Cognitive Algorithms Exam

16.07.2019

Please fill in below your full name, your matriculation number, field of studies and which optional course you attended.

I hereby confirm that I feel capable to participate in this exam.

_____

Name	Musterlösung
Field of study	
Matriculation number	
Registration ☐ via Qispos ☐ yellow form ☐ I need a certificate ☐ as an exchange student ☐ as Neben-, Gasthörer	**Optional course** ☐ none ☐ not yet ☐ Python Programming ☐ Math course ☐ Seminar

Exercise	**Total Points**	**Points**
1 Overview	12	
2 Modelling	6	
3 Linear Classification	10	
4 Multilayer Perceptron	9	
5 Cross-Validation	7	
6 Kernels	7	
7 Unsupervised Learning	9	
Total	60	

1 Short overview questions

Hint: When we talk about linear methods (including linear regression), we mean methods that are linear in ω, i.e. $f(x) = \omega^\top x$ where ω can also contain a nonlinear transformation of the data and the offset/bias β.

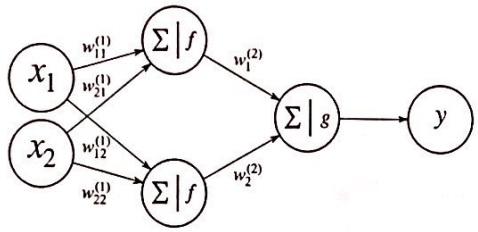


Figure 1: Neural Network

1. **[1 point]** What do we mean when we say that our machine learning model generalizes well? (1-2 sentences)

   It performs well on new / unseen data.

2. **[2 points]**
   The neural network as displayed in Fig. 1 is _____ training algorithm with ___ hidden layer(s).

   - ☒ a supervised
   - ☐ an unsupervised

   - ☒ 1
   - ☐ 2

3. **[1 point]** The neural network as displayed in Fig. 1 is a linear method.
   - ☐ True
   - ☐ False
   - ☒ depends on f
   - ☐ depends on g
   - ☐ depends on f and g

4. **[1 point]** Ridge regression always outperforms linear regression.
   - ☐ True
   - ☒ False

5. **[1.5 points]** Assume the covariance between observations X and their labels Y is less than 0. Which statements are always true?

   - [x] The regression function given by ordinary least squares will have negative slope.
   - [ ] The correlation between X and Y is -1.
   - [ ] Either X or Y has negative variance.

6. **[1.5 points]** Which statements are always true for kernel methods?

   - [x] You need to store all the training data to predict labels of new data points.
   - [x] For every valid kernel function k, there is a feature map ϕ, such that $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.
   - [ ] Kernel ridge regression gives better results and is faster to compute than ridge regression.

7. **[3 points]** Name 3 algorithms we discussed in the lecture that are able to solve a classification problem:

   NCC, Perceptron, LDA, MLP, ...

8. **[1 point]** When does it make sense to use a kernelized algorithm? (1-2 sentences)

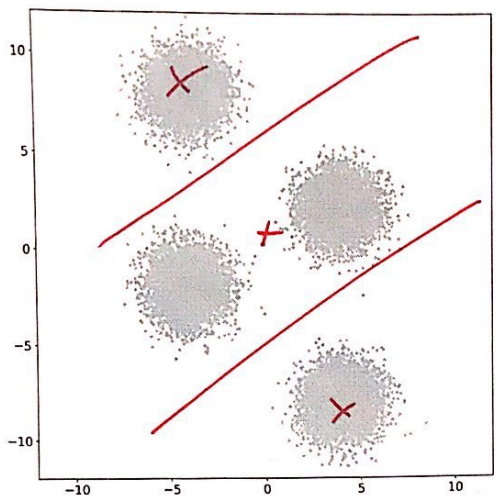
   - $D \gg N$
   - symbolic data
   - linearly separable in feature space, but not in input space

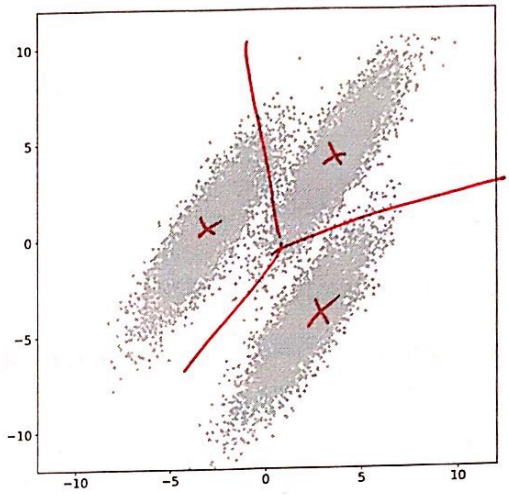
9. **[1 point]** "When the amount of data increases, overfitting is more likely." True or false? Explain in one or two sentences.

   False : Overfitting is less likely because the model is more stable

2 Modelling

1. [4 points] In the Figure below you find two exemplary unlabeled data sets. Draw for each a plausible result of the k-means algorithm after 50 iterations, i.e. mark plausible clusters and cluster centers found by k-means for $k = 3$. Initial cluster centers were randomly drawn from the set of data points.

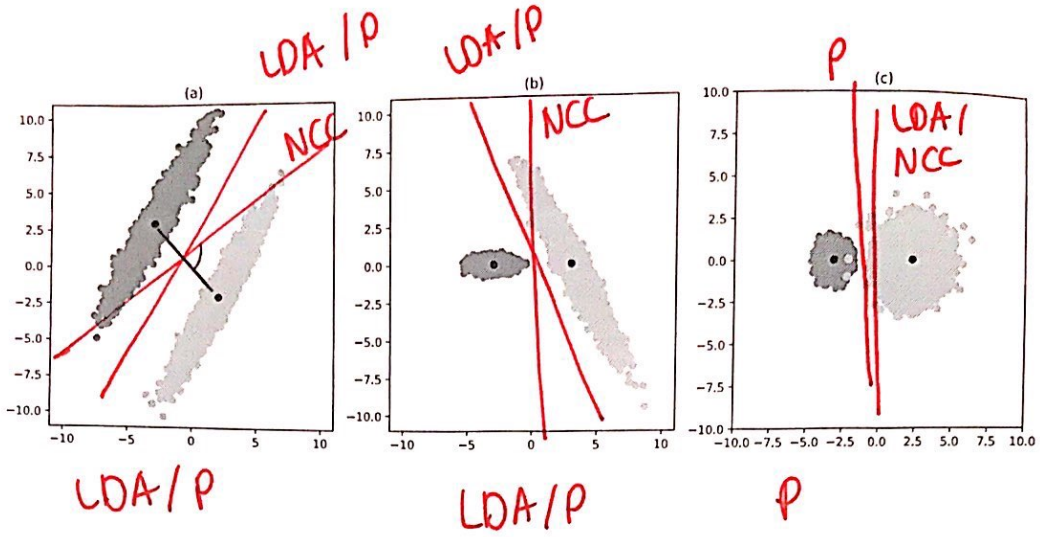




2. **[2 points]** For your bachelor thesis, you apply a classification algorithm to very high dimensional data you have recorded. Your supervisor is concerned that your features are still very correlated and suggests to apply an unsupervised algorithm to reduce the dimensionality of your data before applying the classification algorithm. What does she mean by that? Explain briefly in 1-2 sentences and state one possible algorithm which we discussed in the lecture.

correlated features → redundancy
→ remove dimensions without loosing information

PCA



3 Linear Classification

In the Figure above (a)-(c) you can see three different examples of a binary classification task. Each class consists of 5000 samples.

1. **[3 points]** Draw for each of the three examples the decision boundaries given by **LDA** and **NCC**, and one possible decision boundary given by **Perceptron** (thus, in total 9 decision boundaries). Make sure to label the decision boundaries with the respective algorithm.

2. **[1.5 points]** State for each example (a)-(c) which algorithm provides the best solution. Write your answer below the plots.

3. **[3.5 points]** You are given seven different covariance matrices. State for each class (in the figure above), which covariance matrix corresponds to the given dataset. Two covariance matrices do not belong to any data and one covariance matrix belongs to two datasets.

- $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ c, right
- $\begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$ c, left
- $\begin{bmatrix} 0.05 & 0 \\ 0 & 0.5 \end{bmatrix}$ None
- $\begin{bmatrix} 0.5 & 0 \\ 0 & 0.05 \end{bmatrix}$ b, left
- $\begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}$ a
- $\begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix}$ None
- $\begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix}$ b, right

- a, left/dark
- a, right/light
- b, left/dark
- b, right/light
- c, left/dark
- c, right/light
- None

4. **[2 points]** Based on your results on Task 3, calculate the within class scatter S_W for classification problem (b).

If you did not manage to assign a covariance matrix, you can assume the following matrices:

cov. of b, left/dark: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ (1)

cov. of class b, right/light: $\begin{bmatrix} e & f \\ g & h \end{bmatrix}$ (2)

Hint:

$$S_W = \frac{1}{N_+} \sum_{i=1}^{N_+} (x_{+_i} - w_+)(x_{+_i} - w_+)^T + \frac{1}{N_-} \sum_{i=1}^{N_-} (x_{-_i} - w_-)(x_{-_i} - w_-)^T \quad (3)$$

$$w_+ = \frac{1}{N_+} \sum_{i=1}^{N_+} x_{+_i} \qquad w_- = \frac{1}{N_-} \sum_{i=1}^{N_-} x_{-_i} \quad (4)$$

$$\begin{bmatrix} a+e & b+f \\ c+g & h+d \end{bmatrix}$$

$$\begin{bmatrix} 2.5 & -3 \\ -3 & 5.05 \end{bmatrix}$$

4 Multilayer Perceptron

1. **[2 points]** Name 2 popular activation functions that are used in neural networks.

   sigmoid, tanh, ReLU, softmax, ...

2. **[2 points]** What role does the learning rate play when training a neural network? What can happen when you choose your initial learning rate to be too small?

   - step size at weight update
   - influences speed (how fast the weights change)
   - too small: training is very slow, possible to land in a bad local minimum

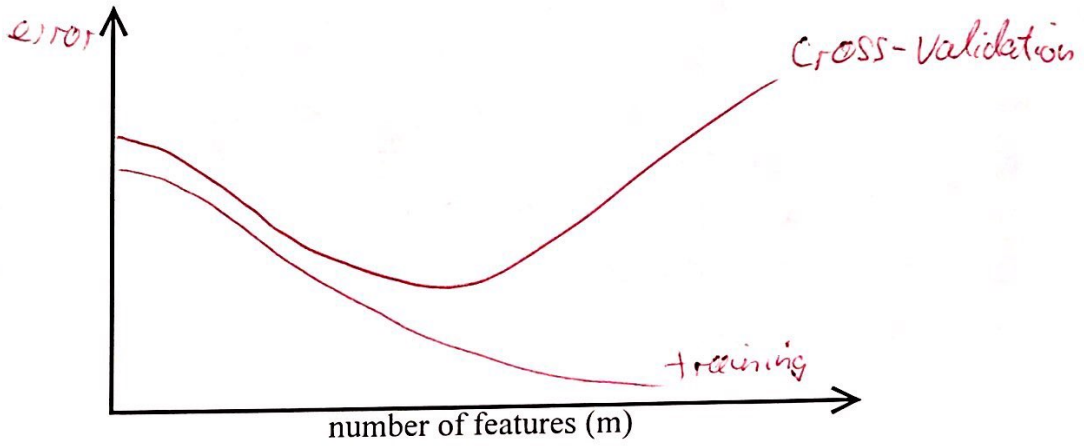
3. **[5 points]** Before we can use a multilayer perceptron for a given task, we have to train it. This training procedure (here: stochastic gradient descent) is composed of different steps, that you can find below. However, the order of the steps is not correct. Please bring the steps in the correct order as it has been done for the first step.

   3 ☐ FOR EACH input vector
   10 ☐ END FOR EACH
   2 ☐ REPEAT until stopping criterion is fulfilled
   11 ☐ END REPEAT
   7 ☐ compute the error of the neurons in the hidden layer
   9 ☐ update the hidden layer weights
   1 1. Initialize all weights
   4 ☐ compute the activation of each neuron of the hidden layer
   8 ☐ update the output layer weights
   6 ☐ compute the error of the output neuron
   5 ☐ compute the activation of the output layer neurons

5 Cross-Validation

1. **[2 points]** Suppose you model the non-linear relationship between a one-dimensional input x and a one-dimensional output y as an mth order polynomial, i.e. $y = w_0 + w_1 x + w_2 x^2 + \ldots + w_m x^m$. The number of training points is fixed, and you estimate the parameters $w_0, w_1, \ldots, w_m$ by linear regression.

   Draw a graph showing two curves: training error vs. the number of features m and cross-validation error vs. the number of features m, annotate both curves.



2. **[2 points]** Find the bugs in the cross-validation algorithm below and correct them.

Algorithm 1: Cross-Validation

Require: Data $\{(x_1, y_1) \ldots, (x_N, y_N)\}$, Number of CV folds F

1: # Split data in F overlapping folds disjoint
2: **for** Fold $f = 1, \ldots, F$ **do**
3: # Train model on folds $\{1, \ldots, F\} \setminus f$ (all folds but f are used for training)
4: # Compute prediction on fold f
5: **end for**
6: **return** average prediction error

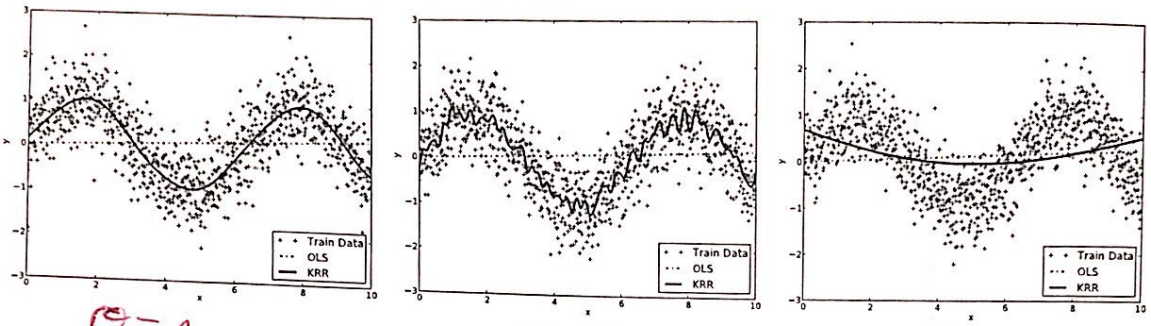
3. **[3 points]** Below you find 3 equal plots with data points. Sketch possible solutions from a polynomial regression, one that underfits (a), overfits (b) and one good fit (c)

(a) underfitting (b) overfitting (c) good fit

6 Kernel methods and Kernel Ridge Regression

1. **[3 points]** We used a Kernel Ridge Regression with a Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$ on training data that follows a sine-function. Below you find the results for three different kernel widths. Indicate which of the following three labels corresponds to which of the plots:

   $\sigma = 10,$ $\sigma = 1,$ $\sigma = 0.1$



$\sigma = 1$ $\sigma = 0.1$ $\sigma = 10$

2. **[1 point]** Explain intuitively how the kernel width σ affects the learned model.

- kernel width is a measure for distance of data points with high influence
- large kernel width → distant data points are considered → under fitting
- small kernel width → only short-distant points considered → over fitting

3. **[3 Points]** You are given the following feature map

$$\phi(x)^\top = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

where $x \in \mathbb{R}^2$.
Show that the dot product defines a kernel function, i.e. show that

$$\phi(x)^\top \phi(y) = k(x, y) = \left(x^\top y\right)^2$$

where again $x, y \in \mathbb{R}^2$.

$$\phi(x)^\top \phi(y) = \begin{pmatrix} x_1^2, \sqrt{2}x_1x_2, x_2^2 \end{pmatrix} \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{pmatrix} = x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2$$

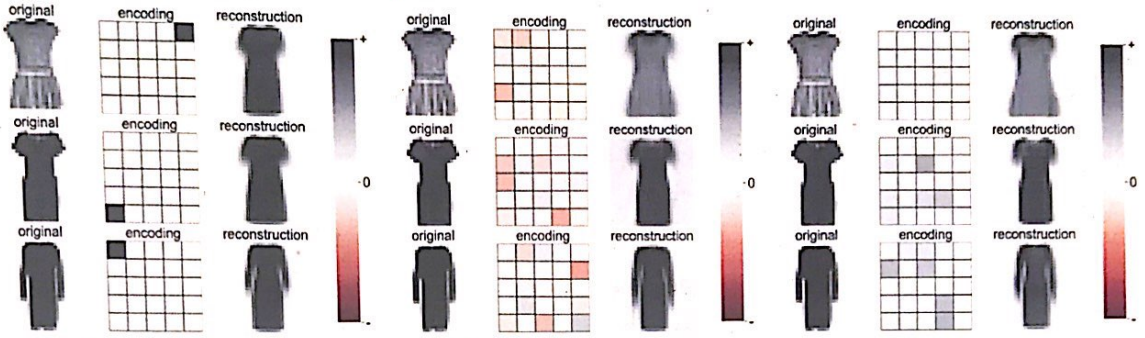
$$k(x,y) = (x^\top y)^2 = \begin{pmatrix} (x_1, x_2) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \end{pmatrix}^2 = (x_1y_1 + x_2y_2)^2 = x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2$$

Same. $\phi(x)^\top \phi(y) = k(x,y)$ □

7 Unsupervised Learning

1. **[3 points]** Fashion-MNIST is a dataset of Zalandos article images – consisting of 70.000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. We applied k-means Clustering (k-means) Nonnegative Matrix Factorization (NMF) and Principal Component Analysis (PCA) on the class *Dress*. Below, you find three figures with the results of those three algorithms. Each figure shows the original image, the encoding (lower dimensional representation) and the reconstruction of 3 exemplary images. Assign the three algorithms to the corresponding figures. Write your answer below the plots.

/ 1 point each



k means PCA NMF

2. **[3 points]** Briefly explain your decision, i.e. what characteristic(s) leads you to the corresponding algorithm. State at least one per algorithm.

/ one point each

k-means : only one dimensional representation

PCA : negative entries

NMF: positive entries and more than 1 non-zero

3. **[3 points]** Write down the steps of PCA up to the point where you receive a lower dimensional representation of your input data (formulas or pseudocode are not necessary but welcome).

1 - compute covariance matrix (C)

.5 - compute eigenvectors / eigenvalues of C

.5 - select eigenvectors corresponding to k largest eigenvalues (W)

1 - project data $H = W^T X$