Exam for ML for Physicists 2021/2022

Name/Sciper:

Instructions:

- Duration of the exam: 3 hours, 26.01.2022 from 16h15 to 19h15. Rooms CM1120, CM1121.
- Material allowed: 2 pages (i.e. one sheet recto-verso or 2 one-sided sheets) of personal notes. Calculator. 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 65.

Problem A: Lighthouse Problem [11 points]

A lighthouse is placed close to a straight coastline at a position α along the shore and a distance β out to the sea (see Fig. 1). The distance β is known, but the position α is not known.

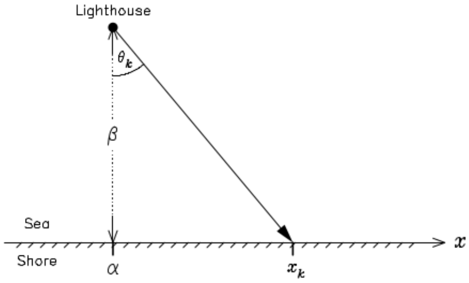


Figure 1: Scheme for the lighthouse problem.

The lighthouse emits a series of short flashes at random intervals and hence in random directions (uniformly at random over the full angle of 2π). These pulses are intercepted on the coast by photo-detectors that record only the fact that a flash has occurred, but not the angle from which it came. N flashes have been recorded so far at positions x_k, $k = 1, \dots, N$.

Goal: From the N observations $x = \{x_k\}_{k=1}^{N}$ estimate the position of the lighthouse α.

1. (1 point) Without writing any probabilities, write a naive estimation of α from data x.

2. (3 points) Prove that the probability $P(x_k|\alpha, \beta)$ of observing a given value x_k conditioned to α, and β is given by the Cauchy distribution

$$P(x_k|\alpha, \beta) = \frac{\beta}{\pi\left[\beta^2 + (x_k - \alpha)^2\right]}.$$

Hint: Recall that if one has a random variable A with probability distribution $P(A)$ and one needs to express a probability distribution of a random variable $B = f(A)$, where f is some known function, one can use $P(B) = \int P(A)\,\delta[B - f(A)]\,\mathrm{d}A$, where $\delta()$ is the Dirac delta function. The integral over B can then be computed by substitution $f(A) \to D$ and using the definition of a Dirac delta that states $g(C) = \int \mathrm{d}D\,\delta(C - D)\,g(D)$ for any test function g.

3. (2 points) What is the mean of the distribution $P(x_k|\alpha, \beta)$ from question 2? Hint: You only need to analyze the corresponding integral when $|x|$ is very large. What does this imply about the naive estimator from question 1.?

4. (3 points) Use the Bayes formula to express the probability $P(\alpha|\beta, x)$ that α has a given value conditioned on the value of β and the N observations x.

5. (2 points) Write the maximum likelihood estimator α^{ML} for α via an equation in which α^{ML} is a root. Use your calculator to decide which of the $\alpha = 2.12; 2.50; 2.85, 3.21$ is the closest to the α^{ML} for $\beta = 1$, $N = 4$, $x = \{1.01; 3.10; 3.27; 2.67\}$.

Problem B: Phase Retrieval [11 points]

Consider the following probabilistic model, input data $X \in \mathbb{R}^{n\times d}$, n samples in d dimensions. The labels are generated as follows

$$y_\mu = \left|\sum_{i=1}^{d} X_{\mu i} w_i^*\right| + \xi_\mu \,,$$

where $\xi_\mu \sim \mathscr{N}(0,\Delta)$ is Gaussian additive noise and $w_i^* \sim \mathscr{N}(0,\sigma^2)$ are the ground truth weights.

The statistician observes the data $\{X_\mu \in \mathbb{R}^d, y_\mu\}_{\mu=1}^n$ and aims to recover back the ground truth weights $w^* \in \mathbb{R}^d$.

1. (2 points) Write the posterior distribution for this problem.

2. (1 point) Write the loss corresponding to the maximum likelihood estimator.

3. (1 point) Write the loss corresponding to the maximum a posterior (MAP) estimator.

4. (2 points) Give one example of an algorithm you could use to compute the MAP estimator.

5. (2 points) Write the minimum mean-squared error (MMSE) estimator.

6. (3 points) Give one example of an algorithm you could use to compute the MMSE estimator.

Problem C: Kawasaki dynamics [5 points]

Consider you need to sample from the following probability distribution

$$P(S) = \frac{1}{Z} e^{-\beta \sum_{i<j} J_{ij} S_i S_j} ,$$

where the variables $S_i \in \{\pm 1\}$, $i = 1, \dots, N$. In this problem, the goal will be to sample at fixed magnetization $m = \sum_{i=1}^{N} S_i/N$, i.e. only configurations that have the same magnetization as the initial one. In order to do that we use the so-called Kawasaki dynamics where the proposed step is to exchange two spins of the opposite sign.

1. (2 points) Recall the spin glass card game that we defined in the course, where we want to reconstruct the values of cards S_i^* based on observation of a symmetric matrix $Y_{ij} = S_i^* S_j^* / \sqrt{N} + \xi_{ij}$ with Gaussian noise $\xi_{ij} \sim \mathcal{N}(0, \Delta)$. Why can it be of interest to sample $P(S)$ at fixed magnetization in the context of solving the spin glass card game problem?

2. (3 points) Write the Metropolis rule for the Kawasaki dynamics defined above.

Problem D: 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. Find families of illnesses from gene expression data related to each illness.
2. From past data about loans and data about customers to whom a bank gave a loan in the past, estimate the risk that a new customer will not properly pay their loan.
3. Based on database of texts in different languages learn to recognize the language of a sentence.
4. Reconstruct the image of lungs in the nuclear magnetic resonance imaging.
5. Based on a dataset of human-produced music, teach computers to produce new pieces of music.

E. Kernel from a feature map [4 points]

Consider $d = 2$ dimensional data, and the following feature map to $p = 3$ dimensional feature space: $\mathbf{x} = (x_1, x_2) \to (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$.

- (2 points) What does a kernel function $K(\mathbf{x}_\mu, \mathbf{x}_\nu)$ calculate?

- (2 points) Write the expression that corresponds to the kernel defined by the above feature map.

F. Course questions: [20 points]

- (3 points) Classify the following learning methods as supervised (S) or unsupervised (U):
  - k-Nearest Neighbors
  - Support Vector Machines
  - Principal Component Analysis
  - Multi-layer Neural Network
  - Mixture of Gaussian Clustering
  - Auto-encoder
- (2 points) Write the probability distribution that a restricted Boltzmann machine uses to generate new data samples.

- (2 points) Write a suitable loss function for sparse linear regression.

- (2 points) Describe at least one method for dimensionality reduction.

- (2 points) What is the main idea behind simulated annealing?

- (2 points) How does the stochastic gradient descent algorithm work?

- (3 points) Write the function represented by a feed-forward fully connected neural network with two hidden layers used for supervised classification into k classes. Input size d, size of the first hidden layer p, size of the second hidden layer r. Use the ReLU activation function in both the hidden layers. Draw a schema of the same network for $k = 3$, $d = 3$, $p = 4$, $r = 2$.

- (4 points) Draw the scheme of a multi-layer autoencoder, describe how it works, and mention one possible application.

G. Edge detecting CNN [5 points]

In this exercise, you will craft a convolutional neural network f that detects edges in $1D$ inputs. Let the inputs be $x \in \{0,1\}^d$ of the form $x_k = (\underbrace{1,1,\dots,1}_{k \text{ times}},0,\dots,0), \quad k \in \{0,\dots,d\}$. For this input we say that the edge (i.e. where we switch from 1 to 0) is at position k. The complete CNN will take x_k as input and output $f(x_k) = k$.

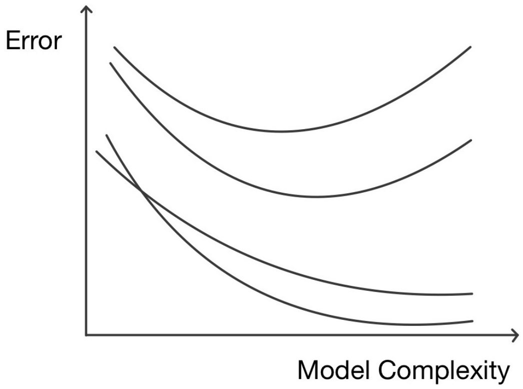
1. (3 points) Build a convolutional layer $A : \{0,1\}^{d+2} \mapsto \{0,1\}^{d+1}$ such that $A(\tilde{x}_k)_l = \delta_{kl}$, in other words the only nonzero output should correspond to the position of the edge. $\tilde{x}$ here is a padded version of x: specify how to make the padding. To build A use a single filter of size 2, $a = (a_1, a_2)$.

2. (2 points) Now define $f(x) = w^T A(\tilde{x})$. Find a suitable vector $w \in \mathbb{R}^{d+1}$ such that $f(x_k) = k$.

H. Bias-variance trade-off [4 points]

Assume that you have two data sets that contain i.i.d. samples from the same distribution, call them S1 and S2. S1 contains 5000 samples, whereas S2 contains 100000 samples. You randomly split each of the data sets into a training and a testing set, where eighty percent of the data is assigned to the training set. You then train and test on a family of predictors of increasing complexity.

1. (2 points) In the figure below, there are four curves, two that show the training error as a function of the model complexity (for S1 and S2) and two that show the test error as a function of the model complexity (for S1 and S2). Label each of the 4 curves clearly (train/test error, S1/S2).



2. (2 points) Draw and label the same curves for S1, S2 with the double-descent phenomenon.