Machine Learning 2018

Resit Exam

2 July 2018, 12:00–14:45

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.

All questions are multiple choice. Only one answer is correct for each question.

1. A *lazy algorithm* is a machine learning method that simply stores the data and refers back to it during evaluation, instead of training to establish a good model which can be stored independent of the data. Which of the following methods is a lazy algorithm?
   A Linear classification
   B Decision trees
   C k-Nearest neighbors
   D None of the above

2. I want to predict house prices, from a set of examples, based on two attributes: surface area and the local crime rate. I create a scatterplot with the surface area of the house on the horizontal axis and the crime rate on the vertical. I plot each house in my dataset as a point in these axes. What have I drawn?
   A the model space
   B the loss curve
   C the feature space
   D the output space

3. How are random search and gradient descent related?
   A Gradient descent is an approximation to random search.
   B Random search is an approximation to gradient descent.
   C Gradient descent is like random search but with a smoothed loss surface.
   D Random search is like gradient descent but with a smoothed loss surface.

4. In the slides, we get the advice that "sometimes your loss function should not be the same as your evaluation function." Why not?
   A The evaluation function may not provide a smooth loss surface.
   B The evaluation function may be poorly chosen.
   C The evaluation function may not be linear.
   D The evaluation function may not be computable.

5. It is common practice in machine learning to separate out a training set and a test set. Often, we then split the training data again, to get a *validation* set. Which is **false**?
   A The validation set is not used until the end of the project.
   B We do this avoid multiple testing on the test set.
   C We use the validation set for hyperparameter optimization.
   D The test set is ideally used only once.

6. Which answer describes the *precision*?

   A The proportion of the actual positives that were classified as positive.
   B The proportion of the instances classified as positive that are actually positive.
   C The proportion of the actual negatives that were classified as negative.
   D The proportion of the instances classified as negative that are actually negative.

7. Imagine a machine learning task where the instances are customers. You know the phone number for each customer and their occupation (one of seven categories). You're wondering how to turn these into features. Which is **false**?

   A You can extract several useful categoric features from the phone number.
   B The phone number is an integer, so you should use it as a numeric feature.
   C Whether to use the occupation directly or turn it into a numeric feature depends on the model.
   D For some models, you may want to turn the occupation into several numeric features.

8. The slides mention two ways to adapt a categoric feature for a classifier that only accepts numeric features: *integer coding* and *one-hot coding*. Which is **true**?

   A One-hot coding always turns one categoric feature into one numeric feature.
   B Integer coding always turns one categoric feature into one numeric feature.
   C Integer coding becomes inefficient if there are too many values.
   D One-hot coding becomes inefficient if there are too few categories.

9. Which is **false**?

   A In PCA, the first principal component provides the direction of greatest variance.
   B PCA is a supervised method.
   C PCA can be used for dimensionality reduction.
   D PCA can be used for data preprocessing.

10. We are performing classification. We represent our instance by the random variable X and its class by the random variable Y. Which is **true**?
    A Generative modeling is training a model for $p(X \mid Y)$ and computing $p(Y \mid X)$ from that.
    B Discriminative modeling is training a model for $p(X \mid Y)$ and computing $p(Y \mid X)$ from that.
    C Generative modeling can only be done through the EM algorithm.
    D Discriminative modeling can only be done through the EM algorithm.

11. *Toxoplasmosis* is a relatively harmless parasitic infection that usually causes no obvious symptoms. Which statement is acceptable from a Bayesian perspective, but not from a frequentist perspective? Note that we don't care whether the statement is *correct*, just whether it fits these frameworks.
    A One in five Dutch people has toxoplasmosis.
    B Being Dutch, the probability that Fred has toxoplasmosis is 0.2.
    C The mean age of people with toxoplasmosis is 54.
    D The probability that a person chosen at random from the Dutch population has toxoplasmosis is 0.2.

12. How does stochastic gradient descent (SGD) differ from regular gradient descent?
    A SGD is used to train stochastic models instead of deterministic ones.
    B SGD trains in epochs, regular gradient descent doesn't.
    C SGD uses the loss over a small subset of the data.
    D SGD only works on neural networks.

13. Which is **false**?
    A Autodiff combines aspects of symbolic differentiation and numeric differentiation.
    B Autodiff computes the gradient but only for a specific input.
    C Autodiff is an alternative to backpropagation.
    D Autodiff boils down to repeated application of the chain rule.

14. Which is **false**?
    A The kernel trick allows us to use support vector machines as a loss function in neural networks.
    B The kernel trick allows us to compute SVMs in a high dimensional space.
    C The SVM algorithm computes the maximum margin hyperplane.
    D The SVM algorithm can be computed without using the kernel trick.

15. Which is **true**?
    A A maximum likelihood objective for least-squares regression does not provide a smooth loss surface.
    B The least-squares loss function for linear regression can be derived from a maximum likelihood objective.
    C Linear regression can be performed with a maximum likelihood objective but the results will be different from the least-squares version.
    D The loss function for logistic regression is derived from assuming a normal distribution on the residuals.

16. Which statement is **false**? [bonus question, due to multiple correct answers]
    A The entropy is the expected codelength using an optimal code.
    B The relative entropy is the KL divergence minus the entropy.
    C The KL divergence is the difference in expected codelength between the optimal code and another.
    D The KL divergence is the relative entropy minus the entropy.

17. What is the relation between the k-Means algorithm and the EM algorithm?
    A The EM algorithm is a simplified version of the k-Means algorithm.
    B k-Means is a simplified version of the EM algorithm.
    C k-Means is to k-Nearest neighbors as the EM algorithm is to Support Vector Machines.
    D k-Means is to k-Nearest neighbors as Support Vector Machines are to the EM algorithm.

18. I'm training a neural network. I notice that during training, the loss on the training data goes to zero, but the loss on the validation set doesn't get any better than chance. Which is true?
    A The model is overfitting. A good solution is to increase the model capacity.
    B The model is overfitting. A good solution is to add L2-regularization.
    C The model is suffering from vanishing gradients. A good solution is to use sigmoid activations.
    D The model is suffering from vanishing gradients. A good solution is to increase the batch size.

19. When training generative models, *mode collapse* is an important problem. Which is **false**?
A Generative Adversarial Networks are a way to train generative models, while avoiding mode collapse.
B Variational Autoencoders are a way to train generative models, while avoiding mode collapse.
C Generative Adversarial Networks avoid mode collapse by learning a network that maps each instance to a latent variable.
D Variational Autoencoders avoid mode collapse by learning a network that maps each instance to a latent variable.

20. Which is **false**?
A Decision trees do not deal with categorical data naturally. To use such data we must convert it to one-hot vectors.
B Decision trees do not deal with numeric data naturally. To use such data we must choose a value to split on.
C The standard decision algorithm (without pruning) operates *greedily*: once it has chosen a split, it will never reconsider that decision.
D When splitting a numeric feature, we must choose a threshold value to split on.

21. When building a decision tree, we choose which *feature* to split on, one after the other. With categoric features it makes no sense to split on the same feature twice. With numeric features it does. Which explanation is *entirely* correct? Let C be a categoric feature and N be a numeric feature.
A In the second split on C there will be no examples left. In the second split on N we can add noise to change the values.
B In the second split on C there will be no examples left. In the second split on N, we can set the threshold at a different value from the first.
C In the second split on C, all examples reaching that node will have the same value for C. In the second split on N we can add noise to change the values.
D In the second split on C, all examples reaching that node will have the same value for C. In the second split on N, we can set the threshold at a different value from the first.

22. Sarah has a large dataset of many recipes and many ingredients. She doesn't know anything about the recipes except which ingredients occur in each, and she doesn't know anything about the ingredients except in which recipes they occur. She would like to predict new recipe/ingredient pairs for ingredients that could be added to existing recipes. Which is **true**?

    A She could model the recipes as instances with their ingredients as a single categorical feature, and solve the problem with a decision tree.
    B She could model the ingredients as instances and their recipes as a single categorical feature, and solve the problem with a decision tree.
    C She could model this as a matrix decomposition problem.
    D None of the algorithms described in the course are applicable.

23. Which is **false**?

    A Word2Vec creates embedding vectors of tokens in a sequence.
    B Recurrent Neural Networks are Neural Networks with cycles, that allow them to operate on sequences.
    C The advantage of Markov models over Recurrent Neural Nets is that they have a potentially unbounded memory.
    D LSTMs tend to have better memories than plain RNNs because of the use of forget gates.

24. Why is it *especially* important to choose your test and validation sets carefully when training on sequential data?

    A Unlike with non-sequential models, if you evaluate on your training data, you risk overfitting.
    B If your learning rate is too low, you risk selecting hyperparameters that gave good performance by random chance.
    C If you remove users randomly from the recommendation set, you will be training on incomplete movie representations (and vice versa).
    D If you sample your test data randomly from the sequence, you may be training on data that is in the future compared to some of your test instances.

25. Shortly after AlphaGo beat Lee Sedol, DeepMind introduced *AlphaGo zero*, a Go engine that could learn from only self-play. Which of the following was **not** a change introduced to make AlphaGo better?

    A Use random search instead of policy gradients.
    B Introduce residual connections between blocks of layers.
    C Combine the policy net and the value net into a single network with two outputs.
    D Use Monte Carlo Tree-Search as a way to generate a better policy than the one implemented by the neural net.

We want to train the following model:

$$f(x_i) = w{x_i}^2 + v x_i + b$$

with parameters w, v and b. We derive the gradient of the loss with respect to w as follows:

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

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

$$= \frac{1}{2} \sum_i \frac{\partial (f(x_i) - y_i)^2}{\partial (f(x_i) - y_i)} \frac{\partial (f(x_i) - y_i)}{\partial w} \tag{3}$$

$$= \frac{1}{2} \sum_i 2 (f(x_i) - y_i) \frac{\partial (f(x_i) - y_i)}{\partial w} \tag{4}$$

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

27. To get from line (3) to line (4), we use the
A Chain rule **B** Product rule **C** Exponent rule **D** Sum rule

28. Fill in the definition of f and work out the derivative with respect to w. Which is the correct result?
A $\frac{1}{2} \sum_i w{x_i}^2 + v x_i + b - y_i$
B $\sum_i w{x_i}^2 + v x_i + b - y_i$
C $\frac{1}{2} \sum_i w{x_i}^4 + v{x_i}^3 + b{x_i}^2 - y_i{x_i}^2$
D $\sum_i w{x_i}^4 + v{x_i}^3 + b{x_i}^2 - y_i{x_i}^2$

We have the following training set:

	x_1	x_2	label
a	0	0	F
b	2	0	F
c	1	1	F
d	2.5	1	F
e	3	2	T
f	6	2	F
g	5	3	T
h	8	3	T

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

$$c(x_1, x_2) = \begin{cases} T & \text{if } x_1 + 0 \cdot x_2 - 4 > 0 \\ F & \text{otherwise.} \end{cases}$$

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

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

31. If we draw a coverage matrix (as done in the slides and book), what proportion of the cells will be red?
    A $\frac{1}{8}$ **B** $\frac{2}{8}$ **C** $\frac{1}{15}$ **D** $\frac{2}{15}$

32. 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 makes a conditional independence assumption that is not usually realistic.
    D It assumes that all classes are independent, conditional on the features.

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

	p	q
a	1/4	6/8
b	1/4	1/8
c	2/4	1/8

33. What are their entropies?
    A $H(p) \approx 1.5$, $H(q) \approx 1.1$
    B $H(p) \approx 1.5$, $H(q) \approx 1.8$
    C $H(p) \approx 1.9$, $H(q) \approx 1.1$
    D $H(p) \approx 1.9$, $H(q) \approx 1.8$

34. What are their cross-entropies?
    A $H(p, q) \approx 1.5$,$H(q, p) \approx 1.9$
    B $H(p, q) \approx 2.4$,$H(q, p) \approx 1.9$
    C $H(p, q) \approx 1.5$,$H(q, p) \approx 1.1$
    D $H(p, q) \approx 2.4$,$H(q, p) \approx 1.1$

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

$$f(x) = \sin(x^2)\cos(x^3)$$

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

$$\begin{aligned} f &= ab \\ a &= \dots \\ b &= \dots \\ c &= x^2 \\ d &= x^3 \end{aligned}$$

35. What should be in the place of the dots?
    A $a = \sin(c)$, $b = \cos(d)$
    B $a = \sin(d)$, $b = \cos(c)$
    C $a = \sin(c^2)$, $b = \cos(d^3)$
    D $a = \sin(d^2)$, $b = \cos(c^3)$

36. Work out the *local* derivatives. Which is the correct expression for the gradient?
    A $b\cos(c)2x + a\sin(d)3x^2$
    B $b\cos(c)2x + a\sin(d)3x$
    C $b\cos(c)2x - a\sin(d)3x^2$
    D $b\cos(c)2x - a\sin(d)3x$

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
B	A	A	A	Yes
A	A	A	B	Yes
B	A	A	A	Yes
A	A	A	B	Yes
B	B	B	A	No
B	B	A	A	Yes
B	A	B	A	No
A	B	A	B	Yes
A	A	B	B	No
B	B	B	A	No
A	B	B	B	No
A	B	B	B	No

37. In standard decision tree learning (as explained in the lectures). Which would be the first feature chosen for a split?
A x_1 **B** x_2 **C** x_3 **D** x_4

38. If we remove that feature from the data, which would be chosen instead?
A x_1 **B** x_2 **C** x_3 **D** x_4

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

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

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

40. The term $KL(q,p)$ uses some shorthand. Which is **false**?
A p represents the distribution of the latent variable $\boldsymbol{z}$ given the observed variable $\boldsymbol{x}$.
B q represents the function $q(z \mid x)$ we use to approximate $p(z \mid x)$.
C The KL-divergence ($KL(\cdot,\cdot)$) expresses how different two distributions are.
D In the EM setting we can't minimize $KL(q,p)$ directly, so we use $L(q,\theta)$ as a lower bound for $\ln p(\boldsymbol{x} \mid \theta)$.

Thank you for your effort. Please check that your name and student number are filled in on the answer sheet.

[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.