Machine Learning

Practice Exam B

WITH ANSWERS

February 7, 2020

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, and take it with you afterwards. The exam will also be made available on Canvas later today, together with the correct answers. We recommend that you copy your final answers onto the booklet, so you can check how you did.

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 answer contains only *unsupervised* methods and tasks?

   A Clustering, Linear regression, Generative modeling
   B k-Means, Clustering, Density estimation ✓
   C Classification, Clustering, Expectation-Maximization
   D Logistic regression, Density estimation, Clustering

2. The two most important conceptual spaces in machine learning are the *model space* and the *feature space*. Which is true?

   A Every point in the model space represents a loss function that we can choose for our task.
   B Every point in the model space represents a single instance in the dataset.
   C Every point in the feature space represents a single feature of a single instance in the dataset.
   D Every point in the feature space represents a single instance in the dataset.✓

3. The ALVINN system from 1995 was a self-driving car system implemented as a classifier: a grayscale camera was pointed at the road and a classifier was trained to predict the correct position of the steering wheel based on the behavior of a human driver. In this example, which are the instances and which are the features?

   A The instances are the different cars the system is deployed in and the features are the angles of the steering wheel.
   B The instances are the angles of the steering wheel and the features are the different cars the system is deployed in.
   C The instances are the frames produced by the camera and the features are the pixel values.✓
   D The instances are the pixel values and the features are the frames produced by the camera.

4. What is the relation between the *loss landscape* and its *gradient*?

   A The gradient points in the direction that the loss increases the fastest.✓
   B The gradient points in the direction that the loss decreases the fastest.
   C The gradient is the region of the loss landscape where the loss is the lowest.
   D The gradient is the region of the loss landscape where the loss is the highest.

5. Which is a legitimate reason to prefer random search over gradient descent as a search method?
   A The loss surface is complicated, so I want the size of the steps to change as I approach a minimum.
   B I need to be sure that I've found a global minimum.
   C My model has multiple layers, so I want to use backpropagation.
   D My model is not differentiable.✓

6. What is the benefit of a convex loss surface?
   A It allows us to use the backpropagation algorithm.
   B It allows us to use evolutionary methods.
   C It ensures that there are no local minima other than the global minimum.✓
   D It allows gradient descent to escape local maxima.

7. Rob trains a k-nearest neighbors classifier. He withholds 20% of his data as a test set and uses the rest as his training data. He runs the training algorithm twenty times, for $k = 1$ to $k = 20$. For each, he computes the accuracy on the test set. Het gets the best accuracy for $k = 17$, so he reports this accuracy as an estimate of the performance of k-nearest neighbors on his data. What *fundamental* mistake has Rob made?
   A Rob is checking a linear range of values for hyperparameter k when a logarithmic range would be better.
   B He is using arbitrary values for k. He should use a grid search.
   C The test set should always be bigger than the training set.
   D By reusing his test set, he may be inflating his performance estimate and overfitting to an arbitrary value of k.✓

8. Accuracy is a very simple and effective performance metric, but in certain cases, we should be careful. Imagine a spam classifier that automatically deletes emails detected as spam. The user receives about one spam email for every legitimate email. Why should we be careful optimizing for accuracy?
   A Because we have very high class imbalance.
   B Because we have very high cost imbalance. ✓
   C Because the data arrives irregularly.
   D Because this is an online learning problem.

9. Maria is fitting a regression model to predict the year in which a particular piece of instrumental music was written. The prediction is based on various features like average and variance of loudness, rhythm, key etc. She realizes that she has many *outliers*: for instance, the atonal music of the 1920s produces extreme variations in loudness, and John Cage's piece *4'33"* from 1952 is entirely silent. What should she do?

   A She should remove these instances entirely. Removing outliers will make it easier to fit the data with a normal distribution.

   B She should remove these instances from the training data, but leave them in the test data.

   C She should leave these instances in the training data, but remove them from the test data.

   D She should leave these instances in. They are important examples of the data distribution.✓

10. When we want to model the throwing of a single die, using probability theory, we define a *sample space* and an *event space*. Which is true?

    A "Rolling an even number" is an element of the event space."Rolling a 1" is an element of the sample space.✓

    B "Rolling an even number" is an element of the sample space."Rolling a 1" is an element of the event space.

    C For this example, the event space is continuous, the sample space is discrete.

    D For this example, the sample space is continuous, the event space is discrete.

11. Let $f(\mathbf{x}) = \sigma(\boldsymbol{w}\mathbf{x} + b)$ be a logistic regression model. We interpret $f(\mathbf{x})$ as the probability that $\mathbf{x}$ has the positive class. If $\mathbf{x}$ actually has the negative class, what is the cross-entropy loss for this single example?

    A $-\log f(\mathbf{x})$

    B $-\log(1 - f(\mathbf{x}))$✓

    C $-\log f(\mathbf{x}) - \log(1 - f(\mathbf{x}))$

    D $\log f(\mathbf{x}) - \log(1 - f(\mathbf{x}))$

12. Frank is a researcher in the 1960s. He's just read about a new model called the *perceptron*, which is a highly simplified simulation of a single brain cell. Frank decides that if a brain is powerful because it chains together multiple brain cells, he should try to chain together multiple perceptrons, to make a network that is more powerful than a single perceptron. Why won't chaining perceptrons together work in this way?

    A A GPU is needed to compute the output of such a function.
    B The perceptron is a linear function, and the composition of linear functions is still a linear function. ✓
    C Such a model would suffer from *vanishing gradients*.
    D This is equivalent to *hypothesis boosting*, which has been proven to be impossible.

13. In a support vector machine, what are the support vectors?

    A The parameters w^T that are multiplied by the input $\mathbf{x}$ to produce a classification.
    B The parameters b that are added to the input $\mathbf{x}$ to produce a classification.
    C The positive and negative data points that are allowed to fall inside the margin.
    D The positive and negative data points that are closest to the decision boundary. ✓

14. In deep learning, what is the difference between lazy and eager execution (or evaluation)?

    A In lazy execution, the computation graph is compiled and kept static during training. In eager execution, it is built up for each forward pass.✓
    B In eager execution, the computation graph is compiled and kept static during training. In lazy execution, it is built up for each forward pass.
    C In lazy execution, the gradient is computed by numeric approximation, while eager execution uses the backpropagation algorithm.
    D In eager execution, the gradient is computed by numeric approximation, while lazy execution uses the backpropagation algorithm.

15. Why is the ReLU activation function often preferred over the sigmoid activation function, for hidden nodes?

    A It causes more vanishing gradients, which help learning.
    B The sigmoid function cannot be used with gradient descent.
    C Its derivative is almost always either 0 or 1, reducing vanishing gradients. ✓
    D The sigmoid function contains a point where the gradient is not defined.

16. When we apply the chain rule to a complex operation involving tensors, in order to use the backpropagation algorithm, the local derivatives might be something like the derivative of a vector with respect to a matrix. The result is a 3-tensor which is complex to work out, and expensive to store in memory. How do modern machine learning frameworks avoid this problem in their implementation of backpropagation?

    A They approximate the local derivative using random search.

    B They approximate the local derivative using the EM algorithm.

    C They don't compute the local derivative, but the product of the upstream derivative (the loss over the module outputs) with the local derivative ✓

    D They don't compute the local derivative, but the product of the downstream derivative (the module inputs over the network inputs) with the local derivative.

17. How is the log-likelihood like a loss function?

    A The log-likelihood is an approximation of the loss function.

    B The loss function is an approximation of the log-likelihood.

    C When we train a model, we minimize the loss, and when we fit a distribution, we often minimize the log-likelihood.

    D When we train a model, we minimize the loss, and when we fit a distribution, we often maximize the log-likelihood. ✓

18. If we train a generator network by comparing a random output to a random target example from the data and backpropagating the difference, we get *mode collapse*. The problem is that we don't know which random input corresponds to which item in the dataset. How do GANs solve this problem?

    A By training a second network to map the target example to a distribution on the input space.

    B By adding a KL-loss term on the random inputs of the generator network.

    C By training the generator to generate outputs that a second network recognizes as real, and training the second network to distinguish generated outputs from real data. ✓

    D By adding a cycle-consistency loss-term.

19. Sometimes we want to learn a model that maps some input to some output, but in some aspects we also want the model to behave like a generator. For instance, if we train a model to colorize photographs, we don't want it to be purely deterministic: to colorize the label of a beer bottle, it should randomly imagine some colors even if it can't infer the correct colors from the input. Which GAN approach is designed to accomplish this?

    A Vanilla GAN
    B Conditional GAN ✓
    C CycleGAN
    D StyleGAN

20. The Variational Autoencoder (VAE) differs from a regular autoencoder in several aspects. Which is **not** one of them?

    A It includes a discriminator, which tries to tell the difference between data points and samples from the generator. ✓
    B It has an added loss term that ensures that the data looks like a standard normal distribution in the latent space.
    C For a given instance, the encoder produces a distribution on the latent space, instead of a single point.
    D It includes a sampling step in the middle, where a latent vector is sampled from the distribution provided by the encoder.

21. The standard decision tree algorithm doesn't stop adding nodes until all leaves either contain no data instances, or only instances with the same label (or all features have been used). Why is this a problem, and what is the default solution (mentioned in the slides)?

    A It's a problem because the algorithm may never terminate. To solve it, we can use a validation set to see if removing nodes improves performance.
    B It's a problem because the algorithm may never terminate. To solve it, we can remove features from the data, so that fewer splits are available.
    C It's a problem because we may be overfitting on the training set. To solve it, we can use a validation set to see if removing nodes improves performance.✓
    D It's a problem because we may be overfitting on the training set. To solve it, we can add features to the data, so that more splits are available.

22. Boosting is a popular method to improve a model's performance. Why do we rarely see boosting used in research settings (unless specifically studying ensembling methods)?

    A In research we want to measure the relative performance with respect to the baseline. If we apply boosting to our model, we should apply boosting to the baseline as well.✓
    B Boosting cannot be applied in combination with a validation split, which is required in research.
    C Boosting makes it difficult to compute a confidence interval over the accuracy, which is required in research.
    D Boosting requires some information from the test set to be used in training, which is not allowed in research.

23. In recommender systems, what is *implicit feedback*?

    A Ratings given by a single "like" button rather than a more fine-grained system.
    B Recommendations that take the temporal structure of the data into account
    C Associations between users and items assumed from user behavior.✓
    D Recommendations derived from manually crafted item features rather than learned ones.

24. Word2Vec and matrix factorization are both *embedding* methods that make it possible to learn about a large set of featureless objects. How do they do this?

    A By taking known features for each object and mapping these to a low-dimensional representation.
    B By taking known features for each object and mapping these to a high-dimensional representation.
    C By embedding these objects into a Euclidean space, with each object represented by a vector. ✓
    D By embedding these objects into a Euclidean space, with each object represented by a scalar.

25. What is batch normalization?

    A An operation in a neural network that normalizes the output of a layer so that it is normally distributed over the current batch. ✓
    B An operation in a neural network that normalizes the output of a layer so that it is uniformly distributed over the batch.
    C A hyperparameter selection technique that sets the batch size to a value that ensures a normal distribution in the gradients of a neural network.
    D A hyperparameter selection technique that sets the batch size to a value that ensures a uniform distribution in the gradients of a neural network.

We want to train the following model:

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

with parameters v, w and b, where x_i and $f(x_i)$ are scalars. The dataset provides target values t_i. We derive the gradient of the loss with respect to v as follows:

$$\begin{aligned}
\frac{\partial \frac{1}{2} \sum_i (f(x_i) - t_i)^2}{\partial v} &= \frac{1}{2} \frac{\partial \sum_i (f(x_i) - t_i)^2}{\partial v} && (1)\\
&= \frac{1}{2} \sum_i \frac{\partial\,(f(x_i) - t_i)^2}{\partial v} && (2)\\
&= \frac{1}{2} \sum_i \frac{\partial\,(f(x_i) - t_i)^2}{\partial\,(f(x_i) - t_i)} \frac{\partial\,(f(x_i) - t_i)}{\partial v} && (3)\\
&= \frac{1}{2} \sum_i 2\,(f(x_i) - t_i) \frac{\partial\,(f(x_i) - t_i)}{\partial v} && (4)
\end{aligned}$$

26. To get from line (1) to line (2), we use the

    A Exponent rule
    B Product rule
    C Constant factor rule
    D Sum rule ✓

27. To get from line (3) to line (4), we use the

    A Exponent rule ✓
    B Product rule
    C Constant factor rule
    D Sum rule

28. Fill in the definition of $f(x_i)$ and work out the derivative with respect to v. Which is the correct result?

    A $-\frac{1}{2} \sum_i \left(-v{x_i}^2 + w x_i + b - t_i\right)$
    B $- \sum_i \left(-v{x_i}^2 + w x_i + b - t_i\right)$
    C $-\frac{1}{2} \sum_i x_i \left(-v x_i + w{x_i}^2 + b - t_i\right)$
    D $- \sum_i x_i \left(-v x_i + w{x_i}^2 + b - t_i\right)$ ✓

We have the following training set:

	x_1	x_2	label
a	0	1	Neg
b	2	2	Neg
c	1	4	Neg
d	4	5	Neg
e	3	6	Pos
f	6	8	Pos
g	5	3	Pos
h	8	7	Pos

For the following questions, it helps to draw the data and the classification boundary in feature space.

We use a linear classifier defined by

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

29. If we turn c into a *ranking* classifier, how does it rank the points, from most Negative to most Positive?

A a b c d e f g h
B a b g c d e h f
C a b g c d h e f
D a c b e d g f h ✓

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), what proportion of the cells will be red?

A $\frac{3}{15}$ **B** $\frac{6}{15}$ **C** $\frac{1}{8}$ **D** $\frac{1}{16}$✓

Here are two distributions, p and q, on the members of a set $X = \{a, b, c, d\}$. We will use binary entropy (i.e. computed with base-2 logarithms). Note that in the computation of the entropy $0 \cdot \log_2(0) = 0$, but otherwise, $\log_2 0$ is undefined as usual.

	p	q
a	$4/8$	$1/8$
b	$2/8$	$4/8$
c	$1/8$	$2/8$
d	$1/8$	$1/8$

32. What are their entropies?

A $H(p) = 2.25$, $H(q) = 1.75$
B $H(p) = 1.75$, $H(q) = 2.25$
C $H(p) = 2.25$, $H(q) = 2.25$
D $H(p) = 1.75$, $H(q) = 1.75$ ✓

33. What are their cross-entropies?

A $H(p, q) = 2 + 3/8$, $H(q, p) = 2 + 1/4$✓
B $H(p, q) = 2 + 1/4$, $H(q, p) = 2 + 3/8$
C $H(p, q) = 2$, $H(q, p) = 2 + 1/4$
D $H(p, q) = 2 + 1/4$, $H(q, p) = 2$

34. The KL divergence, a kind of distance between two probability distributions, is strongly related to the entropy. Which is **false**?

A The KL divergence $\text{KL}(p, q)$ is the bits wasted by using p as a compressor for elements sampled from q. ✓

B The KL divergence $\text{KL}(p, q)$ is the bits wasted by using q as a compressor for elements sampled from p.

C The KL divergence $\text{KL}(p, q)$ is the cross-entropy $H(p, q)$, with the entropy $H(p)$ subtracted so that the divergence is zero if the arguments are equal.

D The KL divergence $\text{KL}(p, q)$ is the cross-entropy $H(p, q)$, with the entropy $H(p)$ subtracted so that we measure compression relative to the optimum.

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

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

with respect to x.

First, we break the function up into modules:

$$\begin{aligned} a &= x^2 \\ b &= x^3 \\ c &= a + 1 \\ d &= b + 2 \\ f &= \frac{c}{d} \end{aligned}$$

35. What should be in the place of the dots?

A $a = x^3$, $b = x^2$, $f = c/d$

B $a = x^2$, $b = x^3$, $f = c/d$ ✓

C $a = x^3 + 2$, $b = x^2 + 1$, $f = d/c$

D $a = x^2 + 1$, $b = x^3 + 2$, $f = d/c$

36. In terms of the *local* derivatives. Which is the correct expression for the gradient $\frac{\partial f}{\partial x}$?

A $2xd^{-1} - 3x^2c^{-2}$

B $2xd^{-1} + 3x^2c^{-2}$

C $2xd^{-1} - 3cx^2d^{-2}$ ✓

D $2xd^{-1} + 3cx^2d^{-2}$

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.

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

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. 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.

$$\begin{aligned}
L(q, \theta) + \mathrm{KL}(q, p) &= \mathbb{E}_q \ln \frac{p(\mathbf{x}, \mathbf{z} \mid \theta)}{q(\mathbf{z} \mid \mathbf{x})} + -\mathbb{E}_q \ln \frac{p(\mathbf{z} \mid \mathbf{x}, \theta)}{q(\mathbf{z} \mid \mathbf{x})} \\
&= \mathbb{E}_q \ln p(\mathbf{x}, \mathbf{z} \mid \theta) - \mathbb{E}_q \ln q(\mathbf{z} \mid \mathbf{x}) - \mathbb{E}_q \ln p(\mathbf{z} \mid \mathbf{x}, \theta) + \mathbb{E}_q \ln q(\mathbf{z} \mid \mathbf{x}) \\
&= \mathbb{E}_q \ln p(\mathbf{x}, \mathbf{z} \mid \theta) - \mathbb{E}_q \ln p(\mathbf{z} \mid \mathbf{x}, \theta) \\
&= \mathbb{E}_q \ln \frac{p(\mathbf{x}, \mathbf{z} \mid \theta)}{p(\mathbf{z} \mid \mathbf{x}, \theta)} \\
&= \mathbb{E}_q \ln \frac{p(\mathbf{z} \mid \mathbf{x}, \theta) p(\mathbf{x} \mid \theta)}{p(\mathbf{z} \mid \mathbf{x}, \theta)} \\
&= \mathbb{E}_q \ln p(\mathbf{x} \mid \theta) \\
&= \ln p(\mathbf{x} \mid \theta)
\end{aligned}$$

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

40. Why does this decomposition help us?

A It eliminates q, which we cannot compute.
B It shows that $L(q, \theta)$ is a lower bound on the quantity we want to maximize.✓
C It eliminates expectations, which we cannot compute.
D It shows that $p(\mathbf{x} \mid \theta)$ is a lower bound on the KL divergence.

Thank you for your effort. **Please check that you've put your name and student number on the answer sheet.**