# Machine Learning I

## Exam Date: 05.08.2019

Duration: 150min.

Max. Points: 100pts.

## Question 1: Multiple Choice **[10 Points]**

If you disagree with a statement, write up to 100 words on why it is wrong. (Answer right 1 point, answer wrong -1 point)

1. Until now, no kernel for regression was found.
2. K-Means solves the clustering dilemma, i.e., how many centroids to choose.
3. K-Means is a supervised learning algorithm.
4. Although neural networks are the current cutting-edge technology, SVMs are still frequently used.
5. Backpropagation is used as a means of regularization.
6. Cross-Validation can be used to compare the performance of different algorithms on the same task.
7. The concept of the learningrate gives stochastic gradient descent an advantage over gradient descent.
8. There is no deep learning approach for unsupervised learning.
9. PCA results in loss of relevant data and increases RMSE.

   (Note, that this was a false statement, since a loss of relevant data does not have to increase the RMSE.)
10. Early stop is used to apply regularization on an ANN.

## Question 2: PCA **[10 Points=3+3+4]**

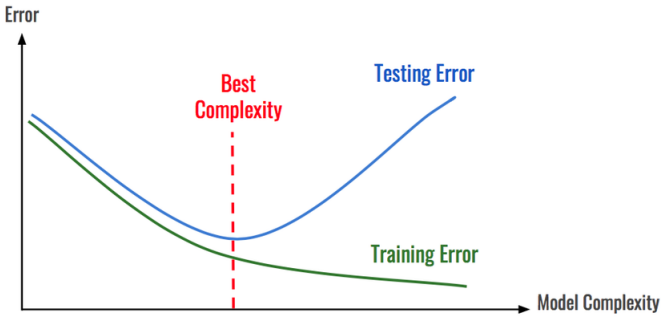
(a) Draw linear PCA into Plot1[1]

(b) After applying PCA, is the data still separable by a linear classifier?

(c) Discuss applying PCA before fitting a SVM.

[1] Class 1 as a data-cloud at the lower left corner, class 2 as a data-cloud at the uppper right corner. There are two outliers, laying in the cloud of the other class respectively such that the data is still linear seperable. But after applying the PCA they are projected onto a line s.th. they are within the points of the other class → not seperable by a linear classifier after applying PCA.

## Question 3: Over- and Underfitting [10 Points=2+3+3+2]




Plot2: Standard curves for illustrating over- and underfitting. (Caution, spoilers.)

(a) Draw the overfitting-, underfitting- and the just-right-region into Plot2[2].

(b) Justify your choices.

(c) Why is underfitting a smaller problem than overfitting?

(d) Name a method to avoid overfitting and a method against underfitting.

## Question 4: Classification [20 Points= 2+5+5+3+5]

Given the hinge-loss and the squared hinge-loss, which is

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(w^T x_i))^2, C > 0$$

.

(a) Is Plot3[3] separable by a linear classifier/SVM?

(b) Calculate the gradient of the squared hinge loss.

(c) Show that the squared hinge-loss is convex.

(d) Why is it good to know that a loss is convex?

(e) Draw the hyperplanes for hinge-loss and squared hinge-loss into Plot3 and explain your choices.

[2] Stolen from here.

[3] Again, class 1 as a data-cloud at the lower left corner, class 2 at the upper right corner. One outlier of class 1 lays within the data of class 2.

## Question 5: Regression

**[20 Points= 5+5+5/4+5/6]**

Given following model with regularization parameter $\alpha$:

$$\alpha||w||^2 + (1-\alpha)||w||^4 + C||y - X^T w||^2$$

(a) Describe the influence of $\alpha$ on the regularization behavior.

(b) Calculate the gradient with respect to $w$.

(c) Calculate the optimal solution for $\alpha = 1$.

(d) Describe pseudocode of SGD for given model.

(Note, that a detailed pseudocode was necessary to get full points. For example, one had to formulate the gradient dependent on the batch-size and adjust the cost parameter $C$ according to the batch-size.)

## Question 6: CNN

**[15 Points=6+3+6]**

(a) Write pseudocode for a max-pooling on 8-bit greyscale Image with Inputs $M \in \{0, ..., 255\}^{pxp}$, stride $s$ , patchsize $m$.

(b) A $7 \times 7$ matrix with values was given. Compute the result of max-pooling on given matrix with patchsize 3 and stride 2.

(c) Discuss the effectiveness of dimensionality reduction to avoid overfitting when using a very deep neural network for binary classification (max. 350 words).

## Question 7: Kernels

**[15 Points=3+4+8]**

(a) Explain the kernel trick in 100 words or less.

(b) Show that if $k_1$ and $k_2$ are kernels, $n > 0$ and $n \in \mathbb{N}, \alpha, \beta > 0$ with $\alpha, \beta \in \mathbb{R}$:

$$k_3 = (\alpha \cdot k_1 + \beta \cdot k_2)^n$$

is a kernel.

**Hint:**You may use the following Theorem: The Hadamard-(elementwise-)Product of two positive semi-definite matrices is also positive semi-definite.

(Note, that you need to use the theorem inductively to get the full points when showing that $K^n$ is a kernel if $K$ is a kernel.)

(c) Let $X \in \{0, 1, \ldots, m\}^{dxn}$, $m >> d, n >> 0$ and $K_{i,j} = k(X_{.i}, X_{.j})$ where $X_{.l}$ is the $l$-th column of $X$ and $k(x, y)$ a kernel. What is the minimal $n$ such that $K$ is singular?

(Note, that, in order to get full points, one does not only need to proof that $n$ does not work, but also why every smaller $n$ works.)