Machine Learning — Final Exam

1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
5	7	6	7	5	6	8	7	3	6	9	8	**??**

Do not write anything above this line

Name:

Student ID: Signature:

- Only write on the sheets given to you by supervisors. If you need more paper, ask the supervisors.
- Pages 15-18 can be used as scratch paper.
- All sheets (including scratch paper) have to be returned at the end.
- **Do not unstaple the sheets!**
- Wherever answer boxes are provided, please write your answers in them.
- Please write your student ID (*Matrikelnummer*) on every sheet you hand in.
- **Only use a black or a blue pen (no pencils, red or green pens!).**
- You are allowed to use your A4 sheet of handwritten notes (two sides). **No other materials (e.g. books, cell phones, calculators) are allowed!**
- Exam duration - 120 minutes.
- This exam consists of **??** pages, **??** problems. You can earn **??** points.

1 Decision Trees

Problem 1 [(1+4)=5 points] You are developing a model to classify games at which machine learning will beat the world champion within five years. The following table contains the data you have collected.

x_1 (Team or Individual)	x_2 (Mental or Physical)	x_3 (Skill or Chance)	y (Win or Lose)
T	M	S	W
I	M	S	W
T	P	S	W
I	M	C	W
T	P	S	L
I	M	C	L
T	P	C	L
T	P	C	L
T	P	C	L
I	P	S	W

You can look up the value of $\log_2(x)$ in this table:

x	0.10	0.2	0.25	0.33	0.50	0.66	0.75	0.8	1.0
$\log_2(x)$	-3.32	-2.32	-2.0	-1.60	-1.0	-0.60	-0.42	-0.32	0.0

a) Calculate the entropy $i_H(y)$ of the class labels y.

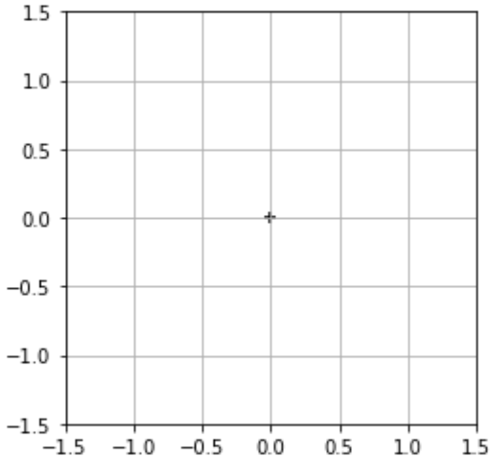
b) Build the optimal decision tree of depth 1 using entropy as the impurity measure. Which attribute is selected as the root of the decision tree?

2 KNN

Problem 2 [(1.5+5.5)=7 points]

a) Let $\boldsymbol{x} \in \mathbb{R}^2$. Draw the unit circle for the following norms. Make sure to clearly label which circle corresponds to which norm.

- L_1-norm: $||\boldsymbol{x}||_1 = \sum_i |x_i|$
- L_2-norm: $||\boldsymbol{x}||_2 = \sqrt{\sum_i x_i^2}$
- L_∞-norm: $||\boldsymbol{x}||_\infty = \max_i |x_i|$



b) Construct a binary classification dataset that consists of 4 data points, that is specify $\boldsymbol{x}_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$ (i.e. write the coordinates and labels) for each data point i, such that:

Performing leave-one-out cross validation (LOOCV) with a 1-NN (one nearest neighbor) classifier using $\underline{L_1 \text{ distance}}$ yields 0% misclassification rate. Meanwhile, performing LOOCV with a 1-NN classifier using $\underline{L_\infty \text{ distance}}$ on the same dataset yields misclassification rate of 50%.

Hint: Remember the shape of the unit circles.

3 Probabilistic Inference

Problem 3 [(6)=6 points] A kangaroo starts from a random location $\boldsymbol{x}_0 \in \mathbb{R}^2$ in the jungle and after one jump reaches the location $\boldsymbol{x}_1 \in \mathbb{R}^2$.

The prior over the start location $\boldsymbol{x}_0$ is the standard bivariate normal distribution

$$p(\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_0 \,\middle|\, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{I}_{2\times 2}\right),$$

where $\boldsymbol{I}_{2\times 2}$ denotes the 2 by 2 identity matrix.

The conditional distribution $p(\boldsymbol{x}_1|\boldsymbol{x}_0)$ is a normal distribution with mean $\boldsymbol{x}_0$ and identity covariance

$$p(\boldsymbol{x}_1|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{x}_0, \boldsymbol{I}_{2\times 2})$$

Assume that we observe $\boldsymbol{x}_1$, the position of the kangaroo after the jump.

Write down the closed-form expression for $p(\boldsymbol{x}_0|\boldsymbol{x}_1)$. Make sure that you obtain a valid probability distribution (i.e. it integrates to one). Show your work.

Important: You are not allowed to simply use facts about conditionals of multivariate normal distribution (e.g. from Bishop's book). Derive the result starting from the Bayes formula.

4 Regression

Problem 4 [(7)=7 points] Consider the following one-dimensional regression problem:

$$p(y_i \mid w, x_i, \tau) = \mathcal{N}(y_i \mid wx_i, \tau^2)$$
$$p(w \mid \sigma) = \mathcal{N}(w \mid 0, \sigma^2)$$

where $x_i, y_i \in \mathbb{R}$, and $\tau > 0$ and $\sigma > 0$ are variance parameters. You fit the parameter w using, e.g., the MLE or the MAP approach on a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_N)\}$ of N i.i.d. instances.

Your task is to qualitatively describe what happens to the quantities in each column as the parameters in each row vary. Specifically, in each cell of the table below you have to write one and only one of the following three options: *(i) increases*, *(ii) decreases* or *(iii) no change.*

For example in the top left cell you have to specify whether the quantity $\text{Var}(p(w|\sigma))$ increases, decreases or does not change as we increase the value of the parameter σ.

	$\text{Var}(p(w\|\sigma))$	$\|w_{MLE} - w_{MAP}\|$	$\left\|\mathbb{E}_{p(w\|\mathcal{D})}[w] - w_{MAP}\right\|$
σ increases	increases	decreases	no change
σ decreases	decreases	increases	no change
N increases	no change	decreases	no change

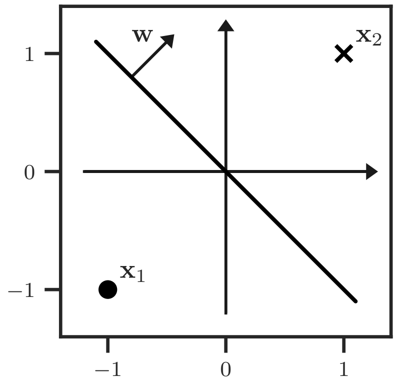
In the table above

- $\text{Var}(p(w|\sigma))$ denotes the variance of the distribution $p(w|\sigma)$
- w_{MLE} and w_{MAP} denote the maximum likelihood estimate and the MAP estimate of the parameter w respectively.
- $\mathbb{E}_{p(w|\mathcal{D})}[w]$ is the expectation of the posterior distribution over w.
- $| \cdot |$ denotes the absolute value.

column 1: 0.5pt. for the first two cells and 1pt. point for the last cell
column 2: 1pt. for each cell
column 3: 2pt. if all cells are correct otherwise 0

Classification

Problem 5 [(5)=5 points] Consider the classification problem in the figure below. There are two points, $\mathbf{x}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$ with class $y_1 = 0$ and $\boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with class $y_2 = 1$. Further you are given a logistic regression model with weight vector $\boldsymbol{w} = \begin{bmatrix} \log 2 \\ \log 2 \end{bmatrix}$. Here, log denotes natural logarithms, i.e. base e. Assume that there is no bias term.



Prove or disprove that the weight vector $\boldsymbol{w}$ is the MAP estimate for a logistic regression model with a Gaussian prior on $\boldsymbol{w}$ with precision $\lambda = 1$.

5 Constrained Optimization

Problem 6 [(3+3)=6 points] Consider the following optimization problem

$$\begin{aligned}\min_{x_1,x_2} 2x_1 - 3x_2\\ x_1 \geq 2\\ x_2 \geq 2\\ x_1 + x_2 \leq 12\\ -x_1 + 2x_2 \geq -3\\ -5x_1 + 3x_2 \leq -4\end{aligned}$$

a) Draw the set of feasible points

b) Solve the optimization problem, i.e. find the minimizer (x_1^*, x_2^*).

1.5pt. for the correct minimizer
1.5pt. for the explanation

6 SVM

Problem 7 [(7+1)=8 points]

a) Consider training a hard-margin SVM using a linear kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$ on a linearly separable training set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_N, y_N)\}$. Let s denote the number of support vectors we would obtain if we would train on the entire dataset. Furthermore, let ε denote the leave-one-out cross validation (LOOCV) misclassification rate.

Does the following inequality hold? Justify your answer.

$$\varepsilon \leq \frac{s}{N}$$

b) Consider a setting similar to the previous problem, except that we now we use SVM with an arbitrary valid kernel k. Assume that the data is linearly separable in the feature space corresponding to the kernel. Does $\varepsilon \leq \frac{s}{N}$ hold in this case? Justify your answer.

7 Kernels

Problem 8 [(7)=7 points] Let $\mathcal{M}$ denote the set of all real-valued matrices of arbitrary size.

Prove or disprove that the function $k : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is a valid kernel.

$$k(X, Y) = \min\{\text{rank}(X), \text{rank}(Y)\}$$

8 Deep Learning

Problem 9 [(1+1+1)=3 points] You are given a dataset containing images $\boldsymbol{x}_i \in [0,1]^D$ (all the pixel values are normalized between 0 and 1) and respective class labels $y_i \in \{1, ..., C\}$. You implement a fully connected neural network with two hidden layers, *tanh* activations and L_2 regularization on all weights excluding biases.

a) Consider two strategies for initializing the weights of your neural network.

   1) Sample the weights from Uniform$(-10, 10)$

   2) Sample the weights from Uniform$(-1, 1)$

   Which choice (1 or 2) is more reasonable, given that we are training the network with backpropagation? Justify your answer.

b) When training neural networks, why do we usually stop training when the loss on the validation set starts to increase?

c) After training has finished, your model has <u>high</u> training loss and <u>high</u> validation loss. What should you do? Justify your answer.

9 Dimensionality Reduction

Problem 10 [(6)=6 points] Let the matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ represent N data points of dimension $D = 10$ (samples stored as rows). We applied PCA to $\boldsymbol{X}$. By using the $K = 5$ top principal components, we transformed/projected $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N \times K}$. We computed that $\tilde{\boldsymbol{X}}$ preserves 70% of the variance of the original data $\boldsymbol{X}$.

Suppose now we apply PCA on the following matrices:

a) $\boldsymbol{Y}_1 = \boldsymbol{X}\boldsymbol{S}$ where $\boldsymbol{S} = \lambda \boldsymbol{I}$, with $\lambda \in \mathbb{R}$ and $\boldsymbol{I} \in \mathbb{R}^{D \times D}$ is the identity matrix

b) $\boldsymbol{Y}_2 = \boldsymbol{X}\boldsymbol{R}$ where $\boldsymbol{R} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$

c) $\boldsymbol{Y}_3 = \boldsymbol{X}\boldsymbol{P}$ where $\boldsymbol{P} = \operatorname{diag}(+5, -5, \dots, +5, -5)$ is a $D \times D$ diagonal matrix

d) $\boldsymbol{Y}_4 = \boldsymbol{X}\boldsymbol{Q}$ where $\boldsymbol{Q} = \operatorname{diag}(1, 2, 3, \dots, D-1, D)$ is a $D \times D$ diagonal matrix

e) $\boldsymbol{Y}_5 = \boldsymbol{X} + \boldsymbol{1}_N \boldsymbol{\mu}^T$ where $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{1}_N$ is an N-dimensional column vector of all ones

f) $\boldsymbol{Y}_6 = \boldsymbol{X}\boldsymbol{A}$ where $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ and $\operatorname{rank}(\boldsymbol{A}) = 5$

and obtain the projected data $\tilde{\boldsymbol{Y}}_1, \dots \tilde{\boldsymbol{Y}}_6 \in \mathbb{R}^{N \times K}$ using the principal components corresponding to the top $K = 5$ largest eigenvalues of the respective $\boldsymbol{Y}_i$.

What fraction of variance of each $\boldsymbol{Y}_i$ will be preserved by each respective $\tilde{\boldsymbol{Y}}_i$? Justify your answer.

The answer "cannot tell without additional information" is also valid if you provide a justification.

10 Gaussian Mixture Models

Problem 11 [(2+5+2)=9 points] Consider two random variables $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^D$ distributed according to two different Gaussian mixture models

$$p(\boldsymbol{x}|\boldsymbol{\theta}^X) = \sum_{k=1}^{K_X} \pi_k^X \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^X, \boldsymbol{\Sigma}_k^X),$$
$$p(\boldsymbol{y}|\boldsymbol{\theta}^Y) = \sum_{l=1}^{K_Y} \pi_l^Y \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_l^Y),$$

The first mixture $p(\boldsymbol{x}|\boldsymbol{\theta}^X)$ consists of K_X components with parameters $\boldsymbol{\theta}_k^X = (\pi_k^X, \boldsymbol{\mu}_k^X, \boldsymbol{\Sigma}_k^X)$ for $k \in \{1, \dots, K_X\}$. Similarly, $p(\boldsymbol{y}|\boldsymbol{\theta}^Y)$ consists of K_Y components with parameters $\boldsymbol{\theta}_l^Y = (\pi_l^Y, \boldsymbol{\mu}_l^Y, \boldsymbol{\Sigma}_l^Y)$ for $l \in \{1, \dots, K_Y\}$.

We generate a new random variable $\boldsymbol{z} \in \mathbb{R}^D$ as $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$.

a) Describe the generative process (process of drawing the samples) for $\boldsymbol{z}$.

b) Explain in a few sentences why $p(\boldsymbol{z}|\boldsymbol{\theta^X}, \boldsymbol{\theta^Y})$ is again a mixture of Gaussians.

c) Write down the probability density function $p(\boldsymbol{z}|\boldsymbol{\theta}^X, \boldsymbol{\theta}^Y)$ of $\boldsymbol{z}$.
It's enough to just state the answer (no need to show the derivation).

11 Variational Inference

The exponential distribution with a scale parameter $\alpha > 0$ is defined as

$$\text{Expo}(\theta \mid \alpha) = \begin{cases} \frac{1}{\alpha}\exp(-\frac{1}{\alpha}\theta) & \text{if } \theta \geq 0 \\ 0 & \text{else} \end{cases}, \qquad \mathbb{E}\left[x\right] = \alpha, \qquad \mathbb{E}\left[x^2\right] = 2\alpha^2$$

Problem 12 [(4+2+2)=8 points] Consider the following probabilistic model.

$$p(z) = \text{Expo}(z \mid 1)$$
$$p(x \mid z) = \mathcal{N}(x \mid z, 1)$$

We want to approximate the posterior distribution $p(z \mid x)$ using a variational distribution

$$q(z \mid \beta) = \text{Expo}(z \mid \beta)$$

a) Write down a closed-form for ELBO $\mathcal{L}(\beta)$ and simplify it as far as you can. You can ignore the terms that are constant in β.

b) Is the ELBO convex in β? Justify your answer.

c) Outline the main steps for solving the optimization problem

$$\min_{\beta} \ \mathbb{KL}(q(z \mid \beta) \parallel p(z \mid x))$$

You don't need to perform the actual computations, just clearly describe each step.

Does this optimization problem have a closed-form solution? Why or why not?