DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO!

Technische Universität Kaiserslautern
Fachbereich Informatik
Prof. Dr. Marius Kloft, TA: Rodrigo Alves

Kaiserslautern, 25th September 2019

Final Examination
Machine Learning 1: Foundations

Summer Semester 2019

Exam ID: **1**

Signature: ....

Test Duration:	150 Minutes
Achievable Points:	100

To be completed by corrector:

Task	Possible Points	Score
1	10	
2	10	
3	10	
4	15	
5	20	
6	20	
7	15	
Total Score	100	
Exam Grade	–	

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO!

Exam Information

- Before you begin the test you will have 15 minutes to read through and orient yourself with the test. Once test time begins you will have 150 minutes to complete the test.
- During the orientation period you may raise your hand to indicate that you have a question about the test and a test proctor will come to record your question. At the end of the orientation period the test proctors will answer all collected questions in front of the class.
- You are not allowed to leave the room during the last 30 minutes of the test so as to not disturb other test-takers.
- Only one solution per question will be accepted. Giving multiple answers to a single question may result in receiving no credit for that question.
- Please cross out anything that you do not want the grader to consider during grading.
- Use the space provided after each question to provide your answer.
- For each answer, please clearly indicate which question you are answering.
- You may only use writing implements which are **not erasable**. For example no pencils may be used. Do not use red pens!
- If you have a question during the test please raise your hand so a test proctor may assist you.
- You may not use any electronic resources (e.g. eletronic calculators, mobile phones and laptops) during the test. Using electronic resources may result in receiving no credit for the exam.
- In questions that have a maximum word limit, you must indicate the number of words of your answer in the corresponding field. Items that exceed the maximum limit will not be corrected. Be concise!
- By the best of our knowledge, we have ordered Questions 2 to 7 by the ascending level of difficulty

Pseudocode example

This page contains no questions to be solved. The following is an example of how to answer items involving pseudocode. Based on this format, answer questions 2 (c), 4(b) and 6(d).

Suppose we are running k-means on a collection of data $x_1, \ldots, x_n$. Given centers $c_1, \ldots, c_m$, write pseudocode describing *one* update of centers using the k-means algorithm.

```
Algorithm kMeansOneUpdate($c_1, \ldots, c_m$,$x_1, \ldots, x_n$)
  for $i \in 1, \ldots, n$ do
    clusterAssignment[$i$]= $\arg\min_{j=1,\cdots,m} \|x_i - c_j\|$
  end for
  for $j \in 1, \ldots, m$ do
    $c_j$ = average($\{x_i \mid i \in \{1, \cdots, n\}$, clusterAssignment[$i$] == $j\}$)
  end for
  return $c_1, \ldots, c_m$
```

Question 1 (10 points) TRUE OR FALSE? Justify **ALL** answers (Max. 100 words per item). The negation of one of the items will not be considered a justification. (Right answer = +1 point, Wrong answer = −1 point, No answer = 0 points, Right answer but wrong justification = 0 points — Total minimum = 0 points)

(a) Principal component analysis (PCA) is a supervised learning algorithm.

(b) Autoencoders are a non-linear generalization of PCA.

(c) Cross-validation can be used to tune a hyperparameter.

(d) The "kernel trick" is frequently used to efficiently transform a non-linear learning machine into a linear learning machine.

(e) No matter which training set, all kernelized classifiers will obtain strictly better training accuracy than any non-kernelized classifiers.

(f) In the last decade, artificial neural networks became very popular because they obtained good results in some tasks that are considered hard for shallow learning methods.

(g) Optimizing a function through gradient descent can yield a more accurate solution than stochastic gradient descent. However, stochastic gradient descent often converges faster (in terms of execution time) than gradient descent to a sufficiently good solution.

(h) Transfer learning permits us to use an ANN trained for multi-classification to cluster unknown classes.

(i) Applying the Box-Cox transformation to the labels before training a regression algorithm can improve the regression result.

(j) The k-means clustering algorithm cannot be kernelized.

Question 2 (03 + 02 + 03 + 02 = 10 points)

(a) [03 points] **(Max. 100 words)** Suppose you are training a neural network to solve a classification problem. After 500 iterations of the stochastic gradient descent (SGD) algorithm, you could observe the result shown in Figure 1:

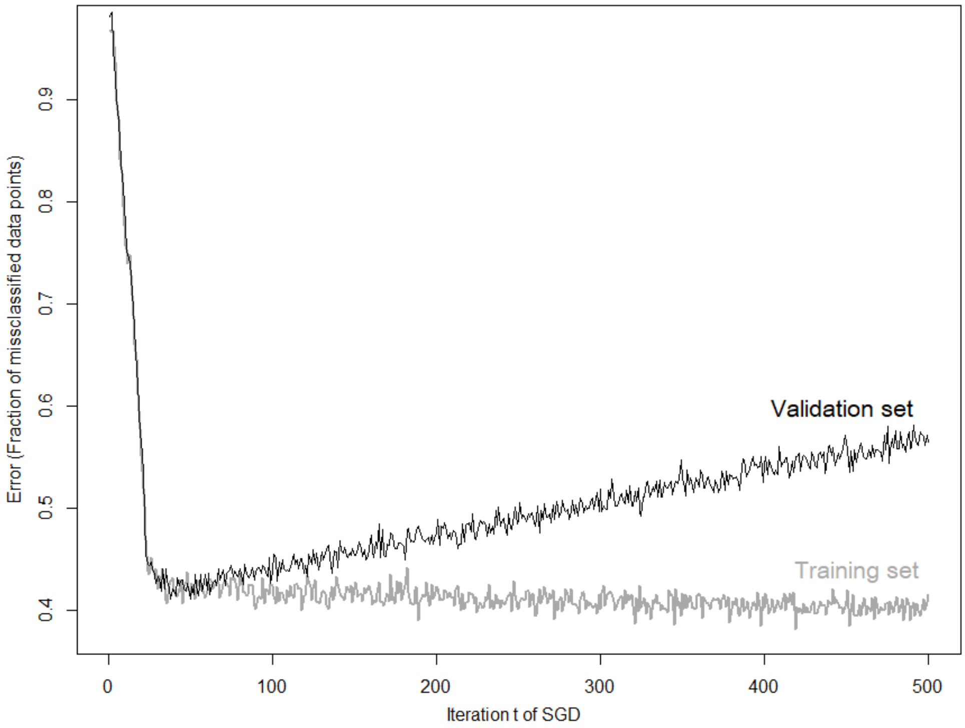


Figure 1: Training and validation error as a function of the iteration t of SGD

Considering Figure 1, use early stopping to determine the lowest value of t that ensures a well-fitted solution. Justify your answer by relying on the concepts of underfitting and overfitting.

Number of words: ________________

(b) [02 points] **(Max. 150 words)** Note that for $t > 300$ the validation error still increasing. Suppose we are using a very deep neural network (consisting of hundreds of hidden layers). For $t > 300$, justify the behaviour of the curves (Figure 1). Base your argumentation on the architecture of the neural network.

Number of words: ________________

(c) [03 points] **(Max. 100 words)** Consider the following statement: *"In deep learning, increasing the regularization strength will oftentimes increase the training loss."* Is the statement TRUE or FALSE? Discuss.

Number of words: ________________

(d) [02 points] **(Max. 100 words)** Give two examples of regularization techniques for neural networks and explain them.

Number of words: ________________

Question 3 (01 + 03 + 04 + 02 = 10 points)

(a) [01 point] The following is a scatter plot of some 2-dimensional input data of a clustering problem.

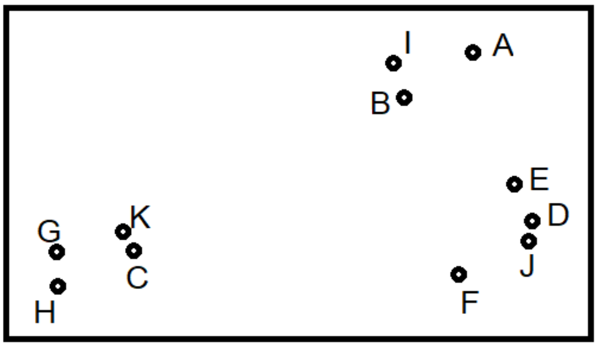


Figure 2: Example of 2-dimensional input data of a clustering problem

Draw into Figure 2 reasonable centroids that could result from running the k-means algorithm until convergence, using $k = 3$.

(b) [03 points] A disadvantage of the k-means algorithm is that we need to define the number of clusters before its execution. One way to solve this problem is by using hierarchical clustering. Based on the data shown in Figure 2, draw the resulting tree of clusters by performing **average linkage** hierarchical clustering.

(c) [04 points] Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be data in a clustering problem. Write pseudocode that implements hierarchical clustering considering **simple linkage**.

(d) [02 points] **(Max. 150 words)** In clustering tasks involving very high-dimensional inputs, would it be advantageous to first apply PCA before running the k-means algorithm (on the dimensionality-reduced data)? Why?

Number of words: _______________

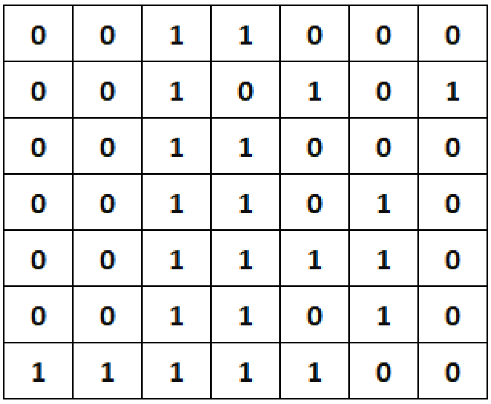
Question 4 (04 + 07 + 04 = 15 points)

(a) [04 points] **(Max. 150 words)** Describe the backpropagation algorithm in five sentences or less.

Number of words: ______________

(b) [07 points] Convolutional neural networks apply multiple cascaded convolutions to solve tasks. Consider a black-and-white image matrix $M \in \{0,1\}^{p \times p}$ (i.e., feature entries in $\{0,1\}$), a stride of s ($0 < s \in \mathbb{N}$) and a convolutional filter matrix F of size $m \times m$, where $0 < m \in \mathbb{N}$. Write down pseudocode to apply the filter F to the image matrix M with a stride s. Use zero padding. **Hint:** *Do not forget the algorithm inputs and outputs.*

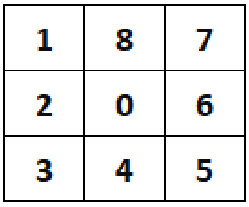
(c) [04 points] Consider the following image matrix (7×7) in black and white.



0	0	1	1	0	0	0
0	0	1	0	1	0	1
0	0	1	1	0	0	0
0	0	1	1	0	1	0
0	0	1	1	1	1	0
0	0	1	1	0	1	0
1	1	1	1	1	0	0

Figure 3: Example of an image matrix of size 7 x 7 in black (0) and white (1).

Now consider the following convolutional filter matrix (3×3)



1	8	7
2	0	6
3	4	5

Figure 4: Example of a convolutional filter matrix (3×3).

Let the stride be 2 and use zero padding. Apply the convolution filter (Figure 4) to the image matrix (Figure 3). Your answer must be a matrix.

Question 5 (05 + 05 + 05 + 05 = 20 points) Consider the following binary classification problem:

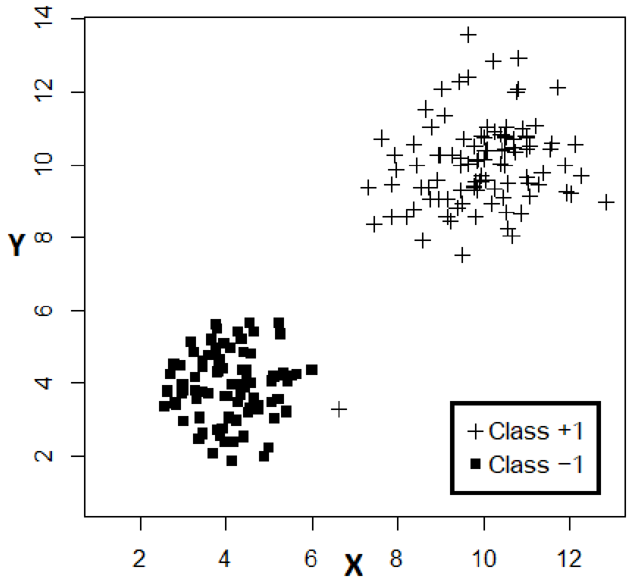


Figure 5: An example of a binary classification problem

The unconstrained linear soft-margin SVMs can be formulated as:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\max\Big(0, 1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b)\Big)}_{J(\mathbf{w})}, \tag{1}$$

where $C > 0$ is a regularization parameter.

(a) [05 points] **(Max. 120 words)** Assume that we are training a linear soft-margin SVM (1) to solve the binary classification problem presented in Figure 5. Draw into Figure 5 the hyperplane H_0 in the case of $C \to 0$. Justify your answer. Do not forget to label H_0.

Number of words: ________________

(b) [05 points] **(Max. 120 words)** Considering the same scenario (that is, training (1) on the data shown in Figure 5), draw into Figure 5 a second hyperplane H_∞ for the case of $C \to \infty$. Justify your answer. Do not forget to label H_∞.

Number of words: ________________

Consider now the classification problem shown in Figure 6. The labels (classes) of the input points $P1$ and $P2$ are unknown.

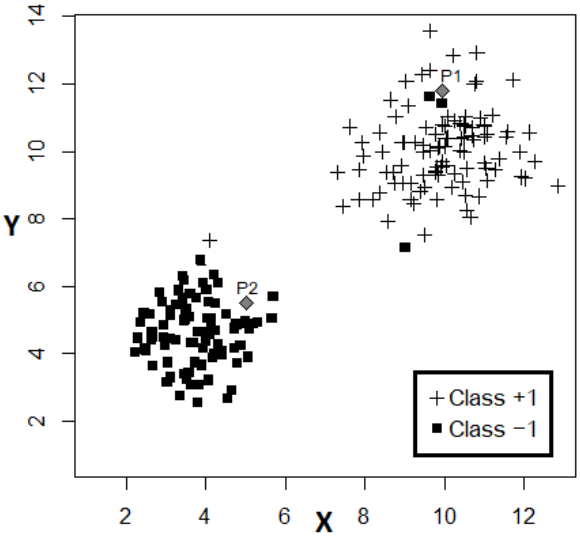


Figure 6: Another example of a binary classification problem

The RBF (radial basis function) kernel, also known as Gaussian kernel, is a function commonly used to kernelize support vector machines (SVMs). To train a kernelized SVM using an RBF kernel, we need to tune a hyperparameter σ (sigma). Recall from the lectures that the following is the kernel function of the RBF kernel:

$$k(\mathbf{x}, \tilde{\mathbf{x}}) := e^{-\frac{1}{2\sigma}||\mathbf{x}-\tilde{\mathbf{x}}||^2} \tag{2}$$

(c) [05 points] **(Max. 150 words)** Assume that we are training an SVM using RBF kernel to solve the binary classification problem presented in Figure 6. Describe the decision boundary when $\sigma \to 0$.

Number of words: _______________

(d) [05 points] **(Max. 150 words)** Discuss what would be the expected classification (-1 or $+1$) for the points $P1$ and $P2$ (Figure 6) in the cases $\sigma \to 0$ and $\sigma \to +\infty$, respectively. Justify your answer.

Number of words: _______________

Question 6 (03+05+05+06 = 20 points) Consider the following optimization problem:

$$\mathbf{w}^* := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{|\mathbf{w}|_1 + C\left\|\mathbf{y} - X^\top\mathbf{w}\right\|^2}_{J(\mathbf{w})}. \tag{3}$$

Here $C > 0$ is a hyperparameter and $|\mathbf{w}|_1 = \sum_{i=1}^{d} |w_i|$. In machine learning, this method is called *LASSO* (least absolute shrinkage and selection operator) and often used for regression.

(a) [03 points] **(Max. 100 words)** Is it possible to compute the solution $\mathbf{w}^*$ of (3) in closed form? Argue why or why not (you do not need to prove it formally).

Number of words: ______________

(b) [05 points] The optimization function of lasso regression may be optimized by iteratively applying a variant of ridge regression:

$$\lim_{k\to\infty} \mathbf{w}^{(k)} = \mathbf{w}^*, \tag{4}$$

where $\mathbf{w}^{(k)} := \arg\min_{\mathbf{w}} J_k(\mathbf{w})$ and

$$J_{k+1}(\mathbf{w}) = \sum_{j=1}^{d} \left(\frac{1}{\left|w_j^{(k)}\right|} \left(w_j\right)^2 \right) + C\left\|\mathbf{y} - X^\top\mathbf{w}\right\|^2. \tag{5}$$

Compute $\nabla_{\mathbf{w}} J_{k+1}$. Note that $\mathbf{w}^{(k)}$ can be treated as constant in (5).

(c) [05 points] Set $\nabla_{\mathbf{w}} J_{k+1} = 0$ to find a closed form for $\mathbf{w}^{(k+1)}$.

(d) [06 points] Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be data in a regression problem and define $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n, X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d\times n}$. Write pseudocode that computes $\mathbf{w}^*$ iteratively using the closed formula calculated in item (c). The algorithm must stop after T iterations. Initialize each element of $\mathbf{w}^{(0)}$ with a different random value between 0 and 1. **Note:** *If you cannot solve items (b) and/or (c) instead use* $\mathbf{w}^{(k)} := \arg\min_{\mathbf{w}} J_k(\mathbf{w})$ *in your code (In this case, the maximum for this item will be 4 points).* **Hint:** *Do not forget the algorithm inputs and outputs.*

Question 7 ($04 + 04 + 07 = 15$ points)

(a) [04 points] **(Max. 90 words)** Give, with a brief justification, two scenarios where it is not beneficial to use a kernel method.

Number of words: ______________

(b) [04 points] If we are given n datapoints $x_1, x_2, \dots, x_n$ and a kernel function $k(\mathbf{x}, \tilde{\mathbf{x}})$, define the corresponding kernel matrix $K \in \mathbb{R}^{n\times n}$. Prove that K is positive semidefinite.

(c) [07 points] Let X be a matrix where the entries are natural numbers between 1 and m ($X \in \{1, 2, 3, \dots, m\}^{d\times n}$), $m \gg 0$, d is the number of features and n is the number of data points ($n \gg 0$) . Let also $K \in \mathbb{R}^{n\times n}$ be the kernel matrix defined by $K_{i,j} = k(X_{\bullet,i}, X_{\bullet,j})$, where $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the kernel function and $\forall l$, $X_{\bullet,l}$ is the l^{th} column of X.

What is the lowest value of n that guarantees the kernel matrix K is singular (not invertible) for any matrix X and for any kernel function k? You must justify (prove) your answer.