Exam Machine Learning for the Quantified Self
26. 06. 2020
8:30 - 11:45

NOTES:

Welcome to the exam of the course Machine Learning for the Quantified Self (XM_40012).

The following tools are permitted:

1. (initially empty) crap paper and pen
2. simple (non graphical) calculator

This is a closed book exam (no other materials are allowed). There are 30 multiple choice questions in total. For each question one correct answer should be selected. Regular questions are awarded 1 point when the correct answer is selected. For correctly answered questions which require more extensive calculations to come to the answer 2 points are given (these are 6 questions in total). In total this means 36 points can be obtained. With random guessing you would obtain 9 points, so this is equal to a 1 as final grade. You pass the exam when you obtain 22.5 points or more (i.e. you obtain half of the points beyond guessing or more).

Avoid ruling out answer purely based on their form (e.g. possible answers are A:5, B:6, C:7, D:none of the above does not mean that D is a less likely answer). Questions about the content of the exam cannot be posed during the exam.

Technical problems?
If you experience technical problems during or prior to your exam, please contact Proctorio via the live chat. This can be done in 2 ways:

- in the lower left corner of your TestVision screen, via the floating widget.
- via the website https://proctorio.com/support (you will see the button 'start live chat' on the website after downloading the extension).

Do not hesitate to make use of Proctorio's chat in case of doubt or problems!

If you have experienced a problem during the exam, we would like to ask you to fill in this form afterwards: https://forms.gle/phDVSm9hJDsmKWbK7. It is then ensured that this message reaches your faculty and the invigilators.

Important:
Do not stop screen sharing during your exam. You will be logged out of your exam. If you are logged out of your exam you can always log back in to your exam by going to TestVision via: https://vu.testvision.nl/online/fe/login_tva.htm

Good luck!

QUESTIONS

1. **Introduction (6 pt)**

   In this COVID-19 era Hugo has become obsessed with tracking apps. He uses an app to track all the people that have been within his vicinity (within 1.5 meters) and also measures how long they were within this distance in minutes. He wants a lot more people to use this app, and also wants all the users to indicate if and when they were infected with the COVID-19 virus. This information can then be used to warn people that might potentially be at risk because they have been in close proximity to infected people.

   (a) **(1 pt)** If we were to use the "Five Factor Framework of Self-Tracking Motivations", which factor would best describe Hugo's motivation?

   (A) improve health
   (B) self-healing → ***correct answer***
   (C) to find new life experiences
   (D) self-entertainment

   Hugo wants to develop a machine learning model which predicts a person being infected with COVID-19 based on the measurements in the app (i.e. who was closeby, how long, and were those people infected).

   (b) **(1 pt)** What kind of machine learning task is this?

   (A) reinforcement learning
   (B) regression
   (C) classification → ***correct answer***
   (D) clustering

   (c) **(1 pt)** Hugo starts collecting data and uses the formal notation we have been using throughout the book, he starts with a dataset of one person, storing only a binary attribute which indicates the presence of a person within 1.5 meters for each time step. Here, "0" indicates no person, "1" one or more persons. Which of the examples below is certainly **not** correct?

   (A) $x_1 = [0]$
   (B) $X = [0, 1, 1]$ → ***correct answer***
   (C) $\mathbf{X} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$
   (D) $X = [person]$

   (d) **(1 pt)** Hugo has become enthusiastic about all lovely sensors on the phone and on top wants to register accelerometer data to detect what kind of activity someone is doing, exercising close together might put one at more risk than sitting on the couch together for example. He is considering two step sizes to aggregate the accelerometer data (i.e. values for Δt): 1 minute, and 20 minutes. Which difference would you expect to see across the accelerometer values we see in the resulting datasets?

(A) The standard deviation of $\Delta t = 1min$ will be lower than the standard deviation of $\Delta t = 20mins$.

(B) The maximum value of $\Delta t = 20mins$ will be higher than for $\Delta t = 1min$.

(C) The number of missing values will be lower for $\Delta t = 1min$ compared to $\Delta t = 20mins$.

(D) None of the above. → ***correct answer***

(e) **(1 pt)** Hugo has by now collected accelerometer data from 5 people for 20 minutes (with a $\Delta t = 1min$) which includes labels of the activities too, and wants to apply machine learning to predict the activity first (later on he plans to use that to work on the task described at the beginning of Question 1). He goes to an AI expert and says he wants to use the latest deep learning model with 10 layers to create a predictive model. The expert warns Hugo that this will not work. Which argument based on supervised learning theory can the AI expert use to back up his statement?

(A) The difference between in sample and out of sample error cannot be guaranteed to be small given the theory of PAC learnability and VC dimensions. → ***correct answer***

(B) The VC dimension of the complex neural network is too low.

(C) The hypothesis space is infinite for the deep neural network, learning theory does therefore not apply at all here, so no guarantees can be given.

(D) None of the above.

(f) **(1 pt)** Consider the following statements:

i. More complex hypothesis spaces result in more overfitting with limited data, but with enough training data (and a complex enough problem) they typically perform better on the test set.

ii. Hyper parameter tuning can be done on the test set.

Which of these statements is correct?

(A) both are correct

(B) only (i) is correct → ***correct answer***

(C) only (ii) is correct

(D) both are incorrect

2. **Outlier Detection (7 pt)**

Consider Figure 1 which shows points in a dataset with two features.

(a) (**1 pt**) We want to apply an outlier detection algorithm to feature X_1 only. Which of the following algorithms is **least** suitable?

(A) Chauvenet's criterion → ***correct answer***

(B) Local Outlier Factor

(C) Mixture Models

(D) Simple Distance-Based Approach

(b) (**1 pt**) We now move to feature X_2 and want to detect outliers. Which of the following algorithms is **not** suitable?

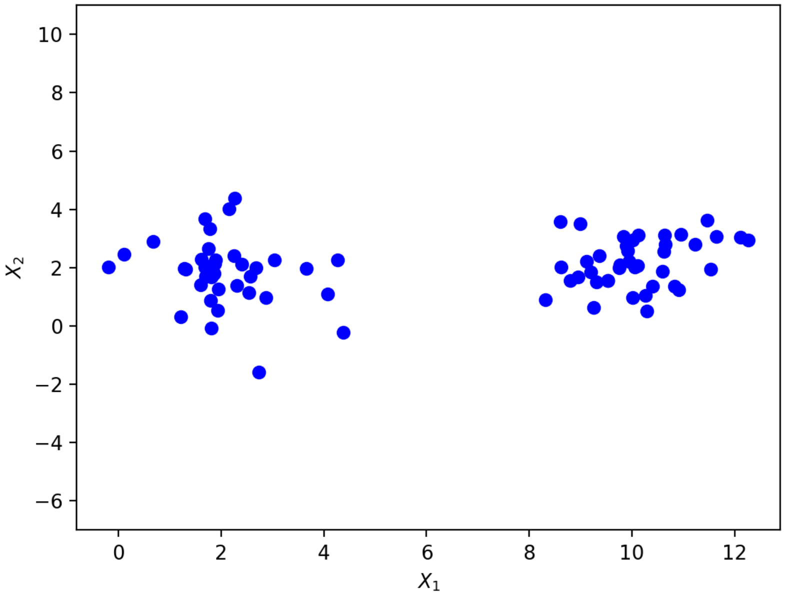

Figure 1: Example dataset - outlier

(A) Chauvenet's criterion

(B) Mixture Models

(C) Simple Distance-Based Approach

(D) None of the above → ***correct answer***

(c) (**1 pt**) In the Local Outlier Factor algorithm, we have the parameter k, which defines the neighborhood. Which statement about the influence of k is correct?

(A) The higher the value for k, the less likely points are considered to be outliers → ***correct answer***

(B) The lower the value for k, the less likely points are considered to be outliers

(C) The value of k does not influence how likely points are considered to be outliers

(D) The value of k follows from the dataset automatically, so we cannot control it

(d) (**1 pt**) Consider the dataset shown in Table 1.

Table 1: Example dataset

Time point	*Heart rate*
0	60
1	
2	
3	?
4	90

What would the value marked with the *?* be in case we impute using simple linear interpolation?

(A) 75

(B) 82.5 → ***correct answer***

(C) 90

(D) None of the above

(e) (**2 pt**) Consider Figure 2.

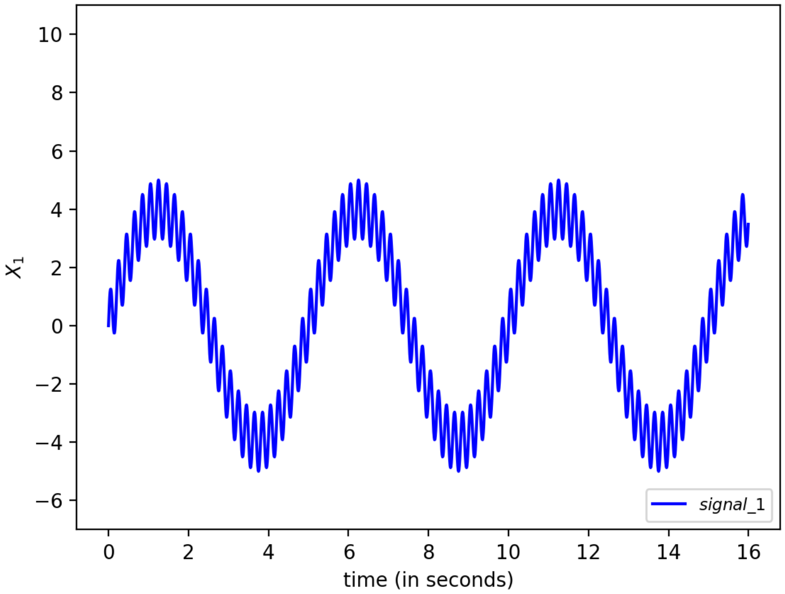

Figure 2: Example dataset - lowpass

We want to get rid of the noise (meaning the high frequency periodic behavior) we see in the dataset. We apply a lowpass filter. To which value should we set the cut-off frequency to filter that noise out?

(A) 10 Hz

(B) 0.1 Hz

(C) 0.3 Hz → ***correct answer***

(D) None of the above

(f) (**1 pt**) Which alternative approach could we use to filter this high frequency noise by considering the notion of time explicitly?

(A) Principal Component Analysis

(B) Kalman Filter → ***correct answer***

(C) Interpolation

(D) Local Outlier Factor

3. **Feature Engineering (9 pt)**

Consider the dataset shown in Table 2.

Table 2: Example dataset

Time point	*Activity_level*	*Mood*
0	Low	Good
1	High	Mwa
2	High	Bad
3	Low	Excellent
4	Low	Bad

(a) **(2 pt)** Assume we apply a window size $\lambda = 1$, what is the support for the pattern $Activity\_level = Low$?

(A) 2/4

(B) 3/4 → ***correct answer***

(C) 3/5

(D) 4/5

(b) **(2 pt)** Assume we apply a window size $\lambda = 1$ and select a minimum threshold for the support of $\Theta = 3/5$ and generate patterns. How many patterns result? Note that we count co-occurs patterns in reverse independent of the order (e.g. $Mood = Good$ (c) $Activity\_level = High$ and $Activity\_level = High$ (c) $Mood = Good$ is the same pattern and counts only once).

(A) 6

(B) 0

(C) 3 → ***correct answer***

(D) None of the above

(c) **(1 pt)** We want to perform a Fourier transformation on our sensory data and assume and select a window size $\lambda = 100$. How many frequencies do we need to explain the measured values in our windows?

(A) 99

(B) 100

(C) 101 → ***correct answer***

(D) None of the above

(d) **(2 pt)** Consider Figure 2 again, without any form of filtering. Assume we apply a Fourier transformation to this complete sequence of 16 seconds. Which frequency do you expect to have the highest value for the amplitude?

(A) 0.1 Hz

(B) 2 Hz

(C) 5 Hz

(D) None of the above → ***correct answer***

(e) **(1 pt)** When composing a final dataset after engineering temporal features we often remove datapoints to avoid too much overlap between each instance in our dataset as we summarize values over time (e.g. we allow a maximum of 50% overlap). How does the increase of our choice for the window size λ impact how many instances we remove given a fixed setting for the amount of overlap allowed?

(A) The larger the window size, the more instances we remove. → ***correct answer***
(B) The smaller the window size, the more instances we remove.
(C) There is not enough information to make a clear statement about that.
(D) None of the above.

(f) **(1 pt)** Put the following components in the right order for an NLP pipeline as we have discussed during the course: (A) Stemming; (B) Stop word removal; (C) Tokenization; (D) Lower case.

(A) CDBA
(B) DCBA
(C) CDAB → ***correct answer***
(D) CBAD

4. **Clustering (6 pt)**

(a) **(1 pt)** We have 5 datasets of different quantified selves. Each of these dataset contain 10 instances. We want to apply a setting we referred in the lecture to as clustering "individual data points". How many datapoints do we have to cluster?

(A) 5
(B) 10
(C) 50 → ***correct answer***
(D) None of the above

Consider two datasets (of different individuals) shown in Table 3.

We want to apply the Dynamic Time Warping (DTW) algorithm to compute the difference between the two time series. Assume we use the absolute difference for comparing two value (e.g. distance between 100 and 80 is 20).

(b) **(2 pt)** What distance should be filled in in Table 4 at the position of the "**?**"?

(A) 20
(B) 40 → ***correct answer***
(C) 100
(D) None of the above

(c) **(2 pt)** What is the value of the shortest path when making the full match?

(A) 20
(B) 40 → ***correct answer***
(C) 120
(D) None of the above

Table 3: Two datasets

Time point	*Heart rate*
Lennart	
1	60
2	60
3	80
4	100
5	80
Tommy	
1	80
2	60
3	80
4	100
5	100

Table 4: DTW answer table

Lennart	t=5 (80)					
	t=4 (100)			?		
	t=3 (80)					
	t=2 (60)					
	t=1 (60)					
		t=1 (80)	t=2 (60)	t=3 (80)	t=4 (100)	t=5 (100)
		Tommy				

(d) **(1 pt)** We want to evaluate the quality of a clustering found and use the silhouette score. We obtain a negative outcome. We want to understand how we could get a negative value. Consider these two statements:

i. This tells us that no clustering can be found that has a silhouette score of 0 or higher.

ii. There are one or more points that are closer to points in another cluster than they are to the points in their own cluster.

Which of these statements is correct?

(A) both are correct

(B) only (i) is correct

(C) only (ii) is correct → ***correct answer***

(D) both are incorrect

5. **Predictive Modeling with Notion of Time (4 pt)**

Below, we will focus on questions related to Predictive Modeling with the Notion of Time.

(a) **(1 pt)** There are several types of Neural Networks that can cope with the notion of time. Given the same number of neurons in the hidden layers, which neural network has the lowest number of weights to be trained?

(A) Regular Recurrent Neural Network
(B) Long Short Term Memory Network
(C) Echo State Network → ***correct answer***
(D) All have an equal number of weights to be trained

(b) **(1 pt)** Select the right answer. When considering the lagged autocorrelation which is relevant for ARIMA models we see consistently high values for this autocorrelation (for any leg) for time series that:

(A) are random
(B) have values that depend on values of one or two time points back
(C) have values that are the cumulative sum over the previous values → ***correct answer***
(D) none of the above

(c) **(1 pt)** Next to Neural Networks we can also use Dynamical Systems Models. In those systems we need to find good values for the parameters of these systems. Which algorithm does parameter optimization using multiple objectives?

(A) NSGA-II → ***correct answer***
(B) Genetic Algorithm
(C) Simulated Annealing
(D) None of the above

(d) (**1 pt**) Consider Figure 3 which shows the errors we obtain on multiple objectives when fitting the parameters of a dynamical systems model.

Which points are **not** dominated?

(A) The cross and the diamond
(B) The circle and the star → ***correct answer***
(C) The star and the diamond
(D) The circle and the diamond

6. **Reinforcement Learning (5 pt)**

We are now going to focus on Reinforcement Learning.

(a) **(1 pt)** Look at the following equation:

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha \cdot (R_{t+1} + \gamma \max_{\mathcal{A}'(S_{t+1})} Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)) \cdot Z_t(S_t, A_t)$$

Which algorithm does this equation belong to?

(A) Q-learning with eligibility traces → ***correct answer***
(B) SARSA with eligibility traces

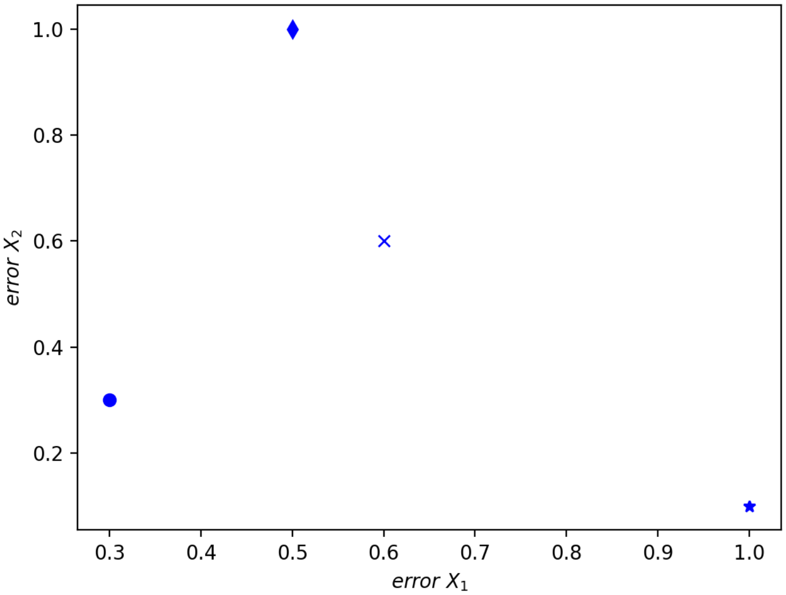

Figure 3: Example error rates on two objectives

(C) Q-learning without eligibility traces

(D) SARSA without eligibility traces

(b) **(1 pt)** Complete the following sentence. SARSA is ...

(A) an on-policy Reinforcement Learning algorithm → ***correct answer***

(B) an off-policy Reinforcement Learning algorithm

(C) the same as Q-learning

(D) not a Reinforcement Learning algorithm

(c) **(2 pt)** We want to use the U-tree algorithm to discretize our state space. We have two features (X_1 and X_2) in the state space that each have a continuous value and find certain Q-values shown in Table 5. Let us ignore the statistical test (the Kolmogorov Smirnov test) that is normally also performed.

Table 5: Example dataset

X_1	X_2	*Q(S,A)-value*
40	20	20
40	40	40
10	40	30

Which feature and branches would be generated in the U-tree algorithm?

(A) X_1 and two branches: ≤10 and >10

(B) X_2 and two branches: ≤20 and >20 → ***correct answer***

(C) X_1 and two branches: $\leq$40 and >40

(D) None of the above

(d) **(1 pt)** In the goal $G(t)$ of Reinforcement Learning the factor γ is present. When we set this value to zero, we ...

(A) focus on long term rewards only.

(B) focus only on instant rewards. → ***correct answer***

(C) will not be able to run a Reinforcement Learning algorithm.

(D) end up with a SARSA algorithm.