Foundations of Data Analysis - Midterm WS2021

November 18th 2021, 11:30 - 13:00

Auxiliary materials

This is an open-book exam, which means that you are allowed use any auxiliary material, such as calculators, computer programs, books or notes.

Plagiarism

You have to solve all questions on this exam **alone**. We will perform plagiarism checks on your submission, which may include asking you to answer additional oral questions up to four weeks after the exam.

Examination time

You have 90 minutes to complete the exam. After the exam, you have 20 additional minutes to hand in your exam.

During the exam

You may print this document and answer the questions directly on these pages, or you may use your own paper. Your answers must be hand-written and all plots must be drawn by hand. Make sure all your answers specify which exercise they belong to. **100 points** is the maximum achievable on this exam. You should start by reading all questions and choose which you want to answer first.

Contact and questions

During the entire exam you will be able to ask questions in the Collaborate video chat session posted on moodle below the exam.
Only in the case that you cannot connect to this video chat, you may call this number instead: +43-1-4277-79621

Handing in your exam

You should hand in your exam as a **single pdf file** in the "Final exam" task in moodle. Please name your file `FDA_LastName_StudentID.pdf`. Alternatively you can hand in your exam by email to `anja.meunier@univie.ac.at` with the subject line "FDA Midterm".

If you have technical problems while uploading, contact us immediately and before the deadline! If you do not hand in your exam on time (13:20), you will fail the exam!

Question 1: Multiple choice [31 points]

(a) Which among those below are generative classifiers? [*Circle all that apply*]

A. LDA
B. random forest
C. kernel SVM
D. decision trees
E. logistic regression [/ 5]

(b) The equation of the separating hyperplane for LDA in 1-D, where μ_0 and μ_1 is the mean for each class and $\sigma_0 = \sigma_1 = \sigma$ the variance, and $P(y = 0) = P(y = 1)$, is given by [*Circle one*]

A. $x = \dfrac{\mu_0 + \mu_1}{2}$
B. $x = \dfrac{\mu_0 + \mu_1}{\sigma}$
C. $x = \dfrac{\mu_0 - \mu_1}{\sigma}$
D. $x = \mu_0 - \mu_1$ [/ 2]

(c) Given below is a dataset $\mathcal{S} = \{(x_i, y_i)\}_{i\in\{1,\dots,m\}}$ with $x_i \in \mathbb{N}$ and $y_i \in \{0, 1\}$,

$$\mathcal{S} = \{(1,0),(2,0),(2,0),(3,0),(4,0),(4,1),(5,0),(6,1),(7,1),(7,1),(7,1)\}$$

Consider the set of threshold functions over $\mathbb{N}$, namely $\mathcal{H} = \{h_a = \mathbf{1}_{x>a} : a \in \mathbb{N}\}$. Which of the following hypotheses $h_a \in \mathcal{H}$ minimises the empirical risk if we choose the 0-1 loss function. [*Circle one*]

A. h_3
B. h_4
C. h_5
D. h_6 [/ 3]

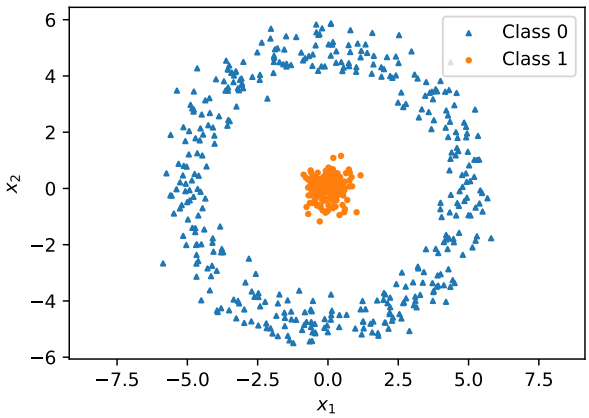
(d) Which of the following statements are true? [*Circle all that apply*]

A. Linear discriminant analysis is used for the binary classification of samples.
B. Logistic regression is capable of working with data that is not linearly separable.
C. Linear discriminant analysis makes use of the l_2 loss function, where $l_2(x, y, h) = (h(x) - y)^2$.
D. Logistic regression is a classification algorithm.
E. Over-fitting can be remedied by increasing the complexity of the hypothesis class
F. K-fold cross-validation is used for tuning the parameters of a hypothesis class.
G. All hypothesis classes with finite VC dimension are PAC learnable.

[/7]

(e) Choose all classifiers that would be capable of achieving an empirical risk of 0 with the following dataset. [*Circle all that apply*]

A. Neural net with 5 hidden layers with linear activation for each layer.
B. Logistic regression
C. SVM with polynomial kernel $k(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^k$ with $k = 1, c \in \mathbb{R}$
D. SVM with polynomial kernel $k(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^k$ with $k = 2, c \in \mathbb{R}$
E. A single decision tree with continuous feature inputs.
F. Linear SVM with features x_1^2 and x_2^2.
G. Linear SVM with features $|x_1|$ and x_2. [/ 7]

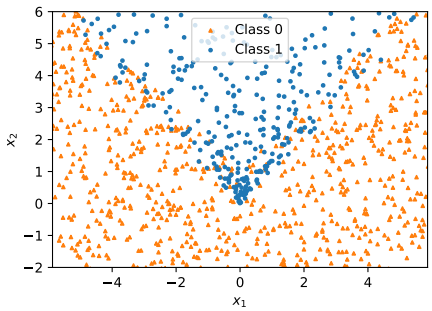


(f) What are the assumptions of linear discriminant analysis (LDA)? [*Circle all that apply*]

A. The prior probabilities for both classes are equal.
B. The covariance matrices of both classes are equal.
C. The means for both classes are identical.
D. The features are normally distributed within each class. [/ 4]

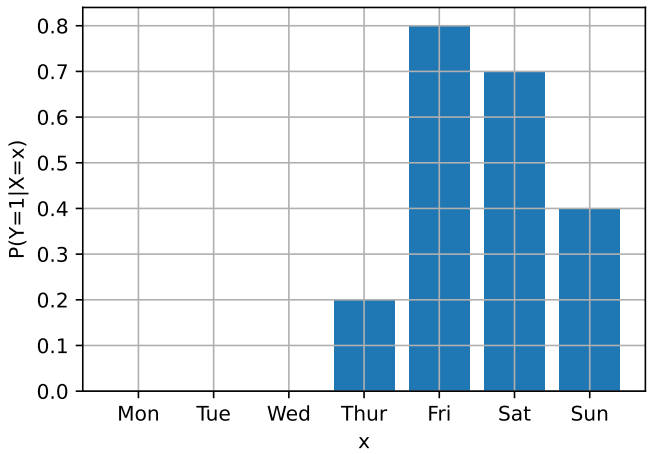
(g) Which would be the most appropriate classifier to use on the following dataset, i.e. the classifier that would generalize best to new data drawn from the same distribution? [*Circle one*]

A. SVM with a soft margin
B. Kernelized SVM with a hard margin
C. SVM with $|x_1|$ and x_2 as features with a soft margin
D. SVM with $|x_1|$ and $|x_2|$ as features with a hard margin [/ 3]



Question 2 [18 points]

(a) The following bar-plot represents the probability of Emma going to the club on a certain day of the week. Here the feature is the day of the week, $x \in \{Mon, Tue, ..., Sun\}$ and the label represents whether they go to the club or not $y \in \{0, 1\}$ ($y = 1$ means Emma goes to the club). Let's say you want to build a classifier the predicts whether Emma goes to the club or not based on the day of the week.



Construct the optimal Bayes classifier $h_B(x)$ for the given data distribution and fill in the table below.

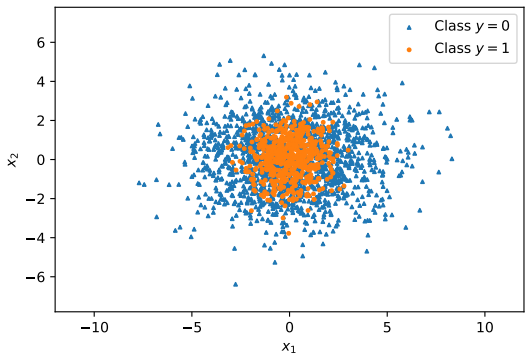
	Mon	Tue	Wed	Thurs	Fri	Sat	Sun
$h_B(x)$							

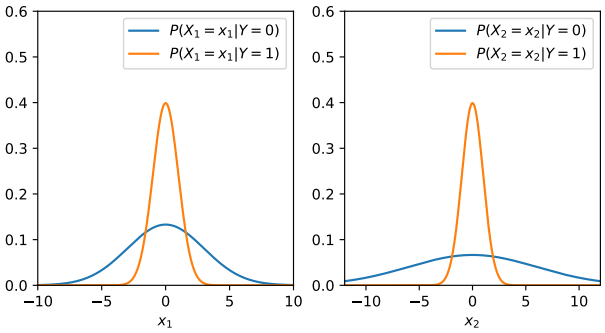
[/ 5]

(b) The Bayes error (also known as Bayes risk) is the minimum error that is achievable by any classifier on a given distribution. The Bayes classifier achieves this minimum risk. Estimate the Bayes error for the classifier constructed in the previous question.

[/ 5]

(c) Consider the plot below where we have m samples in a two-dimensional feature space. In the second figure you see the probability distribution functions for each of the classes along the x_1 and x_2 axis.
Does a Bayes optimal classifier exist for such a distribution? If it does exist, mark the decision boundary (boundaries) on the plots and indicate which regions would be assigned to which class. If it does not exist, explain how the dataset may be altered so that a Bayes classifier exists. Justify your answers with an explanation and, if necessary, equations.





[/ 8]

Question 3 [15 points]

Consider the polynomial Kernel $K(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^k$ for a two dimensional feature space $x \in \mathbb{R}^2$. In this case the inner product refers to the dot product ($\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \boldsymbol{x}^T \boldsymbol{x}'$. Recall that the kernel function can be expressed as an inner-product in a transformed vector space $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \psi(\boldsymbol{x_i}) \psi(\boldsymbol{x_j}) \rangle$ where $\psi : \mathbb{R}^2 \to \mathcal{F}$

(a) For $c \in \mathbb{R}$ and $k = 2$ find the transformation ψ and the space $\mathcal{F}$ it maps into. Express the transformation in the form

$$\psi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} ? \\ \vdots \\ ? \end{pmatrix}$$

Hint: Once you have simplified the expression of the Kernel, try to bring it to the form of a dot product,

$$\begin{pmatrix} d_1^i & d_1^i & \dots \end{pmatrix} \begin{pmatrix} d_1^j \\ d_1^j \\ \vdots \end{pmatrix} = (d_1^i)(d_1^j) + (d_2^i)(d_2^j) + \dots$$

[/ 10]

(b) For values of $c = 0$ and $c \neq 0$, what is the dimensionality of the subspace in $\mathcal{F}$ to which the samples are mapped?

[/ 5]

Question 4 [36 points]

A researcher wants to train a machine learning model which can predict, when their cat will be awake or asleep. To this end, they decide to check on their cat at random times and record the following features and label:

$x_1 \in (0, 24]$	Time of day
$x_2 \in \mathbb{R}^+$	Room temperature
$y \in \{0, 1\}$	Asleep (0) / awake (1)

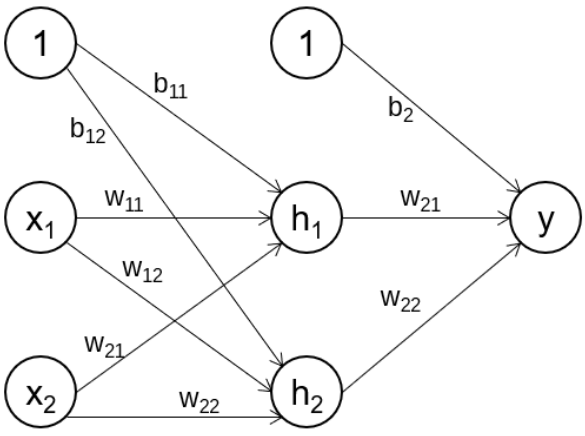
(a) The researcher trains a neural network with two neurons in one hidden layer for this classification problem. The model architecture is shown in the left figure and the corresponding formulae to obtain the values of the hidden neurons h_1, h_2 and the output y are

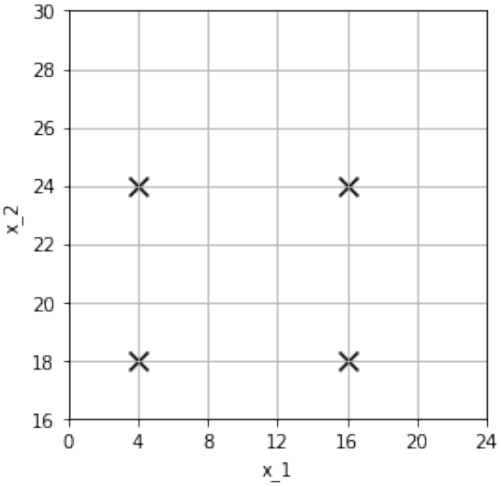
$$h_1(\boldsymbol{x}) = \mathbf{1}_{\{-0.5x_1 - x_2 + 22 \geq 0\}}(\boldsymbol{x})$$
$$h_2(\boldsymbol{x}) = \mathbf{1}_{\{-0.5x_1 - x_2 + 30 \geq 0\}}(\boldsymbol{x})$$
$$y(\boldsymbol{x}) = \mathbf{1}_{\{h_1(\boldsymbol{x}) - h_2(\boldsymbol{x}) - 0.5 \geq 0\}}(\boldsymbol{x}),$$

where

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

is the indicator function. Using the fomulae above, classify the four points in the right figure $\{(4, 18), (4, 24), (16, 18), (16, 24)\}$. Draw the decision boundaries of the hidden neurons h_1 and h_2 in the figure on the right. Then shade the area which is classified as asleep (0) by the neural network model.





[/ 14]

(b) Calculate the VC-dimension of the hypothesis class corresponding to the neural network architecture outlined above.

[/ 10]

(c) As an alternative hypothesis, the researcher suspects that the times during which their cat is awake or asleep follow a strict periodic schedule (i.e. alternating between periods of sleep and waking, all of which have the same duration.) However, they are not sure, what the length of these periods might be. Given that the hypothesis class of sines

$$\mathcal{H}_{\text{sin}} = \{h : h(x_1, x_2) = \text{sign}(\sin(wx_1)), w \geq 0\}$$

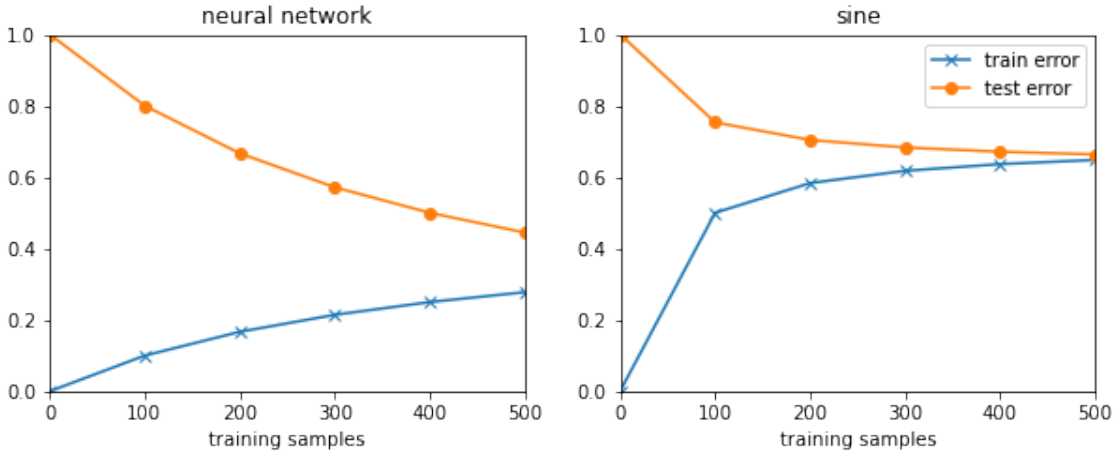
has a VC-dimension of $+\infty$, explain, whether a model obtained via empiric risk minimization on $\mathcal{H}_{\text{sin}}$ would be a reasonable approach to this problem. Justify your answer!

[/ 4]

(d) How could the hypothesis class $\mathcal{H}_{\text{sin}}$ be adapted to obtain a PAC-learnbale hypothesis class $\mathcal{H}'_{\text{sin}}$, which still assumes periodic sleep times?

[/ 4]

(e) The researcher has now trained the neural network, as well as the (adapted) sine model. The figures below show the learning curves of both models. Which model performs better overall? Would you suggest to fit the better model on more training data? Justify your answers!



[/ 4]