Deep Learning 2023

Final Exam

WITH ANSWERS

21 December 2023, 12:15–14:30

The questions start on the next page.

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 exam, your total points for the exam should be between 20 and 25. The exact pass mark will be decided *after* the exam results have been analysed.

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 a single A4 cheat sheet. Everything on the cheat sheet needs to be written by hand.

1. We can use gradient descent to minimize a loss function. Which statement is **true**?
   A Gradient descent takes steps in the steepest descent direction, and this direction is the opposite of the gradient. ✓
   B Gradient descent takes steps in the steepest ascent direction, and this direction is the gradient.
   C Gradient descent takes steps in the steepest descent direction, and this direction is the gradient.
   D Gradient descent takes steps in the steepest ascent direction, and this direction is the opposite of the gradient.

2. 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.

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. 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.

5. Which of the following statements is **false**?
   A A convolution is equivariant to rotation. ✓
   B A convolution is equivariant to translation.
   C Training CNNs with permutations of data may allow for better robustness against transformations of input data.
   D A convolution is not equivariant to scaling.

6. Which of the following operations in a transformer block propagates information across the time dimension?
   A Self-attention ✓
   B Feed-forward layer
   C Layer normalization
   D Residual connection
7. What type of model/method results from only keeping the *decoder* part of an autoencoder?
   A Data manipulator
   B Dimensionality reduction
   C Generative model ✓
   D Transformer
8. Which of the following statements applies to lazy execution:
   a: It is relatively easy to debug because problems in the model occur as the model is executing.
   b: There are many possibilities for optimization.

   A Only statement a.
   B Only statement b. ✓
   C Both statements a and b.
   D Neither statement applies.
9. What is *numeric* differentiation?
   A Building up a computation graph in memory to run backpropagation on.
   B Manipulating mathematical expressions to obtain the derivative as a function independent of the value of the input.
   C Estimating the derivative of a network from two or more forward passes with slightly different inputs. ✓
   D All of the above.
10. What is *symbolic* differentiation?
    A Building up a computation graph in memory to run backpropagation on.
    B Manipulating mathematical expressions to obtain the derivative as a function independent of the value of the input. ✓
    C Estimating the derivative of a network from two or more forward passes with slightly different inputs.
    D All of the above.
11. What is automatic differentiation?
    A Building up a computation graph in memory to run backpropagation on.✓
    B Manipulating mathematical expressions to obtain the derivative as a function independent of the value of the input.
    C Estimating the derivative of a network from two or more forward passes with slightly different inputs.
    D All of the above.

12. What do we call the two terms of the ELBO?
    A Reconstruction error and regularization ✓
    B Recombination error and regularization
    C Recombination error and reorganization
    D Reconstruction error and reorganization

13. When creating embeddings, we want the embedded space to have a meaningful structure. But, what does that mean?
    A The embeddings in the embedded space must have the same spatial structure as the input objects; e.g., a vector for audio, a rectangle for a single-channel image, and so on.
    B The embeddings must encode all information present in the original objects, i.e., it must be a lossless encoding.
    C The embeddings must intelligibly preserve the content of the objects, e.g., the embedding of an image must look like a scaled-down version of that image.
    D Certain directions in the embedded space correspond to features in the represented objects. ✓

14. Which of the following statements are true about recurrent neural networks (RNNs)?
    1. RNNs are causal.
    2. RNNs have a bounded memory.

    A Only 1 ✓
    B Only 2
    C Both 1 and 2
    D Neither 1 nor 2

15. Which of the following statements are true about temporal convolutional networks (TCNs)?
    1. TCNs are causal.
    2. With TCNs, dilation is a solution for many weights of large filters.

    A Only 1
    B Only 2
    C Both 1 and 2 ✓
    D Neither 1 nor 2

16. In transformers for natural language, we add multiple heads to our self-attention, so the model
    A can learn multiple ways in which two words may relate to each other. ✓
    B has more parameters to learn the word meanings.
    C can parallelize the learning process.
    D has key, query and value vectors that differ from one another.

17. Assume that we have a classification task with 5 classes. We build a network with raw output nodes o_i, to which we apply a softmax, yielding outputs y_i, from which we compute a cross-entropy loss l. Which is **true**?

    A It would be better to use a sigmoid activation instead of a softmax, since its backwards pass is easier.

    B It is not necessary to work out $\frac{\partial l}{\partial o_i}$ since we can apply the chain rule. ✓

    C It is not necessary to work out $\frac{\partial y_i}{\partial o_j}$ since we can apply the chain rule.

    D The backwards step of the cross-entropy loss is not included, since it is a linear function.

18. We have a logistic regression model for a binary classification problem, which outputs 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: *log-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 probability so that we can search for a good model using gradient descent.

    C Log-loss is equal to the binary cross-entropy. ✓

    D Log-loss is proportional but not equal to the binary cross entropy.

19. In machine learning, sometimes we want to make use of sequential data. Which of the following is true?

    A Transformers are a special type of recurrent neural network.

    B You need RNNs to learn from sequences of arbitrary length.

    C Due to the symmetries of the dot-product, transformers can not distinguish between two permuted sequences.

    D GRUs have fewer gates than LSTMs. ✓

20. Machine learning on graphs has some known complications. Which of the following is **not** such a complication?

    A Different types of graphs need different processing.

    B Graphs are not suitable in settings with many different relation (edge) types. ✓

    C Traditional graph-learning methods exhibit scaling issues.

    D Information can not spread via message passing between disconnected graph components.

21. We need to calculate derivatives for many machine learning setups to work. Which of the following statements is **true**?

    A Backpropagation starts with applying the chain rule to the *local* derivative of the loss with respect to the parameters.
    B Backpropagation starts with applying the chain rule to the *global* derivative of the loss with respect to the parameters. ✓
    C For tensor backpropagation of a module in our computation graph, we work out the derivative of an input tensor with respect to the output tensor.
    D For tensor backpropagation of a module in our computation graph, we work out the derivative of an output tensor with respect to the input tensor.

22. What statement is **not true** about walk-forward validation?
    A It simulates the performance of a model trained at a specific time.
    B It does not allow evaluation on different data sizes. ✓
    C It prevents the leakage of future data into validation and test set.
    D It keeps the data ordered by time.

23. What is the output dimension of an image with a width of 5 and height of 5, if two filters with kernel size (3×3), a stride of 2 and a padding of 1 are applied to this image (assuming channel, height, width ordering)?
    A $(2 \times 3 \times 3)$ ✓
    B (3×3)
    C $(2 \times 2 \times 2)$
    D (2×2)

24. I want to define a single convolution operation for images. Which information do I *not* need when setting up this model?
    A The kernel size.
    B The height and width of the input. ✓
    C The stride.
    D The padding.

25. I am training a generator network to generate faces. I take a random sample, compare it to a randomly chosen image form 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 Diffusion
    C The information bottleneck
    D Mode collapse ✓

26. What is an important difference between regular recurrent neural networks (RNNs) and LSTM neural networks?

    A All LSTMs have a *forget gate*, allowing them to ignore parts of the cell state right away. ✓

    B All RNNs have a *forget gate*, allowing them to ignore parts of of the cell state right away.

    C LSTMs have a vanishing gradient problem, RNNs don't.

    D RNNs can be turned into variational autoencoders, LSTMs can't.

27. There are several benefits to using Gaussian noise in a diffusion model. Which is *not* one?

    A The KL divergence has a closed-form expression for Gaussians.

    B With Gaussians we can use an exact form of the maximum likelihood loss, instead of a lower bound.✓

    C Gaussian noise allows us to directly compute z_t without going through t diffusion steps.

    D With Gaussian noise, under the right assumptions, our noise can be reduced to a squared Euclidean distance.

28. Which answer contains methods that can *all* be used as sequence-to-sequence layers?

    A Convolutions, RNN, LSTM ✓

    B Gradient boosting, LSTM, Deep Q-Learning

    C Convolutions, Word2Vec, Spherical Gaussian

    D RNN, Deep Q-Learning, Word2Vec

29. In graph embedding methods, we can define the score of an embedding for a triple (s, p, o) as how close the sum of the embeddings for s and p is to the embedding for o. What's this approach called?

    A Tensor factorization

    B Random walk-based embedding

    C RGCN

    D TransE ✓

30. In transformers, what is a benefit of position *encodings* over position *embeddings*?

    A Unlike embeddings, they break the permutation equivariance of self-attention.

    B Unlike embeddings, make self-attention permutation equivariant.

    C They can, in theory, generalize to sequences that are longer than any in their training data.✓

    D They can, in theory, generalize to sequences that have a larger vocabulary of tokens than the model has seen in the training data.

31. 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 a discriminator network pairs of inputs.
    D By learning the latent representation of an instance through an "encoder" network.✓

32. What does "model-agnostic" mean in the context of Explainable AI techniques?
    A The technique is guaranteed to work equally well on all models.
    B The technique can be applied to any type of machine learning model. ✓
    C The technique does not improve the model's recall, only its precision.
    D The technique does not improve the model's precision, only its recall.

33. What is the purpose of a *prototype instance*?
    A To visualize the decision boundaries of a model.
    B To provide a representative example of a particular class or concept. ✓
    C To measure the accuracy of a model.
    D To identify the most important features for making predictions.

34. How does LIME (Local Interpretable Model-agnostic Explanations) achieve interpretability in complex machine learning models?
    A By directly modifying the underlying neural network architecture.
    B By training a separate global model to interpret local predictions.
    C By perturbing input instances and learning a locally faithful interpretable model. ✓
    D By using reinforcement learning to optimize feature importance weights.

35. What is the Information Bottleneck principle in the context of machine learning?
    A A method for compressing training data to reduce storage requirements.
    B A principle that aims to minimize the information loss between input and output while compressing irrelevant features. ✓
    C An algorithm for maximizing the mutual information between input and extracted features.
    D A technique to selectively eliminate irrelevant features from the dataset during preprocessing.

36. How does the Information Bottleneck principle contribute to the training of neural networks?

A By decreasing the computational complexity of the training process.
B By enforcing strict regularization, it restricts the expressiveness of neural networks during training.
C By encouraging the extraction of a minimum sufficient statistic from input data for improved generalization. ✓
D By encouraging the model to memorize training data.

37. Among the following statements regarding Maximum Mean Discrepancy (MMD), which one is **false**?

A MMD is a measure of dissimilarity between two probability distributions in a reproducing kernel Hilbert space (RKHS).
B The "kernel trick" allows MMD to efficiently operate in high-dimensional spaces without explicitly mapping data points.
C MMD is insensitive to the choice of kernel function, as it always uses a fixed, predefined kernel. ✓
D Given a Gaussian kernel, MMD matches all orders of statistics of samples from two probability distributions.

38. In the context of Kullback-Leibler (KL) Divergence, consider two probability distributions, p and q. If the KL Divergence is computed as $D_{KL}(p; q)$, which statement accurately characterizes the interpretation of the result?

A $D_{KL}(p; q)$ is always symmetric regardless of the nature of distributions p and q.
B $D_{KL}(p; q)$ measures the information lost when using distribution q to approximate distribution p. ✓
C $D_{KL}(p; q)$ is unaffected by the differences in the means and variances of p and q.
D $D_{KL}(p; q)$ satisfies the triangle inequality.

39. In the context of deep neural networks, what is the *generalization error*?

A The error incurred during the training phase, measuring the discrepancy between predicted and actual values.
B The discrepancy between the model's performance on the training data and its ability to generalize to new, unseen data. ✓
C A metric exclusively applicable to shallow neural networks, representing the model's capability to generalize across different tasks.
D The error introduced by model complexity, directly proportional to the number of parameters in the neural network.

40. In deep reinforcement learning, what sets Policy Gradient methods apart from value-based approaches?

A Policy Gradient methods focus on deterministic policies and prioritize minimizing temporal differences in sparse reward environments.

B Policy Gradient methods directly optimize the policy function using the gradient of expected cumulative rewards. They excel in handling high-dimensional action spaces and addressing exploration-exploitation challenges. ✓

C Policy Gradient methods are limited to tabular representations, making them less suitable for continuous environments. They prioritize maximizing state values for efficient learning.

D Policy Gradient methods and value-based approaches share similar optimization objectives, differing only in their implementation. They show comparable performance across various reinforcement learning tasks.

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