Machine Learning 2022

Final Exam

WITH ANSWERS

28 March 2022, 18:45–20:45

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

There are 40 questions, worth 1 point each.

To pass the examination, your total points for the exam plus those for the quizzes should be around 50–55. The exact pass mark will be decided *after* the exam 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 Recall questions

1. What is the difference between *classification* and *regression*?

   A Classification predicts an item from a finite set, regression predicts a numeric value. ✓

   B Regression predicts an item from a finite set, classification predicts a numeric value.

   C Classification is unsupervised, regression is supervised.

   D Regression is unsupervised, classification is supervised.

2. What is a valid reason to prefer *gradient descent* over *random search*?

   A My model is easily differentiable.✓

   B I need to be sure that I find the global minimum.

   C My loss function is not smooth.

   D There is some computation between the output of my model and my loss function, which I do not control.

3. We are training a classification model by gradient descent, and we want to figure out which learning rate to use, before comparing the model to other classifiers. We try five learning rate values, resulting in five different models. How do we choose among these five models?

   A We measure the accuracy of each model on the training set.

   B We measure the accuracy of each model on the validation set. ✓

   C We measure the accuracy of each model on the test set.

   D We measure the accuracy of each model on the full dataset.

4. *Undersampling* and *oversampling* are ways to deal with imbalanced classes. Which is **true**?

   A You oversample your majority class.

   B You undersample your minority class.

   C Undersampling leads to duplicate instances in your data.

   D Oversampling leads to duplicate instances in your data. ✓

5. 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 Integer coding always turns one categoric feature into multiple numeric features.

   B One-hot coding always turns one categoric feature into multiple numeric features. ✓

   C Integer coding becomes inefficient if there are too many categories.

   D One-hot coding becomes inefficient if there are too few categories.

6. If somebody says: “There is a high probability that the mean height of Italian women is below 2 meters.” Which is **true**?
   A A strict subjectivist would consider this an improper use of the term probability.
   B A strict Bayesian would consider this an improper use of the term probability.
   C A strict frequentist would consider this an improper use of the term probability.✓
   D In machine learning, we would consider this an improper use of the term probability.

7. How does dropout help with the overfitting problem?
   A By propagating the gradient of the loss back down the network.
   B By randomly disabling nodes in a neural network, to eliminate solutions that require highly specific configurations.✓
   C By ensuring that the output distribution of a neural network is normally distributed if the input distribution is.
   D By converting the scalar backpropagation algorithm to work with tensors.

8. The *soft margin SVM loss* is defined as a constrained optimization objective. We can rewrite this in two ways. Which is **true**?
   A We can rewrite to an unconstrained problem. This allows us to use the kernel trick.
   B We can rewrite to an unconstrained problem. This expresses the solution purely in terms of the dot product between pairs of instances.
   C We can rewrite using KKT multipliers. This allows us to use the kernel trick. ✓
   D We can rewrite using KKT multipliers. This allows some instances to fall inside the margin.

9. Neural networks usually contain *activation functions*. What is their purpose?
   A They are used to compute a local approximation of the gradient.
   B They are applied after a linear transformation, so that the network can learn nonlinear functions. ✓
   C They control the magnitude of the the step taken during an iteration of gradient descent.
   D They function as a regularizer, to combat overfitting.

10. By what method do variational autoencoders avoid mode collapse?
    A By training the “decoder” network through a discriminator.
    B By using a regularizer to steer the network toward the data average.
    C By feeding the discriminator network pairs of inputs.
    D By learning the latent representation of an instance through an “encoder” network.✓

11. I am training a generator network to generate faces. I take a random sample, compare it to a randomly chosen image from the data, and backpropagate the error. When training is finished, all samples from the network look like the average over all faces in the dataset. What name do we have for this phenomenon?

    A Multiple testing
    B Overfitting
    C Dropout
    D Mode collapse ✓

12. In some machine learning settings it is said that we must make a trade-off between *exploration* and *exploitation*. What do we mean by this?

    A That hyperparameter selection (exploration) uses computational resources that can also used in training the model (exploitation).
    B That an online algorithm needs to balance optimization of its expected reward with exploring to learn more about its environment. ✓
    C That an insufficiently thoroughly trained model may be biased against minorities.
    D This refers to the problem of balancing the loss function with the regularization terms in matrix factorization.

13. Which is **false**?

    A To use decision trees on data with categorical features, we must convert those features to one-hot vectors. ✓
    B To use decision trees on data with numeric features, we must choose a threshold value to split on, for every split.
    C When training a decision tree on only categorical features, there's no use in splitting again on a feature you've already split on.
    D When training a decision tree on numeric features, it can often be useful to split on a feature you've already used before.

14. Some models are built on the *Markov assumption*. What do we mean by this?

    A We can apply backpropagation to neural networks by unrolling them.
    B The probability of a word does not depend on the current class for which we are predicting the probability.
    C The operation of an LSTM cell depends only on its predecessors through two inputs.
    D A word is conditionally dependent only on a finite number of words preceding it.✓

2 Combination questions

15. Which is (primarily) a **supervised** machine learning method?
    A Principal Component Analysis
    B Support Vector Machines ✓
    C Variational Autoencoders
    D None of the above

16. We are fitting a regression model using the least squares loss. We have seen two different forms of the loss function: $\sum_i (y_i - t_i)^2$ and $\frac{1}{2}\sum_i (y_i - t_i)^2$ (where y_i is the model output and t_i is the true value given by the data). Which is **true**?
    A The global minima of these two loss functions occur at different points in the model space .
    B If we work out the solution analytically, when we set the gradient equal to zero, the constant factor $\frac{1}{2}$ in the second loss function changes the parameters of the optimal solution.
    C If we use these loss functions with gradient descent, it makes no difference which we use, the behavior is exactly the same.
    D If we use these loss functions with gradient descent, there is a small difference in which we use, but if we scale the learning rate appropriately, the difference will disappear.✓

17. We are choosing a new basis for our data. We decide to use an orthonormal basis. What is the advantage of having an orthonormal basis?
    A It ensures that the basis vectors are equal to the principal components.
    B It ensures that the inverse of the basis matrix is equal to its transpose. ✓
    C It ensures that the basis vectors are orthogonal to the principal components.
    D It ensures that the data is automatically whitened in the new basis.

18. We are considering using either gradient descent or random search for a problem. Which is **true**?
    A For both, which optimum they find depends on the initial starting point.✓
    B Gradient descent can get stuck in a local optimum, random search cannot.
    C Gradient descent cannot get stuck in a local optimum, random search can.
    D Gradient descent is more efficient than random search and can always be applied, so we always prefer gradient descent over random search.

19. Which property is common to both logistic regression and support vector machines?
    A For both, the decision boundary is chosen by minimizing the number of misclassified examples.
    B Both are usually optimized by alternating optimization.
    C Both require backpropagation to work out the gradient efficiently.
    D They both focus mostly or only on the points closest to the decision boundary. ✓

20. Imagine we have a naive Bayes classifier. In our dataset we have two binary features (categorical with two possible values) and two classes. How many pseudo-observations do we need to add if we want to apply Laplace smoothing?
    A 1
    B 2
    C 4 ✓
    D 8

21. One can choose between the likelihood function or the log likelihood function as a loss function. Which is usually preferred, and why?
    A Both result in a maximum at the same point in model space, but the log-likelihood is often easier to work with. ✓
    B The likelihood function. When this is maximised we have the best fitting model whereas the log likelihood results in a worse model.
    C The log likelihood function. The squared errors are minimized only when the log-likelihood is maximized.
    D The likelihood function. The squared errors are minimized only when the likelihood is maximized.

22. We have a logistic regression model for a binary classification problem, which predicts class probabilities q. We compare these to the true class probabilities p, which are always 1 for the correct class and 0 for the incorrect class. The slides mention two loss functions for this purpose: logarithmic loss and binary cross-entropy. Which is **true**?
    A Log-loss does not lead to a smooth loss landscape, so we approximate it by cross-entropy so that we can search for a good model using gradient descent.
    B Cross-entropy loss does not lead to a smooth loss landscape, so we approximate it by log-loss so that we can search for a good model using gradient descent.
    C Log-loss is equal to the binary cross-entropy $H(p, q)$. ✓
    D Log-loss is equal to the binary cross-entropy $H(q, p)$.

23. We want to represent color videos in a deep learning system. Each is a series of frames, with each frame an RGB image. Which is the most natural representation for *one* such video?

A As a 1-tensor.
B As a 2-tensor.
C As a 3-tensor.
D As a 4-tensor. ✓

24. In support vector machines, how is the *maximum margin hyperplane criterion (MMC)* related to 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.

25. Which of the following is **not** a method to prevent overfitting?

A Boosting ✓
B Bagging
C Dropout
D L1 regularization

26. What problem, if it exists for a single model, **cannot** be solved by training an ensemble of such models?

A High bias.
B High variance.
C High overfitting.
D High training time. ✓

27. I have a dataset of politicians in the European parliament and which past laws they voted for and against. The record is incomplete, but I have some votes for every law and for every politician.

    I would like to predict, for future laws, which politicians will vote for and which will vote against. I plan to model this as a recommender system using matrix factorization.

    Which is **true**?
    A This is not a good model, because there are too many classes and not enough instances.
    B This is not a good model, because there are not enough classes, and too many instances.
    C I would have to deal with the cold start problem, because for the future laws I don't have any voting information. ✓
    D I would have to deal with the cold start problem, because the voting record for past laws is incomplete.

3 Application questions

We want to train the following model:

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

with scalar parameters w and b, and scalar argument x_i. The dataset provides target values t_i. We derive the gradient of the loss with respect to w as follows:

$$\frac{\partial \frac{1}{2} \sum_i \left(w{x_i}^2 - b - t_i\right)^2}{\partial w} = \frac{1}{2} \frac{\partial \sum_i \left(w{x_i}^2 - b - t_i\right)^2}{\partial w} \tag{1}$$

$$= \frac{1}{2} \sum_i \frac{\partial \left(w{x_i}^2 - b - t_i\right)^2}{\partial w} \tag{2}$$

$$= \frac{1}{2} \sum_i \frac{\partial \left(w{x_i}^2 - b - t_i\right)^2}{\partial \left(w{x_i}^2 - b - t_i\right)} \frac{\partial \left(w{x_i}^2 - b - t_i\right)}{\partial w} \tag{3}$$

$$= \frac{1}{2} \sum_i 2 \left(w{x_i}^2 - b - t_i\right) \frac{\partial \left(w{x_i}^2 - b - t_i\right)}{\partial w} \tag{4}$$

28. To get from line (1) to line (2), we use the
    A chain rule.
    B product rule.
    C constant factor rule.
    D sum rule.✓

29. To get from line (2) to line (3), we use the
    A chain rule. ✓
    B product rule.
    C constant factor rule.
    D sum rule.

30. Finish the derivative. Which is the correct result?
    A $\frac{1}{2}\sum_i x_i \left(w{x_i}^2 - b - t_i\right)$
    B $\sum_i x_i \left(w{x_i}^2 - b - t_i\right)$
    C $\frac{1}{2}\sum_i {x_i}^2 \left(f(x_i) - t_i\right)$
    D $\sum_i {x_i}^2 \left(f(x_i) - t_i\right)$ ✓

We have the following training set:

	x_1	x_2	label
a	2	8	Pos
b	3	7	Pos
c	5	6	Neg
d	6	5	Neg
e	8	4	Pos
f	9	2	Pos
g	1	1	Neg

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

We use a linear classifier defined by

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

31. 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
    B f e d c b a g
    C g a b c d e f ✓
    D g f e d c b a

32. How many ranking errors does the classifier make?
    A 3
    B 4 ✓
    C 5
    D 6

33. If we draw a coverage matrix (as done in the slides), what proportion of the cells will be red?

    A $\frac{3}{12}$ **B** $\frac{4}{12}$ ✓ **C** $\frac{3}{16}$ **D** $\frac{4}{16}$

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

$$f(x) = \frac{3\cos x}{2\sin x}$$

with respect to x.

First, we break the function up into modules:

$$\begin{aligned} a &= \sin x \\ b &= \cos x \\ c &= 2a \\ d &= 3b \\ f &= \frac{d}{c} \end{aligned}$$

34. What should be in the place of the dots?
    - **A** $a = \cos x,\ c = 2d,\ f = d/c$
    - **B** $a = \cos x,\ c = 2d,\ f = c/d$
    - **C** $a = \sin x,\ c = 2a,\ f = c/d$
    - **D** $a = \sin x,\ c = 2a,\ f = d/c$ ✓

35. In terms of the *local* derivatives. Which is the correct expression for the derivative $\frac{\partial f}{\partial x}$?
    - **A** $-\frac{2d\cos x}{c^2} - \frac{3\sin x}{c}$ ✓
    - **B** $-\frac{2d\cos x}{c^2} + \frac{3\sin x}{c}$
    - **C** $-\frac{2\cos x}{d} - \frac{3\sin x}{c}$
    - **D** $-\frac{2\cos x}{d} + \frac{3\sin x}{c}$

Consider the following function:

$$\mathbf{y} = \mathbf{x} \otimes \mathbf{v} + \mathbf{b}$$

Here $\mathbf{x}$, $\mathbf{y}$, $\mathbf{v}$ and $\mathbf{b}$ are vectors and $\otimes$ represents element-wise multiplication.

This function is part of a larger computation graph, resulting in a scalar loss l. We want to implement this computation in an automatic differentiation (AD) system, as discussed in the lectures.

36. Work out the scalar derivative of y_i over x_j. Which is the correct solution?

A $\frac{\partial y_i}{\partial x_j} = \begin{cases} v_j & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$ ✓

B $\frac{\partial y_i}{\partial x_j} = \begin{cases} v_j & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$

C $\frac{\partial y_i}{\partial x_j} = \begin{cases} b_j & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$

D $\frac{\partial y_i}{\partial x_j} = \begin{cases} b_j & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$

The AD system will give us a vector $\mathbf{y}^\nabla$, such that $y_i^\nabla = \frac{\partial l}{\partial y_i}$. Using $\mathbf{y}^\nabla$, we want to efficiently compute a vector $\mathbf{x}^\nabla$ such that $x_i^\nabla = \frac{\partial l}{\partial x_i}$.

37. Which operation computes this vector for us?

A $\mathbf{x}^\nabla = \mathbf{y}^\nabla \otimes \mathbf{v} + \mathbf{b}$

B $\mathbf{x}^\nabla = \mathbf{y}^\nabla \otimes \mathbf{v}$ ✓

C $\mathbf{x}^\nabla = \mathbf{y}^{\nabla^\mathbf{T}} \mathbf{v} + \mathbf{b}$

D $\mathbf{x}^\nabla = \mathbf{y}^{\nabla^\mathbf{T}} \mathbf{v}$

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.

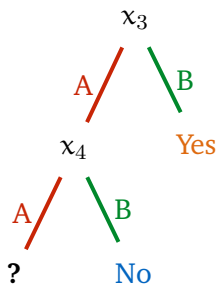
x_1	x_2	x_3	x_4	y
A	A	A	B	Yes
A	A	B	A	Yes
A	A	A	A	Yes
A	A	B	A	Yes
A	B	A	A	Yes
B	B	A	A	No
A	A	A	B	Yes
B	B	B	B	No
B	A	B	B	No
B	B	A	B	No
B	B	B	B	No
B	B	B	A	No

38. 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

39. 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

Consider the following (partial) decision tree:



We can place either x_1 or x_2 on the open node (indicated by the question mark).

40. At this point, what is the information gain for x_2?
A 0
B 0.2516✓
C 0.5
D 0.8753

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