Exam Machine Learning for the Quantified Self
20. 08. 2021
8:30 - 11:15

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 6 questions in total with 2 points). 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 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)**

   In these extraordinary times people tend to spend more time at home. As a result, the popularity of board games has gone up substantially. Kevin is a Quantified self enthusiast and this increase in board game playing has sparked his enthusiasm. He has decided to start developing an app around board games. The app collects sensory data from a wrist band with a variety of sensors (accelerometer, heart rate, skin conductance) and combines this with sensory information from the phone (accelerometer, magnetometer, and gyroscope). The goal of the app is twofold: (1) determine when someone is playing a board game and if so, what board game the person is playing, and (2) determine whether someone is winning the game or not.

   (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 → ***correct answer***
   (D) improve health

   (b) **(1 pt)** Kevin expects the wrist band to collect valuable information on the characteristics of the game being played, e.g. how often and how far tokens are moved, whether cards are picked up, etc. He is in doubt on what value to select for Δt for processing his dataset. Consider the following statements:

   i. A higher value for Δt gives us a more fine grained representation of the sensory data.
   ii. A lower value for Δt gives us a lower number of 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. → ***correct answer***

   (c) **(1 pt)** To train the app properly, we ask users to indicate the board game they are playing. Hereby, Kevin decides to limit the board games of interest to the 10 most popular ones. What is the notation for the data we collect on the labels that is used in the course?

   (A) **X**
   (B) **Y**
   (C) **G** → ***correct answer***
   (D) None of the the other answers

(d) **(1 pt)** Which learning algorithm would **not** be suitable for the task to predict which board game is being played?

(A) linear regression → ***correct answer***

(B) multi-layer perceptron

(C) support vector machine

(D) naive bayes

(e) **(1 pt)** We want to start training a learning algorithm and decide to ignore the order in which people play games over time (while of course still using features that exploit the temporal dimension) and focus on generalizability to new users of the app. According to the terminology in the book, which learning setup would match this scenario best?

(A) Individual level temporal

(B) Individual level non-temporal

(C) Population level with unknown users → ***correct answer***

(D) Population level with unseen data of known users.

(f) **(1 pt)** Let us move to learning theory. Imagine that our collected data only consists of a fixed number of categorical features (and a categorical target) and we apply a decision tree learning algorithm. What can we say about the number of hypotheses?

(A) This is infinite.

(B) This is finite.→ ***correct answer***

(C) We cannot say whether it is infinite or not, this depends on the hyperparameter settings of the decision tree algorithm.

(D) We cannot say whether it is infinite or not, this depends on the VC dimension.

(g) **(1 pt)** How does the number of hypotheses influence the difference between the in-sample and out-of-sample error according to PAC learnability given a fixed size of the dataset and a finite set of hypotheses?

(A) The larger the number of hypotheses, the smaller the difference between the in-sample and out-of-sample error.

(B) The larger the number of hypotheses, the bigger the difference between the in-sample and out-of-sample error. → ***correct answer***

(C) PAC learnability cannot be applied to a finite number of hypotheses.

(D) None of the other answers.

2. **Outlier Detection (6 pt)**

This part concerns outlier detection and removal of noise.

(a) (**1 pt**) Consider the following statements about outlier detection algorithms:

i. The higher we set the value of c in Chauvenet's criterion, the more points we will identify as outlier.

ii. The higher that value for d_{min} in the simple distance-based outlier detection algorithm, the more points we will identify as 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. → ***correct answer***

Consider the data shown in Figure 1.

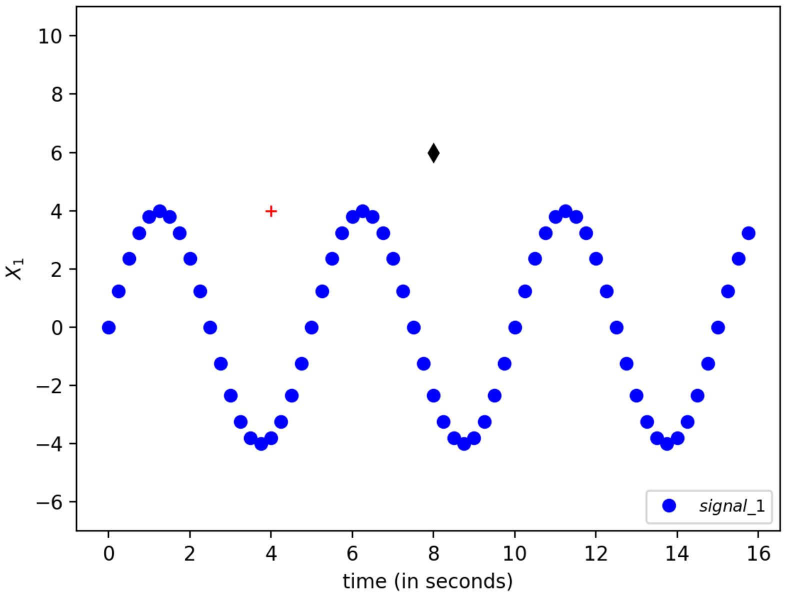

Figure 1: Example dataset - outlier

(b) (**2 pt**) Which points would be likely to be considered outliers when we would apply the local outlier factor algorithm to the data shown in Figure 1?

(A) the point represented by the plus

(B) the point represented by the diamond → ***correct answer***

(C) the points represented by the plus and the diamond

(D) none of the points

(c) (**1 pt**) Imagine that we apply the Kalman filter to the data shown in Figure 1. Which component of the filter will be influenced most by seeing the points indicated by the plus and diamond?

(A) $P_{t|t}$ → ***correct answer***

(B) H_t

(C) B_t

(D) F_t

(d) (**1 pt**) We want to apply a lowpass filter to the figure shown and decide to select a cut-off frequency $f_c = 0.15Hz$. What would the data after application of the filter look like?

(A) The data would be the same, except that the points represented by the plus and diamond would be changed.

(B) The blue points (i.e. all points except for the plus and diamond) would be changed. → ***correct answer***

(C) The data would remain unchanged.

(D) All points in the data would be changed.

(e) (**1 pt**) When we consider the data shown in Figure 1 and we would like to impute missing values, which approach would be least appropriate?

(A) Kalman filter

(B) Interpolation

(C) Mean imputation → ***correct answer***

(D) All are equally suitable

3. **Feature Engineering (8 pt)**

This part concerns feature engineering.

(a) (**1 pt**) When we consider a window size of $\lambda = 10$ for a particular feature X_1, which approach would result in the most added features in our dataset?

(A) Time domain using the mean

(B) Time domain using the standard deviation

(C) Frequency domain using the amplitude of each frequency considered → ***correct answer***

(D) Frequency domain using the power spectral entropy

(b) (**2 pt**) Consider the dataset shown in Table 1.

Table 1: Example dataset

Time point	*Heart rate*	*Temporal feature*
0	60	
1	70	
2	80	
3	80	?
4	90	

We want to create temporal features in the time domain for the feature *Heart rate* by averaging using a window size $\lambda = 2$. What is the value for this temporal feature at time point 3 (i.e. the value at the "?").

(A) 80

(B) $76\frac{2}{3}$ → ***correct answer***

(C) 70

(D) none of the other answers

(c) (**1 pt**) Fill in the right word in the following sentence: in the algorithm of Batal *et al.* increasing the window size λ will ..... the number of found patterns.

(A) most likely increase → ***correct answer***

(B) most likely decrease

(C) certainly not influence

(D) certainly increase

(d) (**2 pt**) Consider the data shown in Figure 2.

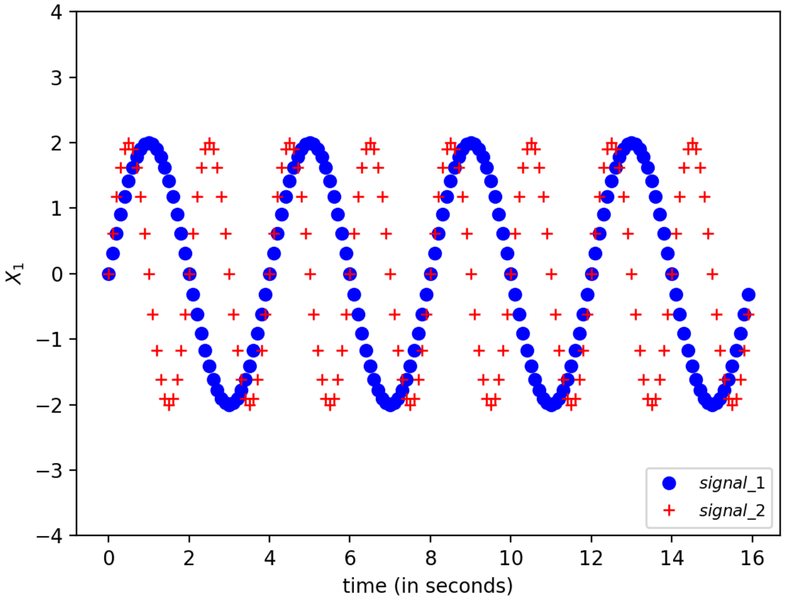

Figure 2: Example dataset - temporal

We see two signals (signal 1 and signal 2). We apply a Fourier transformation to both signals and identify a feature that takes as value the frequency with the highest amplitude. Which signal would have the highest value for this feature?

(A) signal 1

(B) signal 2 → ***correct answer***

(C) both signals have the same value

(D) there are multiple values per signal, therefore they cannot be compared using a single value

(e) (**1 pt**) We now consider text based data. Which of the following algorithms would normally result in the least number of features that we add to our dataset?

(A) Topic modeling → ***correct answer***

(B) Bag of words

(C) TF-IDF

(D) All increase the number of features with the same amount

(f) **(1 pt)** When we are done with the temporal feature engineering we might have substantial overlap between different instances. To avoid too much overlap we set a limit to the amount of overlap allowed. We are in doubt between two different values for the overlap we allow: 50% and 90%. Consider the following statements:

i. In the case of 50% overlap we will end up with more instances compared to the case with 90% overlap.

ii. It is more likely to find highly similar instances in the case of the 90% overlap compared to the 50% overlap case.

Which of these explanations is correct?

(A) both are correct
(B) only (i) is correct
(C) only (ii) is correct → ***correct answer***
(D) both are not correct

4. **Clustering (5 pt)**

This part concerns clustering approaches.

(a) (**1 pt**) When we apply clustering on a person level and have 10 datasets, each covering different persons and each of these datasets containing 100 datapoints. How many points do we have to cluster?

(A) 10 → ***correct answer***
(B) 100
(C) 1000
(D) None of the other answers

(b) **(2 pt)** Look at Figure 3, 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 cross correlation coefficient. We allow for a τ of either zero or one. What is the value of the cross correlation coefficient with the best value for τ?

(A) 2
(B) 5
(C) 7 → ***correct answer***
(D) None of the other answers

(c) **(1 pt)** In Dynamic Time Warping we have several so-called conditions. Which condition states that the time order should be preserved?

(A) The boundary condition
(B) The time constraint condition
(C) The monotonicity condition → ***correct answer***
(D) None of the other answers

(d) (**1 pt**) What is the term dense units used for in subspace clustering?

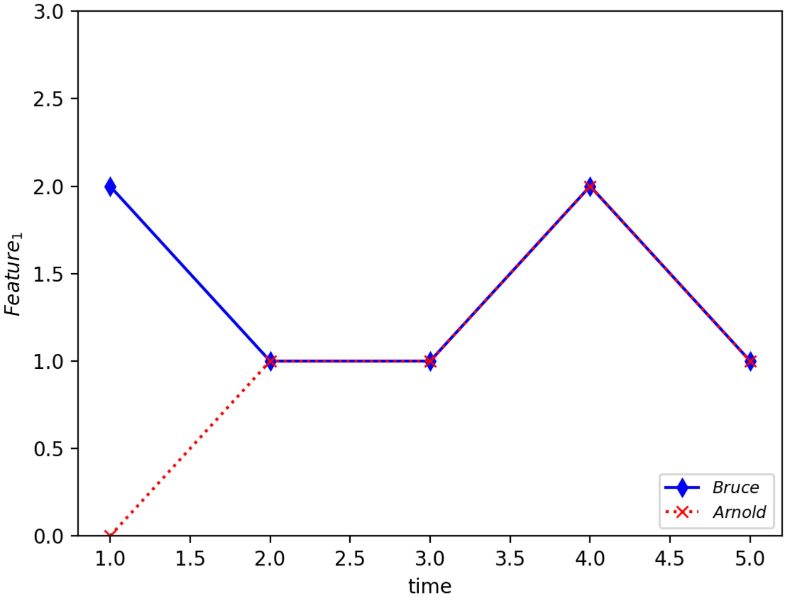

Figure 3: Time series

(A) That indicates whether a unit contains a number of datapoints that exceeds some threshold. → ***correct answer***

(B) That represents the number of connections to other units that contain at least one datapoint.

(C) That represents how many units are created for each feature.

(D) None of the other answers.

5. **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)** Which parameter in the ARIMA model is concerned with differencing?

(A) p

(B) q

(C) d → ***correct answer***

(D) r

(b) **(1 pt)** What is the advantage of LSTM networks over regular RNNs?

(A) They handle the vanishing gradient problem better. → ***correct answer***

(B) They have a simpler structure for the neurons.

(C) They are computationally more efficient.

(D) None of the other answers.

(c) **(1 pt)** Which one of the following algorithms does not use back propagation?

(A) Echo State Networks → ***correct answer***

(B) LSTM networks

(C) RNNs

(D) Multi-Layer Perceptrons

(d) **(2 pt)** Imagine the following dynamical systems model of the relationship between stamina (how much endurance a person has) and the intensity of activities being conducted:

$$\hat{y}_{stamina}(t + \Delta t) = y_{stamina}(t) + \gamma \cdot (y_{activity\_level}(t) - y_{stamina}(t)) \cdot \Delta t \tag{1}$$

$$\hat{y}_{activity\_level}(t + \Delta t) = y_{activity\_level}(t) + \gamma \cdot \Delta t \tag{2}$$

The model basically says that *stamina* increases when the *activity level* is above the current *stamina*. The precise change depends on the parameter γ. The *stamina* decreases when the *activity level* is below the current *stamina*. Furthermore, the *activity level* increases with a fixed value γ. We assume a setting of $\Delta t = 1$. In addition, we have collected a dataset shown in Table 2 about the values for *stamina* and the *activity level*. Finally, we assume the absolute difference to be used as a distance metric (i.e. $E(target) = \sum_{t=0}^{N} |\hat{y}_{target}(t) - y_{target}(t)|$). At time point 0 we assume the model is initialized with the measured values at that time point.

Table 2: Example dataset

Time point	*Stamina*	*Activity level*
0	0.5	0.1
1	0.4	0.2
2	0.3	0.3

Let us consider two setting for γ: $\gamma = 0.1$ and $\gamma = 0.2$. Compute the error for both settings with respect to each objective. What can we say about the dominance of one parameter setting over the other when considering Pareto Efficiency?

(A) $\gamma = 0.1$ dominates $\gamma = 0.2$.

(B) $\gamma = 0.2$ dominates $\gamma = 0.1$.

(C) $\gamma = 0.1$ and $\gamma = 0.2$ do not dominate each other. → ***correct answer***

(D) There is insufficient information to calculate the dominance of one over the other.

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

Below, we will focus on questions related to Reinforcement Learning.

(a) **(2 pt)** We focus on a case where we want to support people that are hopelessly out of shape to regain their shape again by sending them messages to coach them in their work out endeavors. The MDP which includes the states, actions, and rewards is shown in Figure 4. We see two actions, namely an advice to work out and an

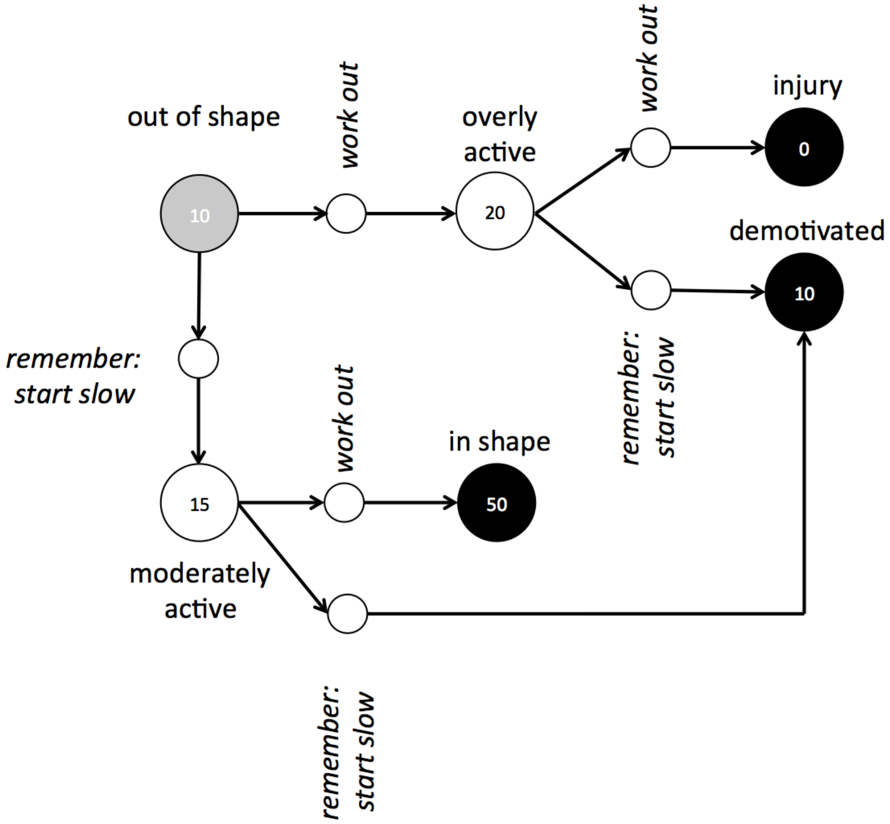

Figure 4: MDP

advise to remember: start slow. What is the value that is eventually learned for $Q(moderately\_active, remember : startslow)$?

(A) 10 → ***correct answer***

(B) 25

(C) 50

(D) None of the other answers

(b) **(1 pt)** The number of states and actions can be very large in Reinforcement Learning. As a result, storing the Q-values for all possibilities might not be possible. What could be a solution to this problem?

(A) Use eligibility traces.

(B) Use SARSA instead of Q-learning.

(C) Make the system such that the Markov property is satisfied.

(D) Use a neural network to predict the values. → ***correct answer***

(c) **(1 pt)** What is the main difference between SARSA and Q-learning?

(A) Q-learning is on-policy while SARSA is off-policy learning.

(B) Q-learning is off-policy while SARSA is on-policy learning. → ***correct answer***

(C) Q-learning uses a table to store the values in while SARSA uses a model.

(D) None of the other answers.

(d) **(1 pt)** What does the Markov property entail?

(A) That the MDP is completely deterministic.

(B) That the probability of moving to a future state only depends on the current state and action. → ***correct answer***

(C) That the probability of moving to a future state depends on the entire history of state and actions, not just the current one.

(D) That the state space is discrete.