TU Wien

194.025 Einführung in Machine Learning

Examiner: Prof. Dr. Thomas Gärtner

Written Exam, 31.01.2024

Group A

The **5 problems** in this exam yield a total of **60 points**. Passing the exam requires to achieve at least **50% of the points**.

Your name (please use block letters)

Your student ID (e.g., e12345678)

Task 1 General Questions (20 P)

For the following questions, please answer in simple sentences (i.e., 1-2 short sentences) or provide bullet points to the answer.

Q: Please select all pre-processing strategies that you would recommend using for numeric data: (1 P)

- ☐ One-hot encoding
- ☐ Scaling
- ☐ Imputing missing values with a dummy string
- ☐ Gradient descent

(optional) **Justification:**

Q: Please select all pre-processing strategies that you would recommend using for categorical data: (1 P)

- ☐ One-hot encoding
- ☐ Discretise continuous values
- ☐ Imputing missing values with a dummy string
- ☐ Re-label ordinal values

(optional) **Justification:**

Q: With vc($\mathcal{H}$) we denote the VC dimension of the hypothesis space $\mathcal{H}$. Let $X = \mathbb{R}^2$ and $\mathcal{H}$ be the set consisting of all areas bounded by a convex polygon. Does $\mathcal{H}$ have an unbounded VC dimension, that is, vc($\mathcal{H}$) $= \infty$? (2 P)

A:

Q: Consider the following statement: In PAC learning, the $\varepsilon > 0$ parameter is is used to state that the returned classifier h has test loss at least ε. Is it true or false? (1 P)

A:

Q: What is the first *principal component* in PCA? (2 P)

A:

Q: Suppose your data has outliers, how could k-means be negatively affected by this? How would you address this situation? (2 P)

A:

Q: When performing PCA, what happens if the eigenvalues of the covariance matrix have roughly the same value? (2 P)

A:

Q: We are given a coin with probability p and $1 - p$ for getting Heads and Tails, respectively. Say we toss it m times getting k heads. What is the maximum likelihood estimate for p?
(1 P)

A:

Q: What is the number of parameters required to fully specify any possible probability distribution on n Boolean random variables? (1 P)

A:

Q: Why do we need non-linear activation functions for (deep) neural networks? (1 P)

A:

Q: In your own words, what does the universal approximation theorem say? (2 P)

A:

Q: Name five of the most important terms/concepts/variables/functions/etc. in reinforcement learning. (2 P)

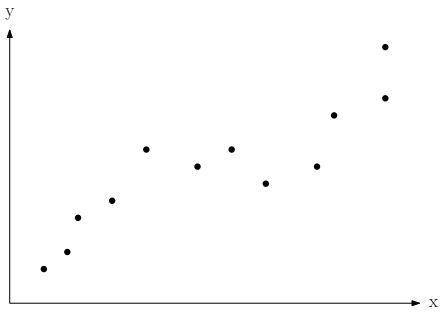
A:

Q: You want to help a robot learn to move around a building, find an electric outlet, recharge when necessary, and pick up trash. Pose the learning problem precisely in terms of reinforcement learning (i.e., rewards, states, environment, transition of states). (2 P)

A:

Task 2 Learning to fit sheep (10 P)

Consider the following data set, where the x-axis describes the weight of a sheep, and the y-axis the diameter of the sheep. Your task is to learn a function which can predict the diameter of a sheep based on its weight in order to find them nicely fitted tuxedos for the ball season.

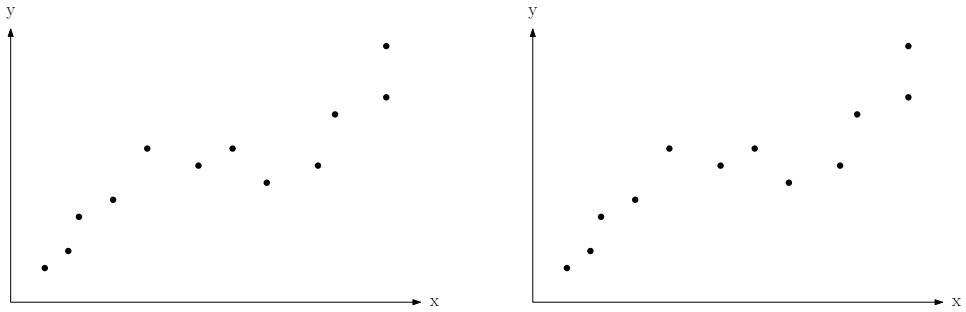


Task 2.1 Learning task (1 P)

What type of learning task is described above?

Task 2.2 Fitting the sheep (4 P)

Please indicate how it would look like if you fit a polynomial of degree 1 (i.e., a line) to the data in the left plot, and a polynomial of arbitrarily high degree which fits the data perfectly in the right plot.



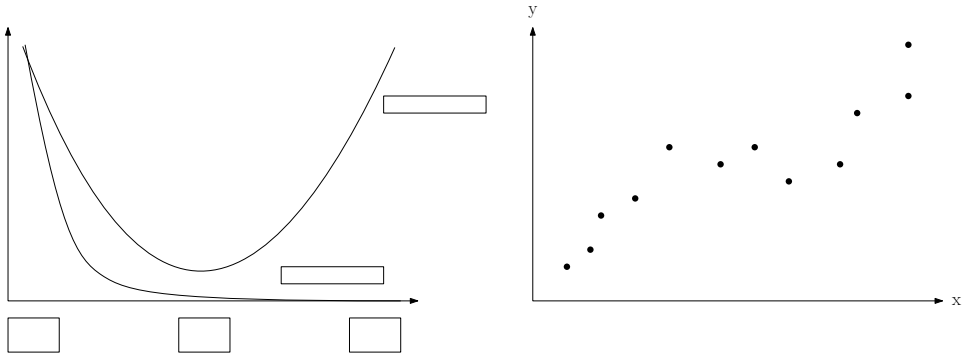
Answer the following questions:

1. Would you say that the degree 1 polynomial is a good predictor? Why or why not?
2. Would you say that the high degree polynomial is a good predictor? Why or why not?

3. How would you assess whether a learned function is a good predictor?
4. Consider the arbitrarily high degree polynomial again. What could you do to obtain a better fit?

Task 2.3 Assessing and improving the fit (3 P)

To make sure we obtain sufficiently well fitted tuxedos for our sheep, we trained multiple machine learning models to fit polynomial functions of different degrees to the data. You can see a summary of the results in the left plot below, where the x-axis describes the model complexity with respect to the degree, and the y-axis depicts the (empirical) risk.



1. Indicate which function graph depicts the **risk**, and which depicts the **empirical risk** (i.e., fill out the blanks in the figure on the left).
2. In each of the three boxes below the x-axis, please indicate what degree a polynomial would likely have to obtain the corresponding risk and empirical risk.
3. Please indicate where overfitting happens, where underfitting happens, and where you would consider a good fit (on the x-axis).
4. Please draw a polynomial which gives a good fit for the data in the figure on the right.

Task 2.4 Risk and empirical risk (2 P)

Answer the following questions:

1. Explain, in your own words, what the risk and the empirical risk are.
2. If we increase the model complexity, how does the behavior of the bias and the variance change?

Task 3 Basic Algorithms (10 P)

Let $D = \{(x_i, y_i)\}_{i=1}^3$ with $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$ be a dataset as follows:

$$(x_i)_{i=1}^3 = \begin{bmatrix} 6 \\ 1.5 \\ 4 \end{bmatrix} \qquad (y_i)_{i=1}^3 = \begin{bmatrix} 0.2 \\ -0.7 \\ -0.2 \end{bmatrix}$$

Task 3.1 Compute the Value of a Loss function on D (1 P)

Let

$$\mathcal{L}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w_1 x_i + w_0)^2$$

be a loss function. Compute $\mathcal{L}(0.2, 1)$ for D.

Task 3.2 Compute the Gradient of the Loss function (2 P)

Compute the gradient $\nabla\mathcal{L}(0.2, 1)$ for D.

Task 3.3 Minimum of the Loss Function (2 P)

For which parameters $\hat{w}_1$ and $\hat{w}_0$ is $\mathcal{L}$ minimal on D? Argue why.

Task 3.4 Matrix Formulation of Linear Regression (3 P)

Recall that one way to solve polynomial regression problems is to compute $\mathbf{w} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$ for suitable $\mathbf{X}$ and $\mathbf{y}$. Using this algorithm, compute $\mathbf{w}$ for

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 3 \\ -4 \\ -2 \end{bmatrix}.$$

Task 3.5 Construct $\mathbf{X}$ for some Dataset (2 P)

Construct the matrix $\mathbf{X}$ for polynomial regression with $d = 3$ on the dataset

$$(x_i)_{i=1}^3 = \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix}$$

such that you can get the optimal parameters $\mathbf{w}$ by computing $\mathbf{w} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$.

Task 4 Evaluation Metrics & Data Set Split (10 P)

In the following tasks, you will be presented with different scenarios and are asked to choose appropriate metrics or compute certain numbers. Please read the scenario carefully to identify the learning task/question.

Task 4.1 Predict Animal Species (3 P)

Scenario: You are given a dataset of 350 pictures. Each picture contains either one cat or one sheep. In total, there are 320 pictures of cats and 30 pictures of sheep. For each picture, **you are given a label that tells you which of the two animals the picture contains**. You create a dataset with the stratified holdout method with an 80% / 10% / 10% split for training / validation / test data.

You have trained a model on a training dataset, that for a given picture predicts the animal shown. Our goal is that our model performs similarly well for both classes of animals. Below, you are given a list of possible metrics.

Tick the metrics that are appropriate for the task. For each metric, provide a short one-sentence explanation of why it is or is not appropriate.

- ☐ Mean squared error:

- ☐ Accuracy:

- ☐ F1 score:

Task 4.2 Predict Animal Height (3 P)

Scenario: You are given a dataset of 350 pictures. Each picture contains either one cat or one sheep. In total, there are 320 pictures of cats and 30 pictures of sheep. For each picture, **you are given a label that tells you the height** of the animal in cm. You create a dataset with the stratified holdout method with an 80% / 10% / 10% split for training / validation / test data.

You have trained a model on the training dataset, that for a given picture predicts the height of the animal.

Tick the metrics that are appropriate for the task. For each metric, provide a short one-sentence explanation of why it is or is not appropriate.

☐ Mean squared error:

☐ Accuracy:

☐ F1 score:

Task 4.3 Hyperparameter Tuning (3 P)

Scenario: You are given a dataset with 1000 samples. Additionally, you are given a model with fixed hyperparameters (meaning that you do **not** have to tune hyperparameters). You intend to train a model with the given hyperparameters and evaluate its performance with different methods.

For the given splitting method, how many models are you **training**? *(Respond with a single integer number per splitting method)*

1. 10-fold cross-validation:
2. Hold-out method with 90% training and 10% test data:
3. Leave-1-out cross-validation:

Task 4.4 Choice of Data Set Split (1 P)

On which of the following split would you tune hyperparameters?

☐ Training set

☐ Validation set

☐ Test set

Task 5 Kernel Methods (10 P)

Recall important properties for kernels from the lecture:

For any symmetric $K \in \mathbb{R}^{n \times n}$ the following are equivalent:

1. K is **positive semi-definite (PSD)**
$$\forall c \in \mathbb{R}^n : c^t K c = \sum_{i,j \in [\![n]\!]} c_i c_j K_{ij} \geq 0$$
2. K can be **factorised**
$$\exists \ell \in \mathbb{N}, F \in \mathbb{R}^{\ell \times n} : K = F^t F$$
3. K has only **non-negative eigenvalues**
$\exists U, D \in \mathbb{R}^{n \times n} : UDU^t = K, U^t U = \mathbf{I}, \forall i \in [\![n]\!] : D_{ii} \geq 0,$
and $\forall i, j \in [\![n]\!] : j \neq i \Rightarrow D_{ij} = 0$

Task 5.1 PSD Matrices? (3 P)

Which of the following matrices are symmetric positive semi-definite?

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \qquad C = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Argue why!

Task 5.2 General Results of the PSD Property (2 P)

Given any symmetric positive-semidefinite Matrix $K \in \mathbb{R}^{123 \times 123}$ are the following matrices

- always,
- never, or
- sometimes but sometimes not

symmetric positive-semidefinite?
(brackets indicate selected range, i.e. rows 45-67 & columns 45-67)

$D = K[45:67, 45:67]$
$E = K[56:78, 67:89]$
$F = -3K[1:3, 1:3]$
$G = KK$

Argue why!

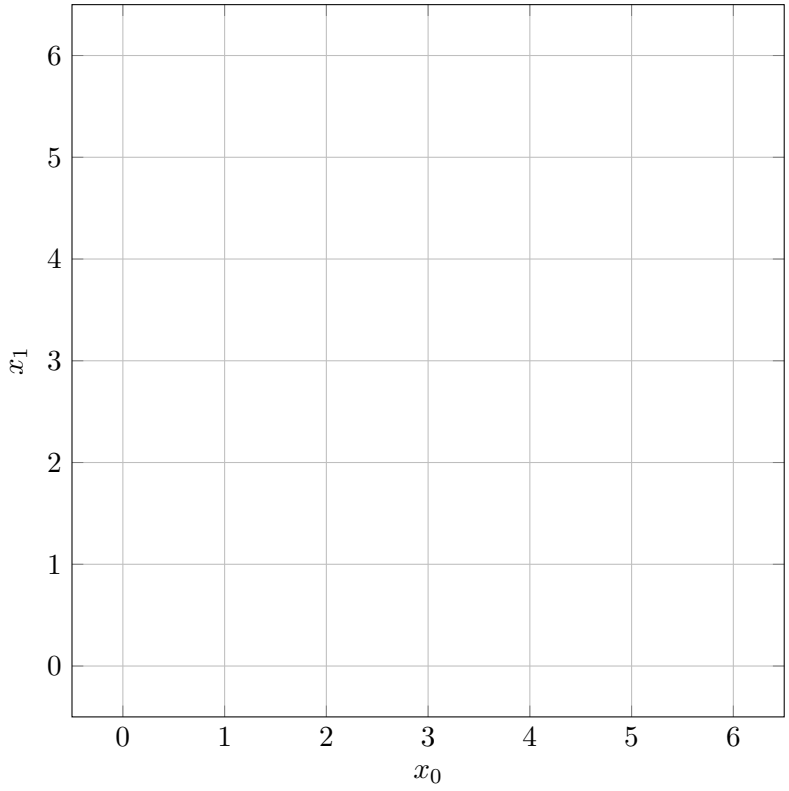
Task 5.3 Maximum-Margin Separation (5 P)

Given

$$D = \left\{ \left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, 1 \right), \left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, -1 \right), \left(\begin{pmatrix} 6 \\ 2 \end{pmatrix}, -1 \right), \left(\begin{pmatrix} 3 \\ 5 \end{pmatrix}, -1 \right), \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, 1 \right) \right\} \subset \mathbb{R}^2 \times \{\pm 1\}$$

what are the support vectors, max-margin seperating hyperplane, and supporting hyperplanes?

Remark: You can use the grid below to plot the data points.



Additional Pages