Exam Machine Learning for the Quantified Self

25. 06. 2021
9:00 - 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) scrap paper and pen
2. simple (non graphical) calculator

Please show all the permitted materials clearly and completely during the desk scan. Please note that conducting a deskscan is mandatory. Do you think the desk scan you did just now was insufficient? Then please show your desk and all permitted materials now.

Technical issues during the exam?

- Please use the Proctorio live chat in the widget, extension or at proctorio.com/support.
- If Proctorio's live chat cannot help you, then please send an email to onlineproctoring.soz@vu.nl.

Note again that 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 (there are 8 questions in total with 2 points). In total this means 38 points can be obtained. With random guessing you would obtain 9.5 points, so this is equal to a 1 as final grade. You pass the exam when you obtain 23 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.

Good luck!

QUESTIONS

1. **Introduction (7 pt)**

   During this online era due to COVID-19 a significantly larger portion of students have troubles with collaborating effectively with their teammates during courses with group assignments. A group of students that recently followed the course Machine Learning for the Quantified Self has therefore decided to develop an app that is able to match students to form groups by considering their daily rhythms. This app is called *RhythmGrouper*. Their idea is that if daily rhythms are sort of similar this gives a better alignment of working times and thus a higher chance of more intensive collaborations. To identify daily rhythms they exploit a number of sensors, namely the accelerometer, the GPS sensor, and the agenda of the person.

   (a) **(1 pt)** If we were to take the "Five-Factor-Framework of Self-Tracking-Motivation", which one would describe the purpose of students using the app best?

   (A) enhance other aspects of life
   (B) self-design
   (C) self-entertainment
   (D) improve health

   (b) **(1 pt)** One of the components of the app is a model to predict the activity of a person based on the sensory data, in order to identify the rhythm. For this, the group has collected an initial dataset. A choice has to be made on the level of granularity (i.e. Δt) with which the dataset is processed. Consider the following statements.

   i. A higher value for Δt gives us a higher variability for feature values in the dataset.
   ii. A lower value for Δt gives us more instances.

   Which of these statements is correct?

   (A) both are correct.
   (B) only (i) is correct.
   (C) only (ii) is correct.
   (D) both are incorrect.

   (c) **(1 pt)** We want to identify a particular feature (the first one) in the dataset resulting from *RhythmGrouper*. Using what mathematical term do we identify this feature?

   (A) X_1
   (B) $\mathbf{X_1}$
   (C) x_1^1
   (D) None of the the other answers

   (d) **(1 pt)** Next to suggesting students to team up, the *RhythmGrouper* app also helps once the group is formed. It provides messages to students when they should start working on their group assignments (there are two messages: (1) work soon, as in get ready to work, and (2) work now). The timing of these messages is determined

using the sensory measurements and optimized with the long term goal to obtain the highest grade possible. A Reinforcement Learning algorithm is used to learn the timing of these messages. What is the term used in Reinforcement Learning for the sensory measurements?

(A) state space for the reinforcement learning algorithm.

(B) action space for the reinforcement learning algorithm.

(C) reward structure for the reinforcement learning algorithm.

(D) eligibility trace for the reinforcement learning algorithm.

(e) **(2 pt)** Consider the MDP shown in Figure 1 for the problem that has just been described. What is the value that is eventually learned for *Q(inactive, work_soon)* given a value $\gamma = 1$?

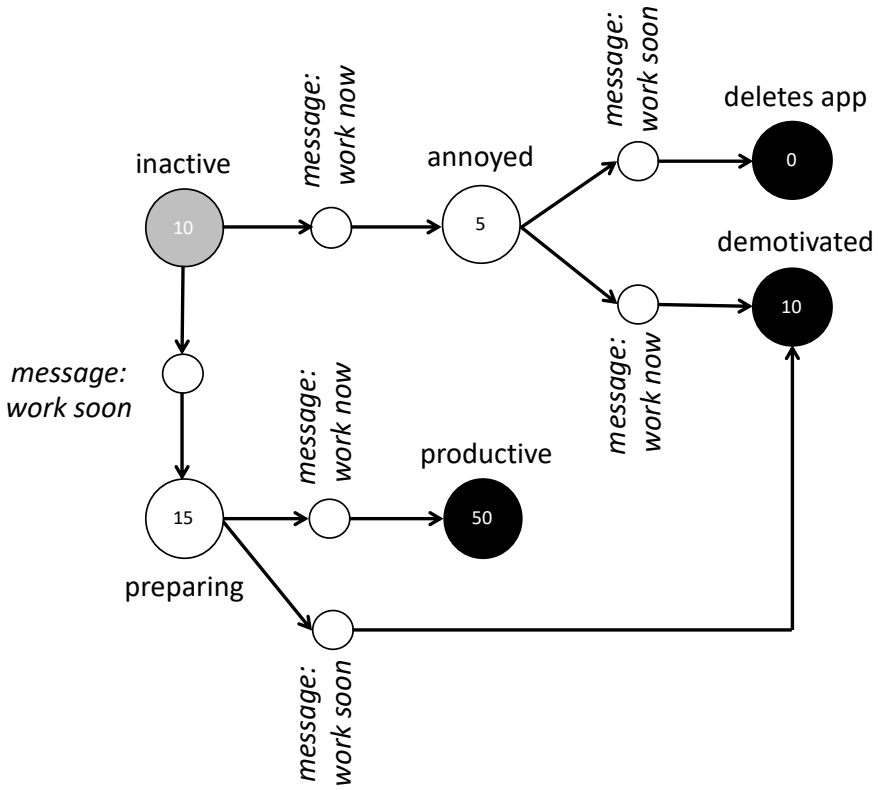

Figure 1: MDP for *RhythmGrouper*

(A) 50

(B) 65

(C) 75

(D) None of the other answers

(f) **(1 pt)** What is the purpose of eligibility traces?

(A) To take into account prior occurrence of state-action pairs in the update of the Q-values

(B) To discretize a continuous state space

(C) To allow the algorithm to train on multiple states at the same time

(D) None of the other answers.

2. **Outlier Detection (8 pt)**

This part concerns outlier detection and removal of noise.

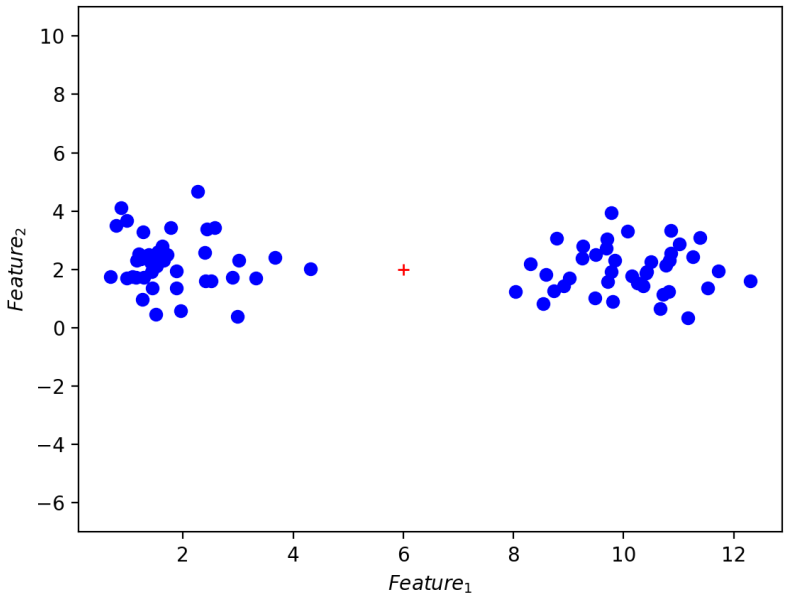

Figure 2: Example outliers

(a) (**1 pt**) Consider Figure 2. When focusing on the red plus and only $Feature_1$, which approach would certainly **not** consider this point as an outlier?

(A) Local Outlier Factor

(B) Simple Distance Based

(C) Chauvenet's criterion

(D) None of the other answers

(b) (**2 pt**) Let us now apply the simple distance-based outlier detection to the data shown in Figure 2 using both $Feature_1$ and $Feature_2$. The figure contains 81 points. We have two parameters to set, d_{min} and f_{min}. Which of the following settings would result in the red plus being an outlier?

(A) $d_{min} = 2$ and $f_{min} = 1$

(B) $d_{min} = 1$ and $f_{min} = 0.5$

(C) $d_{min} = 10$ and $f_{min} = 0.5$

(D) None of the other answers

(c) (**1 pt**) Consider the following statements about the Local Outlier Algorithm:

i. A higher value for the local reachability distance for a point means a lower chance of the point being considered an outlier.

ii. A higher value for the local outlier factor for a point means a lower chance of the point being considered an outlier.

Which of these statements is correct?

(A) both are correct.

(B) only (i) is correct.

(C) only (ii) is correct.

(D) both are incorrect.

(d) (**1 pt**) Let us consider the Kalman filter to detect outliers and impute missing values. Which term represents our estimation of the latent state at t when taking the values for the observations at t into account?

(A) $\hat{s}_{t|t-1}$

(B) $s_{t|t-1}$

(C) $\hat{s}_{t|t}$

(D) None of the other answers

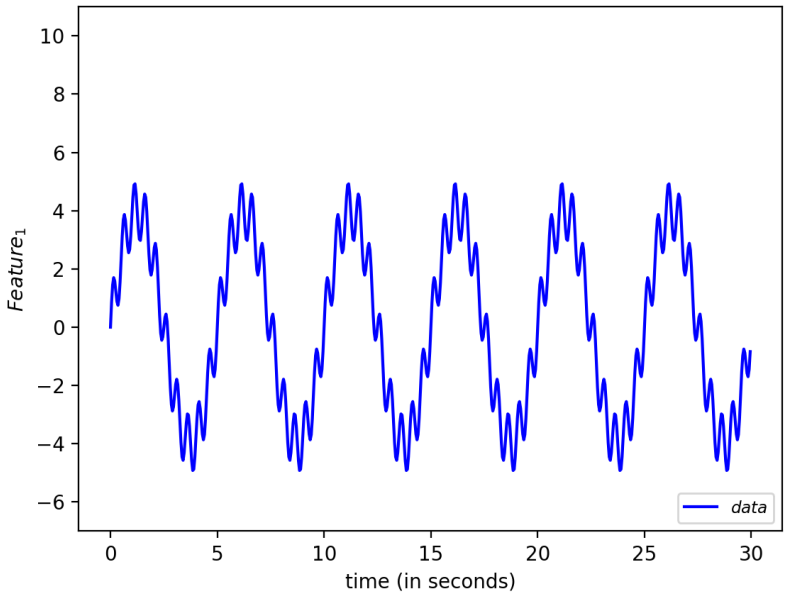

Figure 3: Example time series

(e) (**2 pt**) Consider Figure 3. We see a signal with two different frequencies combined here, we want to filter out the one with the frequency of 0.2 Hz. Which technique could we use for this purpose?

(A) Lowpass filter

(B) Interpolation

(C) Principal Component Analysis
(D) None of the other answers

(f) (**1 pt**) Which approach helps us to resolve missing values?
(A) Lowpass filter
(B) Interpolation
(C) Principal Component Analysis
(D) Local Outlier Factor

3. **Feature Engineering (6 pt)**

This part concerns feature engineering.

(a) (**1 pt**) Consider Figure 3 again. We will now apply a Fourier transformation to the data with $\lambda + 1 = 3$ seconds. Which frequency/frequencies will have an amplitude > 0 when we do not consider a frequency of 0 Hz?
(A) 0.2 Hz and 2 Hz
(B) only 0.2 Hz
(C) only 2 Hz
(D) none of the other answers

Table 1: Example dataset

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

(b) **(2 pt)** Consider the dataset shown in Table 1. Let us focus on the algorithm of Batal *et al.* as explained during the course. Assume we apply a window size $\lambda = 1$, what is the support for the pattern $Mood = Bad$?
(A) 4/5
(B) 3/4
(C) 4/4
(D) None of the other answers

(c) **(2 pt)** We continue with the algorithm of Batal *et al.*. Assume we apply a window size $\lambda = 1$ and select a minimum threshold for the support of $\Theta = 0.6$ and generate patterns. How many patterns result? Note that we count co-occurs patterns independent of the order (e.g. $Mood = Good$ (c) $Activity\_level = Low$ and $Activity\_level = Low$ (c) $Mood = Good$ is the same pattern and counts only once).
(A) 2
(B) 3

(C) 4

(D) None of the other answers

(d) **(1 pt)** How do we handle mixed data (i.e. continuous and categorical data) in combination with Batal *et al.*'s algorithm to derive temporal features?

(A) We can just use the continuous values directly in the algorithm.

(B) We categorize the continuous feature values first, and then apply the standard algorithm.

(C) The algorithm can never work with continuous data, even not with a pre-processing step.

(D) None of the other answers.

4. **Clustering (8 pt)**

This part concerns clustering approaches.

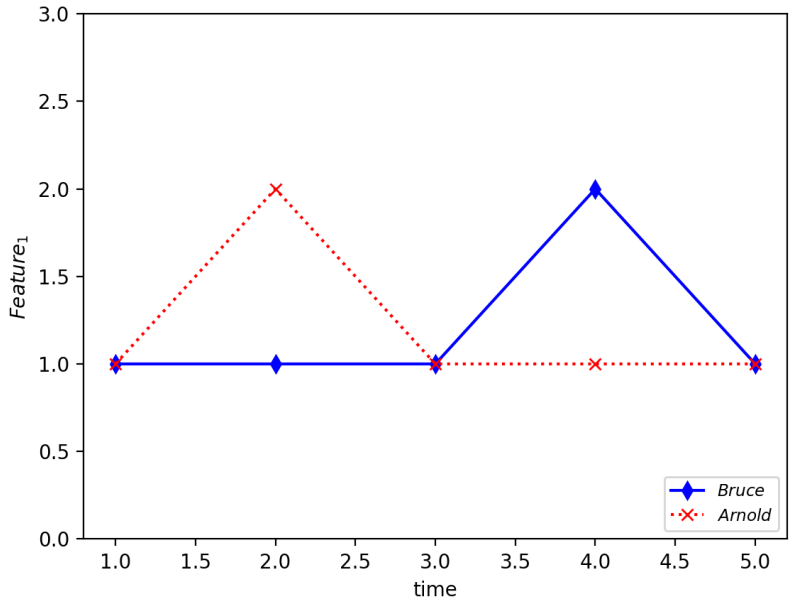

Figure 4: Time series

(a) **(2 pt)** Look at Figure 4, where measurements of both Arnold and Bruce for $Feature_1$ are shown with the measured values indicated by crosses for Arnold and diamonds for Bruce. We want to compute the distance using the Euclidean distance where we take the temporal notion into account. What is the value that results from this computation over the entire time series?

(A) 1

(B) $\sqrt{2}$

(C) 2

Table 2: DTW answer table

Arnold						
	t=5					
	t=4					
	t=3					
	t=2		**?**			
	t=1					
		t=1	t=2	t=3	t=4	t=5
		Bruce				

(D) None of the other answers

(b) **(2 pt)** Rather than using the Euclidean distance we now use Dynamic Time Warping to compute the distance. To compute this, you need to create a table as shown in Table 2. Assume we use the absolute difference for comparing two values (e.g. distance between 1 and 2 is 1). What distance should be filled in in Table 2 at the position of the "**?**"?

(A) 0
(B) 1
(C) 2
(D) None of the other answers

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

(A) 0
(B) 1
(C) 2
(D) None of the other answers

(d) **(1 pt)** When we want to cluster on a person level, which algorithm is least suitable?

(A) Agglomerative clustering
(B) K-means clustering
(C) Divisive clustering
(D) K-medoids clustering

(e) (**1 pt**) Consider the following statements about the Subspace Clustering algorithm:

i. A higher value for ϵ will result in more units.
ii. Subspace clustering always uses all features in the dataset to form clusters

Which of these statements is correct?

(A) both are correct.
(B) only (i) is correct.
(C) only (ii) is correct.
(D) both are incorrect.

5. **Learning Theory and Predictive Modeling without Notion of Time (4 pt)**

Below, we will focus on questions related to learning theory as well as predictive modeling without the notion of time.

(a) **(1 pt)** When we consider PAC learnability, what does this theory say when we consider two datasets (for the same problem and we have the same number of hypotheses) where one dataset has N_1 samples and the other dataset has N_2 samples with N_1 being much smaller than N_2

(A) The difference between the in sample error and out of sample error is larger in the case of N_1

(B) The difference between the in sample error and out of sample error is larger in the case of N_2

(C) The VC dimension of N_1 is larger

(D) The VC dimension of N_2 is larger

(b) **(1 pt)** Assume we have a dataset $\mathbf{X}$ of size 3 and we have a binary classification problem. What is the number of elements in the restriction H_X on a hypothesis set H?

(A) 3

(B) 8

(C) There is not enough information to compute the size of the restriction.

(D) Restrictions do not apply to binary classification problems.

(c) **(1 pt)** We are applying a machine learning algorithm which does not consider the temporal dimension explicitly. Hereby, we use temporal features. To avoid too much overlap between instances we decide to set a limit to the amount of overlap allowed. We try out two different values for the overlap we allow: 50% and 90%. We sample data randomly in our training-and test set and see performance is much better for the case of 90% overlap. Consider the following explanations:

i. There is a lot of overlap between the instances in the training and testset in the setting of 90%, therefore the assessment of the generalizability might be overly optimistic.

ii. For the case of 90% there are more instances to learn from, that could result in better performance.

Which of these explanations could be correct?

(A) both

(B) only (i)

(C) only (ii)

(D) both are not correct

(d) **(1 pt)** For which one of the following can feature selection **not** be used?

(A) Reducing computation time of machine learning algorithms.

(B) Reducing overfitting.

(C) Having more insightful models.

(D) Hyperparameter optimization.

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

We are now going to focus on predictive modeling approaches which take the notion of time into account explicitly.

(a) **(1 pt)** For which parameter in the ARIMA model can we use the Partial Autocorrelation Function to find the appropriate value for that parameter?

(A) p
(B) q
(C) d
(D) r

(b) **(1 pt)** For an Echo State Network we have the following weight matrices: $\mathbf{W^{IN}}$ of size $n \times p$, $\mathbf{W}$ of size $n \times n$, and $\mathbf{W^{OUT}}$ of size $l \times n$. How many weights need to be trained in this network?

(A) $l \times n$
(B) $l \times n + n \times n$
(C) $l \times n + n \times n + n \times p$
(D) $n \times n$

(c) **(1 pt)** What is the purpose of the temperature parameter in Simulated Annealing as explained during the lecture?

(A) To change the balance between exploration and exploitation during a run by gradually moving more and more towards exploitation.
(B) To change the balance between exploration and exploitation during a run by gradually moving more and more towards exploration.
(C) To change the size of the updates of the weights during a run by gradually moving more and more towards bigger jumps.
(D) To change the size of the updates of the weights during a run by gradually moving more and more towards smaller jumps.

(d) **(1 pt)** Consider that we apply the Simple GA and we get two parents: $[0, 0, 1, 1]$ and $[1, 0, 0, 1]$. What could be a child of these two parents after crossover?

(A) $[0, 0, 0, 1]$
(B) $[0.5, 0, 0.5, 1]$
(C) $[1, 0, 1, 0]$
(D) $[1, 1, 1, 1]$

(e) **(1 pt)** Which algorithm is focused on optimizing parameters based on multiple objectives?

(A) NSGA-II
(B) Simple GA
(C) Simulated Annealing
(D) None of the other answers