Exam

Applied Machine Learning Fundamentals

Data Science, WWI 18DS A/B, DHBW Mannheim
Matriculation number:

July 24, 2020

Remarks:

1. Please check whether you have received **all 15 task sheets** (excluding the cover page) **prior** to solving the exam questions. Please turn to the test proctor (‘Pruefungsaufsicht’) in case your exam is incomplete.
2. Do not forget to put your matriculation number on all sheets. **Do not put your name onto the exam (anonymization)!**
3. Write down the answers directly on the task sheets. You may use the empty back sides of the task sheets or the last page of this exam in case you need additional space.
4. You are free to answer the questions either in English or in German. Please do not translate technical terms into German in order to avoid confusion.
5. The exam comprises 60 points in total and has to be solved **within 60 minutes**. The maximum attainable scores per task are always specified. You can use them as an indication for how detailed your answers should be.
6. Please turn off all kinds of communication devices. Only the following auxiliary material is allowed: ❶ Non-programmable calculator ❷ Two-sided hand-written cheat sheet
7. **All violations will be considered attempted cheating!**

Good luck! ☺

1 Miscellaneous

1.1 Tick the correct weight update rule which is used in gradient descent (where $\mathcal{J}$ is the cost function). **(1 p)**

□ $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla \mathcal{J}(\boldsymbol{\theta})$

□ $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla \mathcal{J}(\boldsymbol{\theta})$

□ $\boldsymbol{\theta} \leftarrow \alpha \nabla \mathcal{J}(\boldsymbol{\theta})$

□ All options are incorrect.

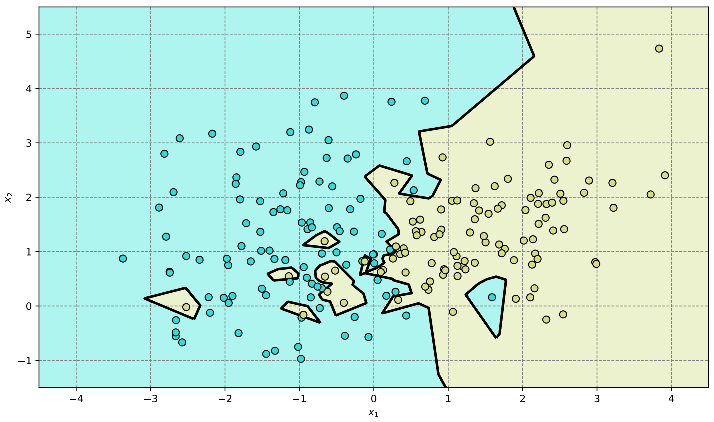
1.2 In gradient descent, what is a suitable value for the learning rate α? What problems do you face when choosing it too low or too high? **(3 p)**

1.3 What is a *gradient* (in the context of machine learning)? **(1 p)**

1.4 You use a k-nearest neighbor classifier and set $k = n$, where n is the total number of data points in the dataset. Which class is predicted by the classifier? **(1 p)**

1.5 What does *'bayes optimal'* mean? **(1 p)**

1.6 What phenomenon is depicted in the image below? Assume the decision boundary was generated by a *k*-nearest neighbor classifier. How can you avoid this problem? **(2 p)**



1.7 Tick the correct statements. **(2 p)**

□ In supervised learning the training data is labeled.

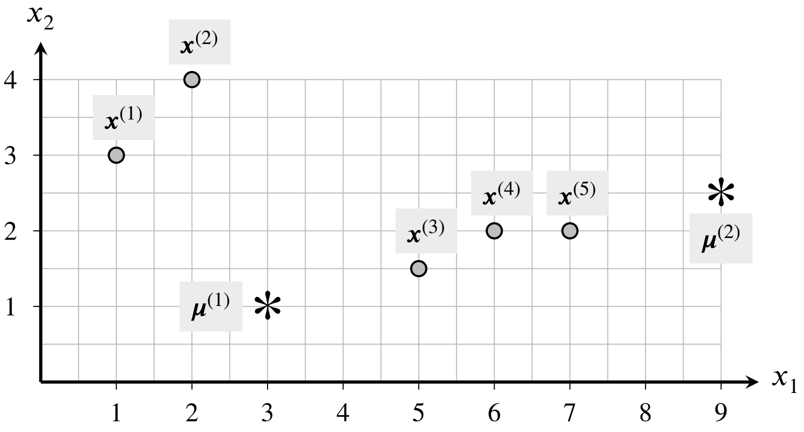
□ In reinforcement learning the agent is told the correct solution after each step.

□ Clustering belongs to the category of unsupervised learning.

□ All options are incorrect.

1.8 You have a dataset consisting of 500,000 data points. Your boss suggests to use a non-parametric method for classification. What does *'non-parametric'* mean? Do you agree with your boss? Explain your answer. **(2 p)**

1.9 The figure below plots a small set of data points $\boldsymbol{x} \in \mathbb{R}^2$ as well as two initialized cluster means denoted by $*$. Perform one iteration of the k-means algorithm using the Euclidean distance: $d(\boldsymbol{x}, \boldsymbol{\mu}) = \sqrt{\sum_{j=1}^{m}(x_j - \mu_j)^2}$. Fill in the table below with your results and mark the updated cluster means in the plot. Has the algorithm already converged? **(5 p)**



i	$\boldsymbol{x}^{(i)}$	$d(\boldsymbol{x}^{(i)}, \boldsymbol{\mu}^{(1)})$	$d(\boldsymbol{x}^{(i)}, \boldsymbol{\mu}^{(2)})$	**Cluster assignment** (1 or 2)
1	(1, 3)			
2	(2, 4)			
3	(5, 1.5)			
4	(6, 2)			
5	(7, 2)			

$\boldsymbol{\mu}^{(1)} = (3, 1)$ $\boldsymbol{\mu}^{(2)} = (9, 2.5)$ $\boldsymbol{\mu}^{(1)}_{new} = ($, $)$ $\boldsymbol{\mu}^{(2)}_{new} = ($, $)$

Has the algorithm converged after the first iteration?

□ Yes

□ No

1.10 Explain in up to three sentences how you can choose the number of principal components for dimensionality reduction. **(2 p)**

Maximum attainable points for task 1: **20 points**

2 Decision Trees and Ensemble Methods

2.1 You are given the dataset listed in the table below. The dataset consists of $m = 3$ attributes and two classes (positive $\oplus$ and negative $\ominus$). Derive a decision tree classifier from the given dataset using the information gain heuristic. Write down all computations necessary and draw the final decision tree. **(10 p)**

A_1	A_2	A_3	C
a	p	x	$\oplus$
a	m	x	$\oplus$
b	m	x	$\ominus$
b	p	x	$\ominus$
a	p	y	$\oplus$
a	p	z	$\ominus$
a	m	z	$\ominus$
b	m	z	$\ominus$
b	m	y	$\ominus$
a	m	y	$\oplus$

2.2 Generally speaking, which class distribution maximizes the entropy function? Consider two classes. **(1 p)**

2.3 Which of the following algorithms is **not** an example of ensemble learning? **(1 p)**

- □ Random forest
- □ AdaBoost
- □ Logistic regression
- □ ExtraTrees
- □ All algorithms are ensemble methods.

Maximum attainable points for task 2: **12 points**

3 Evaluation of Machine Learning Models

3.1 How can the area-under-the-curve (AUC) in ROC space be interpreted? What does an AUC of 0.5 mean in this context? **(2 p)**

3.2 Imagine you have trained a classification model to classify skin tissue samples in either cancerous or healthy. The model should avoid false negatives at all costs (*'false negative'* means predicting cancerous tissue as healthy). Which evaluation metric (precision or recall) would you prefer? Explain your answer. **(2 p)**

3.3 Compute the accuracy score of your model based on the classification results depicted in the confusion matrix below. **(1 p)**

3.4 Compute the metrics precision, recall and F_1-score for class C_1 based on the classification results depicted in the confusion matrix below. **(3 p)**

Conf. mat.		gold			
		C_1	C_2	C_3	Σ
predicted	C_1	25	4	9	38
	C_2	8	31	0	39
	C_3	2	3	18	23
	Σ	35	38	27	100

3.5 Which statements concerning bias and variance are correct? **(2 p)**

□ A decision stump (= decision tree with only one split) suffers from high bias.

□ If a model has high variance, adding more training examples does not help.

□ If a model has high bias, adding more training examples helps.

□ A biased model underfits the training data.

□ None of the above options apply.

3.6 Explain in up to three sentences what cross-validation is and when you should use it. **(2 p)**

Maximum attainable points for task 3: **12 points**

4 Neural Networks / Deep Learning

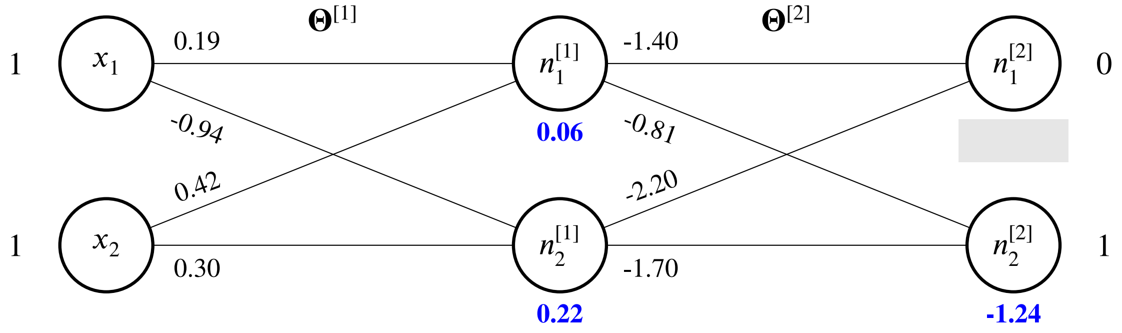
4.1 You want to train a neural network on the *MNIST* dataset to recognize hand-written digits. The images of 10 possible digits have a resolution of 28×28 pixels. The MLP used for the task has the following hidden layer dimensions: $size(L_1) = 64$ units, $size(L_2) = 32$ units. Each layer also has a constant bias input and the classes are one-hot encoded. How many parameters do you have to optimize? **(3 p)**

Number of parameters:

4.2 Your colleague suggests to use neural networks to solve a regression task. Which activation function would you have to use in the output layer of your network in order to achieve the desired results? **(1 p)**

4.3 Why is it necessary to use non-linear activation functions in deep learning? **(2 p)**

4.4 You are given the neural network depicted below. The input is given by $\boldsymbol{x} = [1, 1]$, the desired output is $\boldsymbol{y} = [0, 1]$. The network is trained using the squared error loss function: $\mathcal{J}(\boldsymbol{\Theta}) = (z_{n_k^{[2]}} - y_k)^2$. Some of the (pre-)activations (cf. table) and error gradients (bold numbers below the neurons) were already computed. Fill in the preactivation for $n_2^{[2]}$, the activation for $n_1^{[1]}$, the error gradient for $n_1^{[2]}$, as well as the weight gradient for $\Theta_{11}^{[2]}$. **(4 p)**



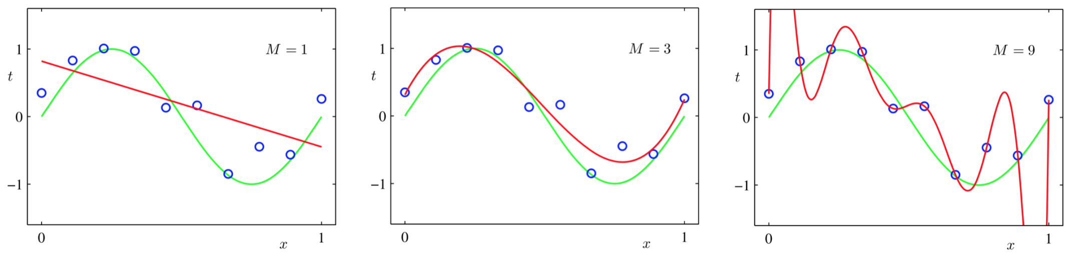
Neuron	Preactivation	Activation	Activation function
$n_1^{[1]}$	0.61		ReLU
$n_2^{[1]}$	-0.64	0.00	ReLU
$n_1^{[2]}$	-0.85	0.30	Sigmoid
$n_2^{[2]}$		0.38	Sigmoid

Weight gradient for $\Theta_{11}^{[2]}$:

Maximum attainable points for task 4: **10 points**

5 Regression

5.1 Your colleague shows you three regression plots (green: true function, red: regression curve, *M* denotes the degree of the polynomial) and asks you which model to use. Justify your decision and explain the issues with the other models. What could you do to mitigate these problems? **(3 p)**



5.2 The squared error cost function results from maximum likelihood estimation assuming uniform noise. **(1 p)**

□ True

□ False

5.3 Why do we need basis functions in regression? Name two common examples of basis functions. **(2 p)**

Maximum attainable points for task 5: **6 points**

Additional space:

Maximum attainable points for the exam: **60 points**