Exam for ML for Physicists 2022/2023

Name/Sciper:

Instructions:

- Duration of the exam: 3 hours, 19.01.2023 from 09h15 to 12h15. Room PO01.
- Material allowed: 2 pages (i.e. one sheet recto-verso or 2 one-sided sheets) of personal notes. Pen and paper.
- Problems can be solved in any order.
- Write your full name on **each** additional sheet of paper you hand in.
- Total number of points is 71.

1 Match tools to tasks. [5 points]

Match one of the tools (a)-(e) to each of the tasks 1.-5. in the list below (1 point for each correct match):

Tools:

(a) Supervised regression, prediction

(b) Supervised regression, estimation

(c) Supervised classification

(d) Unsupervised clustering

(e) Unsupervised generative models

Tasks:

1. Identify 3 different groups of students with similar lunchtime habits based on their camipro payment times.
2. A data center wants to balance its power use. Based on a time series of its past power usage, the engineers want to predict future power use. This way, they can smartly schedule power-intensive computational jobs at lower power usage times.
3. An industrial greenhouse grows tomatoes. It has a reference dataset for what “ripe” and “unready” tomatoes look like. The automated video-enabled picking machine is supposed to determine whether the tomatoes in the greenhouse are ready to be picked or not.
4. From the measurements of a magnetic resonance scanner, you want to reconstruct the image of the patient’s brain.
5. You discovered a big set of diaries from your great grandma in your parent’s basement. You want to automatically generate new texts based on these.

2 Urns and balls [6 points]

1. (2 points) Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random, urn A with probability p, urn B with probability $1 - p$, and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?

2. (2 points) Urn A contains five balls: one black, two white, one green and one pink; urn B contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, and 10 purple. [One fifth of A's balls are black; two-fifths of B's are black.] One of the urns is selected at random, urn A with probability p, urn B with probability $1 - p$, and one ball is drawn. The ball is black. What is the probability that the urn is urn A?

3. (2 points) The inhabitants of an island tell the truth one third of the time. They lie with probability 2/3. On an occasion, one of them (called Alice) tells you a statement. You ask another of them (called Bob) 'Did Alice tell the truth?' and Bob answers 'Yes'. What is the probability that Alice told the truth?

3 Why this particular loss? [12 points]

Consider a regression problem on a dataset with inputs $X \in \mathbb{R}^{n \times d}$ and labels $y \in \mathbb{R}^n$ being solved by minimizing the following loss function over binary weights $w \in \{\pm 1\}^d$:

$$L(w) = \frac{1}{n} \sum_{\mu=1}^{n} \left(y_\mu - \sum_{i=1}^{d} X_{\mu i} w_i \right)^2 \tag{1}$$

1. (3 points) Give an example of an algorithm one can use to estimate the minimizer of the loss. Make sure you take into account that $w \in \{\pm 1\}^d$.

2. (4 points) State the two main assumptions about the data-generative process that are implicitly being made when using such a loss function. Define a generative model of data that would justify using such a loss function.

3. (4 points) Given the generative model you wrote, does the loss above correspond to the maximum likelihood estimator, the maximum a posteriori, or the minimum mean-squared error (MMSE) estimator? Write how do the other two estimators look like for the same problem (the simpler of the two should be given in an explicit form).

4. (1 point) Name one example of an algorithm for the MMSE estimator.

4 Sampling with a field [10 points]

Consider the following probability distribution

$$P(S) = \frac{1}{Z} e^{-\beta \sum_{i<j}\left(J_{ij}S_iS_j - \frac{1}{2\sqrt{N}}S_i^2S_j^2\right) - h\sum_{i=1}^{N} S_i^2},$$

where the variables $S_i \in \{\pm 1, 0\}$, $i = 1, \dots, N$.

Recall the spin-glass card-game that we defined in the course. Consider a variant with card values $S_i \in \{\pm 1, 0\}$. We still want to reconstruct the values of cards $S_i^* \in \{0, \pm 1\}$ based on the observation of a symmetric matrix $J_{ij} = S_i^* S_j^*/\sqrt{N} + \xi_{ij}$ with Gaussian noise $\xi_{ij} \sim \mathscr{N}(0, \Delta)$. We know additionally that the probability that a card in the deck is zero is $1 - \rho$ and the probabilities that it is $+1$, -1 are respectively equal to $\rho/2$, $\rho/2$.

1. (4 points) Write the relation between the above probability distribution and the posterior probability distribution in spin-glass card-game defined above. Concretely, in the context of Bayes-optimal inference (i.e. when computing the MMSE estimator), how is β related to Δ, and h to ρ?

2. (3 points) How does one compute the optimal MMSE estimator of the ground-truth values of the cards S_i given the knowledge of J, Δ, ρ?

3. (3 points) Write the Metropolis rule for the MCMC sampling the probability distribution $P(S)$ where we try to change one spin S_i at a time.

5 Maximum entropy distribution [5 points]

(5 points) Consider a random variable $x \in \mathbb{R}$. We observe its mean μ and variance Δ from the data. Derive the probability distribution $P(x)$ that maximizes the entropy given the constraint that the mean and variance are equal to the observed ones.

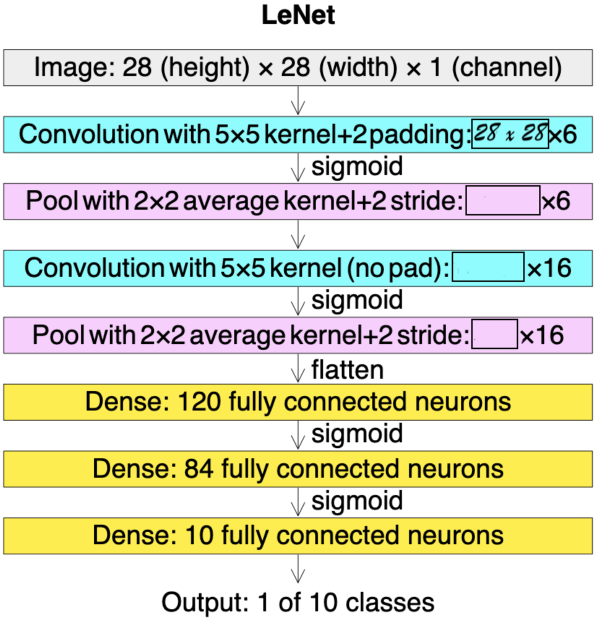
6 Kernels and feature maps [6 points]

- (3 points) Consider $x, x' \in \mathbb{R}^2$, what is a feature map for the kernel $k(x, x') = (x^\top x' + 1)^2$? What is the dimensionality of that feature space?

- (3 points) Write the loss function that kernel regression minimizes (you do not need to consider any regularization).

7 Deep learning architecture [6 points]

1. (6 points) What type of architecture is represented in the following scheme? What is the main advantage of this architecture compared to fully connected neural networks? On the side of the blank squares, write clearly the dimension of the representation after the application of the layer, using the convention Height × Width × Number of channels.



LeNet

Image: 28 (height) × 28 (width) × 1 (channel)

Convolution with 5×5 kernel+2 padding: 28 x 28 ×6

sigmoid

Pool with 2×2 average kernel+2 stride: ×6

Convolution with 5×5 kernel (no pad): ×16

sigmoid

Pool with 2×2 average kernel+2 stride: ×16

flatten

Dense: 120 fully connected neurons

sigmoid

Dense: 84 fully connected neurons

sigmoid

Dense: 10 fully connected neurons

Output: 1 of 10 classes

8 Course questions: [21 points]

1. (3 points) Classify the following quantities as parameters (P) or hyperparameters (H) of a learning algorithm.
   - The weights of a neural network
   - The number of clusters in clustering
   - The learning rate of stochastic gradient descent
   - The number of principal components in a PCA model
   - The class assigned to each sample in a k-means algorithm
   - The number of hidden layers in a neural network
2. (2 points) Given a matrix $X \in \mathbb{R}^{n \times d}$ describe how to compute a rank k matrix $\hat{X}$ that minimizes the quantity

$$\sum_{\mu=1}^{n} \sum_{i=1}^{d} \left(X_{\mu i} - \hat{X}_{\mu i} \right)^2 . \tag{6}$$

3. (3 points) Write the function represented by a feed-forward fully connected neural network with one hidden layer used for regression. Input size d, size of the hidden layer p. Use the tanh activation function in the hidden layer and a bias term in the hidden layer. Draw a schema of the same network for $d = 3$, $p = 4$.

4. (2 points) What is the reason we use the validation error in addition to the training and test error?

5. (2 point) Why is the activation function $\varphi(x) = \text{sign}(x)$ not suitable for neural networks trained by back-propagation?

6. (2 point) Why is the activation function $\varphi(x) = x$ not suitable for neural networks?

7. (2 points) Describe briefly the two steps that are iterated in the k-means clustering algorithm.

8. (2 points) Explain the difference between Stochastic Gradient Descent (SGD) and Gradient Descent (GD). What is the main advantage of using SGD instead of GD?

9. (3 points) How do you set up the last layer of a neural network to do multi-class classification? Write the loss that one typically uses to train it.