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

Kaiserslautern, 5th August 2019

Final Examination
Machine Learning 1: Foundations

Summer Semester 2019

Test Duration: 150 Minutes
Possible Points: 100

To be completed by corrector:

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

Question 1 (10 points) TRUE OR FALSE? If the statement is false, justify your answer (Max. 100 words per item). (Right answer = +1 point, Wrong answer = −1 point, Minimum = 0 points)

(a) The K-means algorithm solves the clustering dilemna (that is, finding the right number of clusters).

(b) K-means is a supervised learning algorithm.

(c) Despite the recent progress in deep learning, SVMs are the state of the art in several applications.

(d) Cross-validation can be used to compare different algorithms on the same task.

(e) Backpropagation is an algorithm used to regularize an artificial neural network.

(f) The concept of learning rate gives stochastic gradient descent an advantage over gradient descent.

(g) It is not possible to use an artificial neural network to solve unsupervised tasks.

(h) No matter which learning problem, reducing the dimensionality of the data matrix results in a loss of relevant information and thus increases the root mean squared error (RMSE).

(i) Early stopping can be used to regularize an artificial neural network.

(j) Most linear classifiers can be kernelized. However, up to now, not a single kernelized regression algorithm is known.

Question 2 $(03 + 03 + 04 = 10$ points)

(a) [03 points] The following is a scatter plot of some 2-dimensional input data of a binary classification problem.

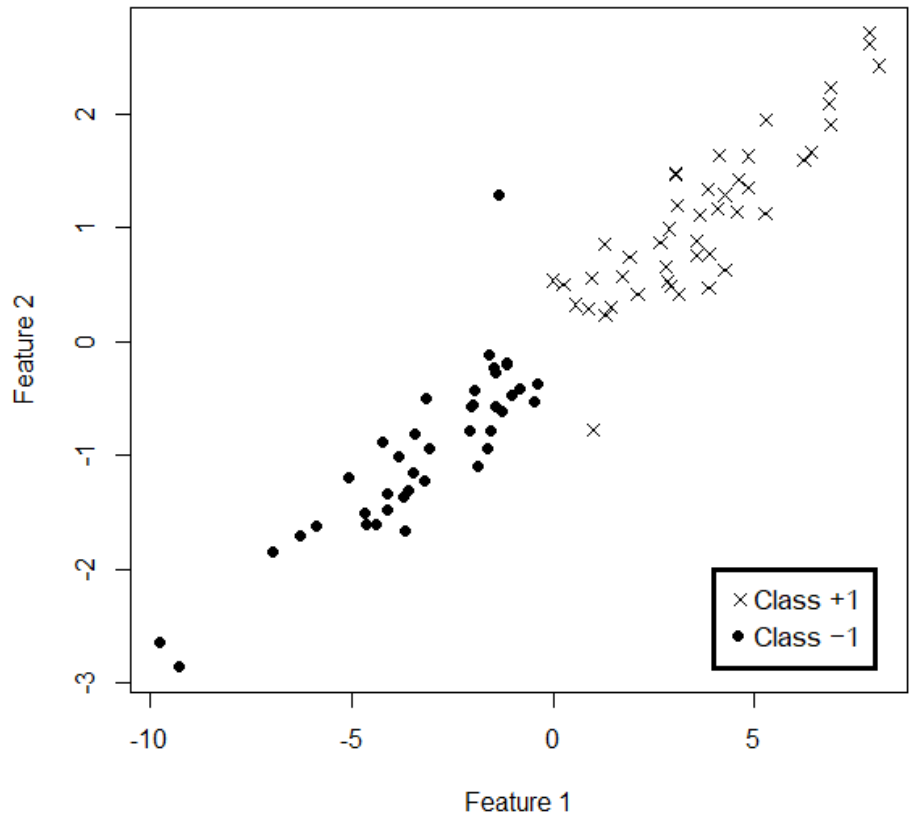

Figure 1: Example of 2-dimensional input data of a binary classification problem

Draw the first principal component into Figure 1.

(b) [03 points] **(Max. 150 words)** In item (a) we can use SVM as a classifier and obtain a good accuracy because the data is linearly separable. After the reduction to one dimension using Principal Component Analysis (PCA) is the data still linearly separable? Why?

Number of words: _____

(c) [04 points] **(Max. 200 words)** In learning tasks involving very high-dimensional inputs, why could it be advantageous to first apply PCA before train a SVM (on the dimensionality-reduced data)?

Number of words: _____

Question 3 $(03 + 02 + 03 + 02 = 10$ points)

(a) [03 points] Suppose you are training a model to solve a regression problem, and you have to tune a hyperparameter K. After several tested values of K you could observe the graph presented in Figure 2.

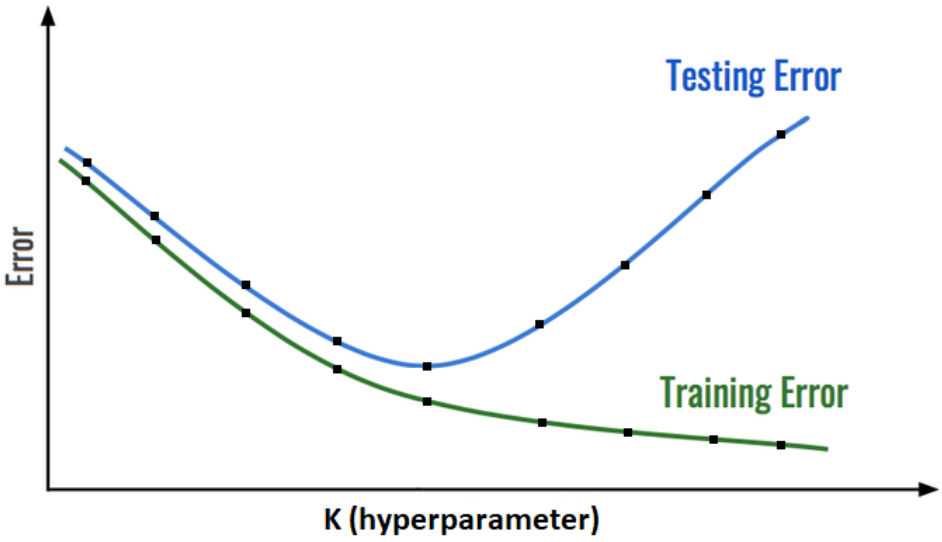

Figure 2: Training and test error in function of hyperparameter K

Divide this graph (Figure 2) into three regions that are best labeled as:

- Underfitting
- Overfitting
- Just right

(b) [02 points] **(Max. 100 words per item)** Justify your choices in item (a).

Number of words: _____

(c) [03 points] **(Max. 200 words)** Why is the overfitting problem typically more critical than the underfitting problem?

Number of words: _____

(d) [02 points] **(Max. 200 words)** Give an example on how to solve the underfitting problem, and give an example on how to solve the overfitting problem.

Number of words: _____

Question 4 $(05 + 05 + 03 + 02 + 05 = 20$ points) The unconstrained linear soft-margin SVMs can be formulated as:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} L(\mathbf{w})}_{J(\mathbf{w})}, \tag{1}$$

where $C > 0$ is a regularization parameter and $L(\mathbf{w})$ is the loss function. A common loss function for SVMs is the hinge loss:

$$L_1(\mathbf{w}) = \max\Big(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i)\Big),$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of d features and $y_i \in \{-1, 1\}$ is the class of the datapoint i. However, we can use a different loss function, for example, the squared hinge loss:

$$L_2(\mathbf{w}) = \max\Big(0, (1 - y_i(\mathbf{w}^\top \mathbf{x}_i))^2\Big).$$

Using the squared hinge loss $L_2(\mathbf{w})$ in place of $L(\mathbf{w})$, Equation (1) is called *square-hinge-loss SVM*.

(a) [05 points] Compute the gradient $\nabla_{\mathbf{w}} J$ of the square-hinge-loss SVM.

(b) [05 points] Show that the SVM using the squared hinge loss is also a convex problem.

(c) [03 points] **(Max. 100 words)** Why is it important for machine learning problems to know if a function is convex?

Number of words: _____

(d) [02 points] Consider the hard-margin SVM as introduced in our lectures (using the standard hinge loss, not the squared hinge loss), and consider the following binary classification problem:

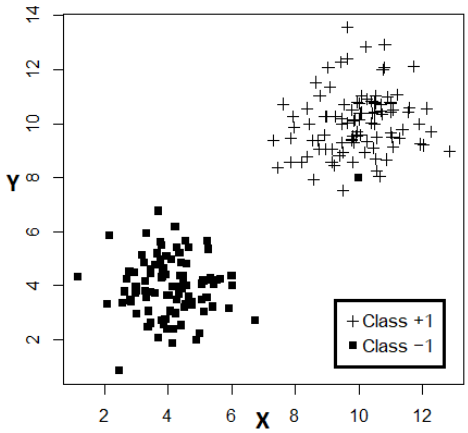

Figure 3: Example of a binary classification problem

Can we solve this problem with hard-margin SVM? Justify your answer.

(e) [05 points] Put C=1 in Equation (1). Let H_1 be the hyperplane that is the solution of (1) when using the hinge loss. Analogously, let H_2 be the hyperplane that is the solution of (1) when using the *squared* hinge loss. Draw two hyperplanes H_1 and H_2 into Figure 3. Do not forget to label the two lines with H_1 and H_2, respectively. Explain your decision.

Question 5 (04+04+04+08 = 20 points) Consider the following optimization problem for regression:

$$\mathbf{w}^* := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \quad \underbrace{\alpha||\mathbf{w}||^2 + (1-\alpha)||\mathbf{w}||^4 + C\left\|\mathbf{y} - X^\top \mathbf{w}\right\|^2}_{J(\mathbf{w})}, \tag{2}$$

where $\alpha \in [0,1]$ and $C > 0$ are hyperparameters.

(a) [04 points] **(Max. 175 words)** Discuss how the value of α affects the regularization procedure.

Number of words: _____

(b) [04 points] Compute $\nabla_{\mathbf{w}} J(\mathbf{w})$.

(c) [04 points] Derive a closed-form solution for the minimizer when $\alpha = 1$.

(d) [08 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 $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n, X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d\times n}$. Depending on the dimensions of the matrix X, using the closed-form solution may be too slow. One possible solution is to use stochastic gradient descent to calculate $\mathbf{w}^*$ in an efficient way. Write a pseudocode that implements SGD for the regression problem (2). **Hint:** *Do not forget to define the inputs and outputs of the algorithm as well as ensure its convergence.*

Question 6 $(06 + 03 + 06 = 15$ points)

(a) [06 points] Convolutional Neural Networks (CNN) are broadly used in computer vision tasks. The concepts of *patch*, *stride* and *pooling* are very important in this context. Write down pseudocode that implements max pooling. Consider as input to your algorithm the following items: an 8-bit gray scale image matrix $M \in \{0, 1, 2 \ldots, 255\}^{p \times p}$ (i.e., feature entries in $\{0, 1, 2, ..., 255\}$), a stride of s $(0 < s \in \mathbb{N})$ and a patch size of $m \times m$, where $0 < m \in \mathbb{N}$. An example of the matrix M (where $p = 7$) can be seen in Figure 4. Your output is a 8-bit gray scale image matrix $M_{pool} \in \{0, 1, 2 \ldots, 255\}^{d \times d}$.

(b) [03 points] Consider the following 8-bit gray-scale image matrix (7×7) in gray scale.

181	235	47	181	182	192	37
175	221	108	228	163	8	85
201	104	236	50	44	67	196
17	205	137	16	148	17	77
242	102	155	143	240	22	194
182	21	170	156	76	50	183
232	20	179	2	8	48	176

Figure 4: Example of an 8-bit gray-scale image of size 7 x 7.

Let the patch size be 3×3 and the stride be 2. Execute max pooling on the matrix presented in Figure 4. Your answer must be a matrix.

(c) [06 points] **(Max. 350 words)** Dimensionality reduction is an effective technique for regularizing shallow learning methods. Suppose you are solving a binary classification problem through a very deep neural network (consisting of hundreds of hidden layers). In this scenario, discuss the effectiveness of dimensionality reduction as a method to avoid overfitting.

Number of words: _____

Question 7 $(03 + 04 + 08 = 15$ points)

(a) [03 points] **(Max. 100 words)** Describe the "kernel trick" and explain why it is useful in machine learning.

Number of words: _____

(b) [04 points] Let k_1 and k_2 be kernels, $0 < n \in \mathbb{N}$ and $\alpha, \beta > 0$. Define k_3 as $k_3(y, x) = \big(\alpha k_1(y, x) + \beta k_2(y, x)\big)^n$. Show that k_3 is a kernel.

Hint: *You may use the Theorem 1*

Theorem 1 *The Hadamard product (elementwise product) of two positive semidefinite matrices is also a semipositive definite matrix.*

(c) [08 points] Let X be a matrix where the entries are natural numbers between 1 and m ($X \in \{1, 2, 3, \ldots, m\}^{d \times n}$), $m \gg 0$, 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 $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 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?