Machine Learning 2018

Final Exam

27 March 2018, 8:45–11:30

The questions start on the next page. **Please do not open the exam booklet until the scheduled starting time.**

Feel free to write on the exam booklet. You may take the questions home with you if you want. The exam will also be made available on Canvas later today, together with the correct answers.

To get a passing grade, you will need to get approximately 25 of the 40 questions correct. The true pass mark will be decided after the results have been analysed: if any questions are deemed to have been too difficult, they will become bonus questions.

Rules:

- You are allowed to use a calculator or graphical calculator.
- You are *not* allowed to use your phone or smartphone.
- The exam is closed-book.
- You are allowed to use the formula sheet provided through Canvas.
- The formula sheet should not have any writing on it, except in the cheat sheet box. Anything inside the box should be written by hand.

1 Questions

1. Which statement is **false**?
   A PCA provides a normalisation that accounts for correlations between features.
   B PCA finds a change of basis for the data.
   C The first principal component is the direction in which the variance is highest.
   D The second principal component is the direction in which the bias is highest.

2. What is the advantage of gradient descent over random search?
   A In gradient descent, parallel searches are allowed to communicate.
   B Gradient descent is less likely to get stuck in local minima.
   C Gradient descent computes the direction of steepest descent, random search approximates it.
   D Gradient descent is easier to parallelise.

3. Why is accuracy a bad loss function to use in gradient descent?
   A It is expensive to compute.
   B It makes the gradient zero almost everywhere.
   C It is unreliable in situations with high class imbalance.
   D The confidence interval is high on small test sets.

4. I'm performing spam classification. I represent each email by three numbers: how often the word *pill* occurs, how often the word *hello* occurs and how often the word *congratulations* occurs. What are these three attributes called?
   A The instances
   B The classes
   C The features
   D The principal components

5. There are many classification problems that are not linearly separable. Which trick often allows us to still separate the classes with a linear classifier?
   A Adding a regularizer.
   B Deriving new features from the existing ones.
   C Using least-squares-loss instead of accuracy-loss.
   D Reducing the dimensionality with PCA.

6. Which is **true**?
   A Grouping models have a higher refinement than grading models.
   B kNN has a higher refinement than linear classification.
   C Linear classification has a higher refinement than decision tree classification.
   D We can estimate the refinement from the confusion matrix.

7. Which metric can you **not** compute from the confusion matrix?
   A Accuracy **B** True positive rate **C** Recall **D** Area under the curve

Here we see the derivation of the gradient of the squared-error loss for linear regression.

$$\frac{\partial \frac{1}{2} \sum_i (f(x_i) - y_i)^2}{\partial b} = \frac{1}{2} \frac{\partial \sum_i (x_i w + b - y_i)^2}{\partial b} \tag{1}$$

$$= \frac{1}{2} \sum_i \frac{\partial (x_i w + b - y_i)^2}{\partial b} \tag{2}$$

$$= \sum_i (x_i w + b - y_i) \frac{\partial (x_i w + b - y_i)}{\partial b} \tag{3}$$

$$= \dots \tag{4}$$

8. To get from line (1) to line (2), we use the
   A Chain rule **B** Product rule **C** Exponent rule **D** Sum rule

9. To get from line (2) to line (3), we use the
   A Chain rule **B** Product rule **C** Exponent rule **D** Sum rule

10. In line 4, the correct result is
    A $\sum_i \left(x_i{}^2 w + b - y_i\right)$
    B $\sum_i x_i (x_i w + b - y_i)$
    C $\sum_i (x_i w + b - y_i)$
    D $\sum_i (x_i w + b - y_i)^2$

We have the following training set:

	x_1	x_2	label
a	0	0.1	T
b	2	0	T
c	1	0.5	T
d	3	1.5	F
e	2	2.5	T
f	1	3	F
g	2	4	F
h	4	3.5	F

For the following questions, it helps to draw the data and the classification boundary in feature space.
We use a linear classifier defined by

$$c(x_1, x_2) = \begin{cases} F & \text{if } 0 \cdot x_1 + x_2 - 2 > 0 \\ T & \text{otherwise.} \end{cases}$$

11. If we turn c into a *ranking* classifier, how does it rank the points, from F to T?
    A g h f e d c a b
    B g h f d e c a b
    C h d g e b f c a
    D a c f b e g d h

12. How many ranking errors does the classifier make?
    A None **B** 1 **C** 2 **D** 3

13. If we draw a coverage matrix (as done in the slides and book), what proportion of the cells will be red?
    A $\frac{1}{2}$ **B** $\frac{1}{4}$ **C** $\frac{7}{8}$ **D** $\frac{1}{16}$

14. Why is the naive Bayes classifier called *naive*?
    A It doesn't follow Bayes' rule, but an approximation.
    B It assumes that all features are independent.
    C It assumes that all features are independent, conditional on the class.
    D It assumes that all features are independent, conditional on the hyper-parameters.

15. Which statement is **true**?
    A The posterior is equal to the prior multiplied by the likelihood, divided by the data distribution.
    B The prior is equal to the posterior multiplied by the likelihood, divided by the data distribution.
    C The prior and the posterior are both equal to the likelihood divided by the data distribution.
    D A prior and a posterior are strictly *frequentist* concepts that cannot be related by probability.

16. What is the difference between a *discriminative* and a *generative* classifier?
    A A generative classifier learns a distribution on the data, a discriminative classifier doesn't.
    B A generative classifier is frequentist, a discriminative classifier is Bayesian.
    C A generative classifier can't be learned by gradient descent.
    D A discriminative classifier must be constructed from neural networks.

17. Which is **false**?
    A Softmax is an activation function that allows us to perform multiclass classification.
    B The derivative of the sigmoid activation is always either 0 or 1.
    C A single-layer neural network with a sigmoid output is the same model as logistic regression.
    D A single-layer neural network with a linear output is the same model as basic linear regression.

18. When do we require the *multivariate chain rule* in automatic differentiation?
    A When the computation graph is not a line.
    B When there are multiple paths between the output of the computation graph and one of the weights.
    C When the loss node in the computation graph has multiple inputs.
    D When there are latent variables in our model.

19. Which is **false**?
    A Backpropagation is symbolic differentiation, like Wolfram Alpha does.
    B Backpropagation is a mixture of symbolic and numeric differentiation.
    C Backpropagation applies the chain rule locally to each node in the computation graph.
    D Backpropagation distributes error back down the network based on weights and activations used during the forward pass.

20. What is the relation between the maximum margin hyperplane criterion (MMC) and the support vectors?
    A The support vectors can be removed from the data once the maximum margin hyperplane has been found.
    B The support vectors determine the hyperplane that satisfies the MMC.
    C The MMC and the support vectors describe different loss functions that we can use to fit a hyperplane.
    D The support vectors provide an approximation to the hyperplane that satisfies the MMC.

21. Which is **false**?
    A The kernel trick can be applied if we rephrase the SVM solution in terms of Lagrange multipliers.
    B The kernel trick allows us to extend our feature space without explicitly computing the extensions.
    C The kernel trick allows us to phrase the SVM algorithm purely in terms of dot products of pairs of instances.
    D The kernel trick allows us to use the SVM loss function in neural networks.

22. The probability density function of the univariate normal distribution is $N(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\mu - x)^2\right]$. What is the function of the factor $\frac{1}{\sqrt{2\pi\sigma^2}}$?

    A It allows us to translate the function to the left or right as desired.
    B It ensures that the total area under the curve is 1.
    C It ensures that the distance between the inflection points is 2.
    D It ensures that the distribution has a definite *scale*.

23. We have a naive Bayes classifier with two binary features, and two classes. We decide to add pseudo-observations for the purposes of smoothing. How many pseudo-observations do we need to add?
    A 1 **B** 2 **C** 4 **D** 8

24. We build a Bayes classifier by fitting an MVN to each class. What would happen to the MVNs if we made this classifier naive?

    A The decision boundary would become linear.
    B Their covariance matrices would become diagonal.
    C They would become Bayes optimal.
    D Their covariance matrices would have unit eigenvalues.

25. Define two random variables: Age, with outcomes $\{\text{child}, \text{teenager}, \text{adult}\}$ and Wealth, with outcomes $\{\text{poor}, \text{rich}\}$. Let these represent two attributes of the same person, selected at random from the Dutch population. Which entropies can we compute over these random variables?

    A We can compute the conditional entropy (of Age given Wealth and vice versa) and the cross-entropy of (Age vs. Wealth and Wealth vs. Age).
    B We can compute the conditional entropy, but not the cross-entropy.
    C We can compute the cross-entropy, but not the conditional entropy.
    D We can compute neither.

26. What happens if I build a two-layer neural network with no activation function on the units in the hidden layer?

    A The whole network becomes equivalent to a linear single layer network.
    B It becomes impossible to backpropagate a gradient to the first layer.
    C The optimization problem becomes non-convex.
    D We must use autodiff to compute the gradient.

Here are two distributions, p and q, on the members of a set $X = \{a, b, c, d\}$.

	p	q
a	$2/8$	$6/16$
b	$3/8$	$1/16$
c	$2/8$	$5/16$
d	$1/8$	$4/16$

27. What are their entropies?
    A $H(p) \approx 1.5$, $H(q) \approx 2.4$
    B $H(p) \approx 1.5$, $H(q) \approx 1.8$
    C $H(p) \approx 1.9$, $H(q) \approx 2.4$
    D $H(p) \approx 1.9$, $H(q) \approx 1.8$

28. What are their cross-entropies?
    A $H(p, q) \approx 2.5$,$H(q, p) \approx 2.2$
    B $H(p, q) \approx 1.9$,$H(q, p) \approx 2.2$
    C $H(p, q) \approx 2.5$,$H(q, p) \approx 2.4$
    D $H(p, q) \approx 1.9$,$H(q, p) \approx 2.4$

We will use the backpropagation algorithm to find the derivative of the function

$$f(x) = \frac{x^2 + 1}{x^3}$$

with respect to x.
First, we break the function up into modules:

$$f = \frac{a}{b}$$
$$a = \dots$$
$$b = x^3$$
$$c = \dots$$

29. What should be in the place of the dots?
    A $a = c + 1$, $c = x^2$
    B $a = c^2 + 1$, $c = x^2$
    C $a = c + 1$, $c = x^3$
    D $a = c^2 + 1$, $c = x^3$

30. Work out the *local* derivatives. Which is the correct expression for the gradient?
    A $\frac{\partial f}{\partial x} = \frac{2x}{b^2} - \frac{2ax}{b^2}$
    B $\frac{\partial f}{\partial x} = \frac{2x}{b} - \frac{2ax}{b^2}$
    C $\frac{\partial f}{\partial x} = \frac{2x}{b^2} - \frac{3ax^2}{b^2}$
    D $\frac{\partial f}{\partial x} = \frac{2x}{b} - \frac{3ax^2}{b^2}$

31. We want to find the minimum of the function $f(x, y) = x^3 + y^2 + 1$, subject to the constraint that $\sin(x) = \cos(y)$. If we use the method of Langrange multipliers, what is our L-function?
    A $L(x, y, \lambda) = 3x^2 + 2y + 1 + \lambda \sin(x) + \lambda \cos(y)$
    B $L(x, y, \lambda) = 3x^2 + 2y + 1 + \lambda \sin(x) - \lambda \cos(y)$
    C $L(x, y, \lambda) = x^3 + y^2 + 1 + \lambda \sin(x) + \lambda \cos(y)$
    D $L(x, y, \lambda) = x^3 + y^2 + 1 + \lambda \sin(x) - \lambda \cos(y)$

32. What is the purpose of a regularizer?

A It determines the number of layers your neural network should have.
B It adds nonlinearity to your neural network.
C It negates the need for a validation set.
D It reduces overfitting.

33. What is the *credit assignment problem*?

A The problem of how to train weights based on incomplete data.
B The issue of how to propagate a negative error down a neural network.
C The problem in reinforcement learning that rewards often come long after the action that was responsible for them.
D The issue that unsupervised learning problems do not have training labels, so the learning algorithm cannot be rewarded for correct behavior

34. We train a generative model through plain gradient descent, using random examples from the data as targets, comparing these against random samples from the models, and backpropagating the error. After training, all samples from the model look like the average over the dataset. We can avoid this by training in a different way. Which is **not** a training method that allows us to avoid this problem?
A Expectation-maximization **B** Random search **C** Variational autoencoders **D** Generative adversarial networks

35. How do cycleGANs avoid mode collapse?

A By adding a cycle-consistency term to the loss function.
B By cycling through different hyperparameters during training.
C By adding a "cycle" network that infers a latent representation.
D By adding a cycle-regularization term to the latent representation.

36. I want to test a matrix factorization method for a basic movie recommendation task. I hold out some users and some movies as a test set. Why won't this work?

A You will not learn representations during training for the users and movies in your test set.
B You will be committing multiple testing.
C Matrix factorization is an *unsupervised* task, so it cannot be evaluated with a test set.
D The training data will contain examples that are in the future from the perspective of your test set.

Consider the following task. The aim is to predict the class y from the binary features x_1, x_2, x_3 and x_4.[1]

x_1	x_2	x_3	x_4	y
A	A	A	B	Yes
A	A	B	A	Yes
A	A	A	B	Yes
A	A	B	A	Yes
B	B	A	B	No
A	B	A	B	Yes
B	A	A	B	No
A	B	B	A	Yes
B	A	B	A	No
B	B	A	B	No
B	B	B	A	No
B	B	B	A	No

37. In standard decision tree learning (as explained in the lectures), without pruning. Which would be the first feature chosen for a split?
A x_1 **B** x_2 **C** x_3 **D** x_4

38. Which features would be chosen on the subsequent nodes?
A x_2 on both. **B** x_3 on both. **C** x_3 and x_4. **D** The algorithm would terminate after the first split.

The EM and VAE algorithms are both based on the following decomposition.

$$\begin{aligned}
L(q,\theta) + KL(q,p) &= \mathbb{E}_q \ln \frac{p(x,z \mid \theta)}{q(z)} - \mathbb{E}_q \ln \frac{p(z \mid x,\theta)}{q(z)} \\
&= \mathbb{E}_q \ln p(x,z \mid \theta) - \mathbb{E}_q \ln q(z) - \mathbb{E}_q \ln p(z \mid x,\theta) \quad \langle a \rangle \\
&= \mathbb{E}_q \ln p(x,z \mid \theta) - \mathbb{E}_q \ln p(z \mid x,\theta) \\
&= \mathbb{E}_q \ln \frac{\langle b \rangle}{p(z \mid x,\theta)} = \mathbb{E}_q \ln \frac{p(z \mid x,\theta)p(x \mid \theta)}{p(z \mid x,\theta)} \\
&= \mathbb{E}_q \ln p(x \mid \theta) = \ln p(x \mid \theta)
\end{aligned}$$

39. What should be in place of $\langle a \rangle$ and $\langle b \rangle$?
A $\langle a \rangle : +\mathbb{E}_q \ln q(z)$, $\langle b \rangle : p(x \mid z,\theta)$
B $\langle a \rangle : -\mathbb{E}_q \ln q(z)$, $\langle b \rangle : p(x \mid z,\theta)$
C $\langle a \rangle : +\mathbb{E}_q \ln q(z)$, $\langle b \rangle : p(x,z \mid \theta)$
D $\langle a \rangle : -\mathbb{E}_q \ln q(z)$, $\langle b \rangle : p(x,z \mid \theta)$

[1] Note that in the computation of entropy, $0 \log 0$ is defined as 0. This may or may not be relevant, depending on how you arrive at the answer to this question. Don't worry if you found an answer without running into this problem.

40. Why does this decomposition help us?

A It rewrites $\mathrm{KL}(q, p)$, which is what we want to maximize.

B It rewrites $\ln p(\mathbf{x} \mid \theta)$, which is what we want to maximize.

C It allows us to learn an approximation p to the function q, which we cannot compute directly.

D It transforms expectations, which we cannot compute, into approximations, based on samples.

Thank you for your effort. Please check that you've put your name and student number on the answer sheet.
The student evaluations will be open after this exam. Please remember to take a few minutes to fill them in. Your feedback is crucial for us.