurons in the hidden layer and that there is 1 output neuron. The activation for all hidden neurons and the output neuron is the sigmoid function σ. The weights of the NN are as follows. Hidden layer:

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

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

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

   Output layer:

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

   What will be the output for the NN on the input $(x_1, x_2) = (1, 1)$? Select the value which is closets to your answer.

   (a) 0.5

   (b) 0.1

(c) 0.6
(d) 0.9

5. Once again consider the NN of the previous question. Assume that for a given input the output is 2/3 and the target output is 1. Moreover assume that one applies stochastic gradient descent and the error function is given by:

$$\frac{1}{2}(y-t)^2$$

What will be the δ for the output neuron?

(a) 2/27
(b) 1/3
(c) -2/9
(d) -2/27

6. Once again consider the above NN. Assume that the error δ of the output neuron is 0.2 and that the output of the hidden neuron 1 is 0.4, the output of the hidden neuron 2 is 0.6 and the output of hidden neuron 3 is 0.1 . What will be the delta ($\delta_2^{(1)}$) of the hidden neuron 2?

(a) 0.048
(b) -0.048
(c) 0.2
(d) 0.002

7. Once again consider the same situation as in the question above (the error δ of the output neuron is 0.2 and that the output of the hidden neuron 1 is 0.4, the output of the hidden neuron 2 is 0.6 and the output of hidden neuron 3 is 0.1), but now we assume that the NN shares the following weights: $w_{1,1}^{(1)} = w_{3,2}^{(1)}$, 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?
Assume that the input $(x_1, x_2) = (1, 1)$.

(a) -0.048
(b) 0.018
(c) 0.03
(d) None of the above

8. 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	14	2
a2	b1	c1	d2	8	2
a3	b1	c1	d1	6	4
a1	b2	c1	d2	8	4
a2	b2	c1	d1	4	2
a3	b2	c1	d2	10	0
a1	b1	c2	d1	2	4
a2	b1	c2	d2	2	8
a3	b1	c2	d1	2	4
a1	b2	c2	d2	2	2
a2	b2	c2	d1	2	6
a3	b2	c2	d2	0	2

What is the posterior probability $P(Y|(a1, b2, c1, d2))$ if the retailer assumes a multinomial distribution for the likelihoods? Choose the alternative which is closest to your answer.

(a) 2/3
(b) 2/15
(c) 3/75
(d) 6/10

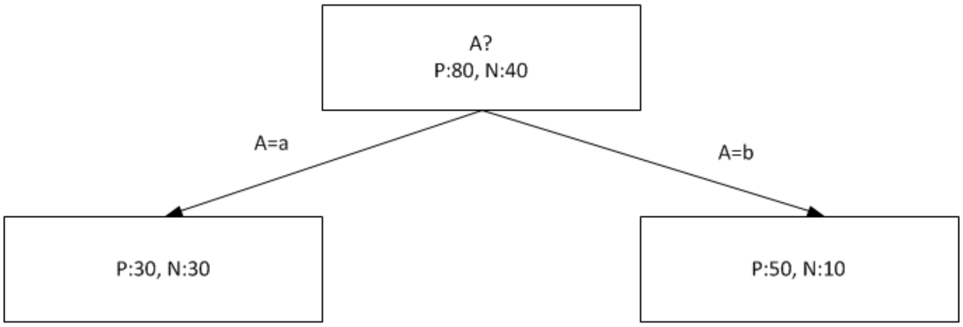
9. Consider the above dataset. What is the entropy of this dataset with respect to the class labels Y and O?

(a) 0.92
(b) 0.08
(c) 0.33
(d) 0.67

10. Consider the above dataset. What will be the gain for attribute (feature) C if one uses the classification error as heuristic?

(a) 0.00
(b) 0.33
(c) 0.26
(d) 0.07

11. 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 $\hat{n}_i$. What are the values of $\hat{n}_1$ and $\hat{p}_2$ in this case? Select the alternative closets to your answer.

(a) $\hat{n}_1 = 30$ and $\hat{p}_2 = 25$

(b) $\hat{n}_1 = 20$ and $\hat{p}_2 = 10$

(c) $\hat{n}_1 = 30$ and $\hat{p}_2 = 50$

(d) $\hat{n}_1 = 20$ and $\hat{p}_2 = 40$

12. Consider the following confusion matrix

		Predicted class		
		C_1	C_2	C_3
Actual Class	C_1	110	15	20
	C_2	16	140	10
	C_3	22	3	130

What is the accuracy of this classifier?

(a) 140/466

(b) 110/148+140/158+130/160

(c) 380/466

(d) 110/145+140/166+130/155

13. Once again consider the confusion matrix of the previous question. What is the recall for class C_2?

(a) 140/158

(b) 140/166

(c) 140/26

(d) 140/28

14. Which of the following terms describes the ratio between correctly classified instances in class C and all instances classified as class C?

    (a) Accuracy
    (b) Precision
    (c) Recall
    (d) Entropy

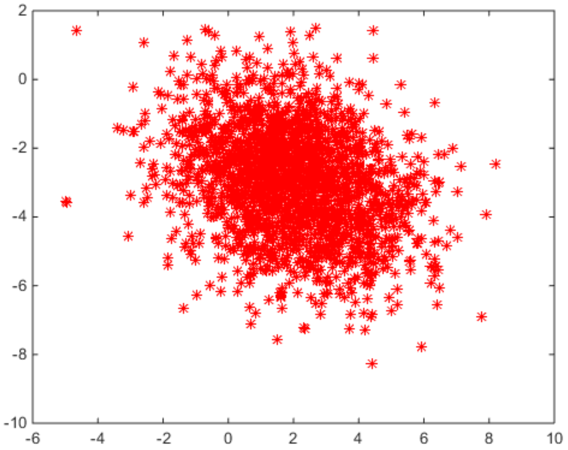
15. Consider a two class classification problem for which we apply a probabilistic approach. The loss matrix for this classification problem is given by:

$$\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$$

    Assume that we apply a classification rule of the form: if $P(C_2|x) > \theta$ then x is classified as C_2. What is the optimal value for θ given the loss matrix above?

    (a) 0.33
    (b) 0.67
    (c) 0.50
    (d) 0.20

16. Consider the following visualization of a two dimensional dataset.



Which direction is closets to the first principal component?

(a) $(1,1)^T$

(b) $(-1,1)^T$

(c) $(1,0)^T$

(d) $(1,0)^T$

17. Consider a 5 dimensional dataset on which one applies PCA. The covariance matrix corresponding to the PCA's has the the following elements on the diagonal $12.0, 8.2, 4.0, 1.1, 0.1$. How much variance is explained by the first two PCAs?

(a) 12.0%

(b) 20.2%

(c) 52.8%

(d) 74.7%

18. Support Vector Machines find a "sparse solution" to the classification problem: they express the discriminant as a function of a subset of the training examples. This means that:

(a) Only a subset of the training data needs to be stored and compared to at testing, making actual classifications faster.

(b) Training is very fast since not all training datapoints are considered

(c) Training is slow since we need to find the best set of training datapoints that together result in the largest margin, but the model is less likely to overfit because it is kept as simple as possible, but speed at test time is unaffected

(d) Both training and testing are faster because fewer datapoints are used

19. Let $\mathcal{D}$ be the set of training data points and $\boldsymbol{\theta}$ the model parameters. Which of the following quantities will Maximum a Posteriori (MAP) learning maximise?

    (a) $p(\mathcal{D}, \boldsymbol{\theta})$

    (b) $p(\mathcal{D}|\boldsymbol{\theta})$

    (c) $p(\boldsymbol{\theta}|\mathcal{D})$

    (d) $p(\mathcal{D})$

    (e) $p(\boldsymbol{\theta})$

20. You try to apply some classification technique to a dataset, and you observe that the error on the training set is low, but the error on the test set is high. Why is that?

    (a) Your method finds a solution that is too simple compared to the actual optimal solution to the problem

    (b) Your method finds a solution that is too complicated compared to the actual optimal solution to the problem

    (c) Your method gets stuck in a bad local optimum

    (d) Any of the above could be the reason

21. You try to apply some classification technique to a dataset, and you get a bad performance on the training set, but a good performance on the test set. What is happening here?

    (a) You are overfitting your training set

    (b) Your technique is not well suited for this problem

    (c) Your training set is way too small

    (d) Your test set is way too small

22. Batch gradient descent (BGD) optimises a function on a training dataset, by following the gradient as computed on the complete training set. As a consequence:

    (a) BGD is guaranteed to find the optimal solution for the function

    (b) BGD can get stuck in local optima, while stochastic gradient descent (SGD) can avoid these more easily because not every step is guaranteed to improve the solution.

    (c) BGD is faster than SGD, because it actually computes the correct gradient and therefore requires fewer iterations to converge

    (d) Iterations of SGD are faster than iterations of BGD, so that SGD is guaranteed to converge faster

23. Maximise $1 - x_1^2 - 2x_2^2$, subject to the constraint that $x_1 + x_2 = 1$. The solution is:

    (a) $x_1 = \frac{1}{3}, x_2 = \frac{2}{3}$

    (b) $x_1 = -\frac{1}{3}, x_2 = \frac{2}{3}$

    (c) $x_1 = \frac{1}{3}, x_2 = -\frac{2}{3}$

    (d) $x_1 = -\frac{1}{3}, x_2 = -\frac{2}{3}$

24. Mixture of Density Networks (MDN) are a way of dealing with non-Gaussian noise on the target values of a neural network. Which of the following statements is correct:

    (a) MDN do not predict the target value for a given input

    (b) MDN predict a distribution over outputs for a given output in the form of a Gaussian Mixture Model

    (c) MDN are not suitable for “inverse problems”

    (d) The number of mixture elements of an MDN must always be two.

25. When using k-Nearest-Neighbours:

    (a) The actual value of k does not matter much, as long as it is odd

    (b) Increasing the value of k reduces the risk of overfitting, but may lead to over-smoothing

    (c) It is impossible to have region in the space where the only strategy is to take a random guess, as long as k is odd

    (d) All of the above are correct

26. Which of the following statements is true? The “dual representation” of a Support Vector Machine (SVM)

    (a) Requires us to define a “kernel function”, but allows us to create non-linear discriminants

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

    (c) 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”

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

    (e) Projects the training data to a high-dimensional “kernel space”

27. The “regularisation” of a model reduces the risk of overfitting by:

    (a) Reducing the number of free parameters, making the model simpler

    (b) Adding a penalty models with many weights, favouring simpler models

    (c) Adding a penalty to large weight values, because large weights tend to result in overfitting

    (d) Constraining different weights to have the same value, to make the model more “regular”

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

    (a) Generative models model the joint probability of data and target, which allow them to compute the probability of the target even if information is missing

    (b) Generative models generate the correct target values by generating new data points

    (c) Discriminative models do not model the probability of the targets, but only the probability of the input data

    (d) Discriminative models only model the value the targets given the input data, not their distribution

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				