Exam Machine Learning for the Quantified Self
16. 07. 2020
12:00 - 14:45

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 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 (7 pt)**

   In the past months with the lockdowns in nearly all countries, the amount of mental health problems has substantially increased, while access to therapists has become increasingly more difficult. To battle these problems, mobile apps have penetrated the market that give people exercises and training how to improve their mental health. *MoodCompanion* is one of these apps. It uses various measurements on the mobile phone to understand someone's context and provide suggestions based on that (e.g. when someone is lying on the couch the whole day the app suggest to perform an outdoor activity). Bruce is one of the users of *MoodCompanion*, using it to battle his mental health problems that have only been worsened by the lockdown situation.

   (a) **(1 pt)** If we were to take the three categories identified by Choe *et al.*, which factor would best describe Bruce's motivation?

   (A) improve health → ***correct answer***
   (B) self-healing
   (C) self-design
   (D) self-entertainment

   (b) **(1 pt)** In order for *MoodCompanion* to use the context of the user in a good way, the developers first trained a machine learning model to predict the type of activity someone is conducting (e.g. lying on the couch, walking in the park, etc.). This activity is being predicted solely based on the accelerometer data (x, y, z-axis) of the smartphone. What can we say about the table $\mathbf{X}$ that is used to train this model?

   (A) The number of rows in $\mathbf{X}$ is 3.
   (B) The number of rows in $\mathbf{X}$ is 4.
   (C) The number of columns in $\mathbf{X}$ is 3. → ***correct answer***
   (D) The number of columns in $\mathbf{X}$ is 4.

   (c) **(1 pt)** When developing the model to predict the activity using the accelerometer data, a choice had to be made between for 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 more instances in our dataset.
   ii. A lower value for Δt gives us more missing values.

   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.

   (d) **(1 pt)** The task of finding the right suggestion for an activity given a certain context is a reinforcement learning problem. Complete the following sentence. The suggestions that can be given for the activity is the ....

(A) state space for the reinforcement learning algorithm.
(B) action space for the reinforcement learning algorithm. → ***correct answer***
(C) reward structure for the reinforcement learning algorithm.
(D) eligibility trace for the reinforcement learning algorithm.

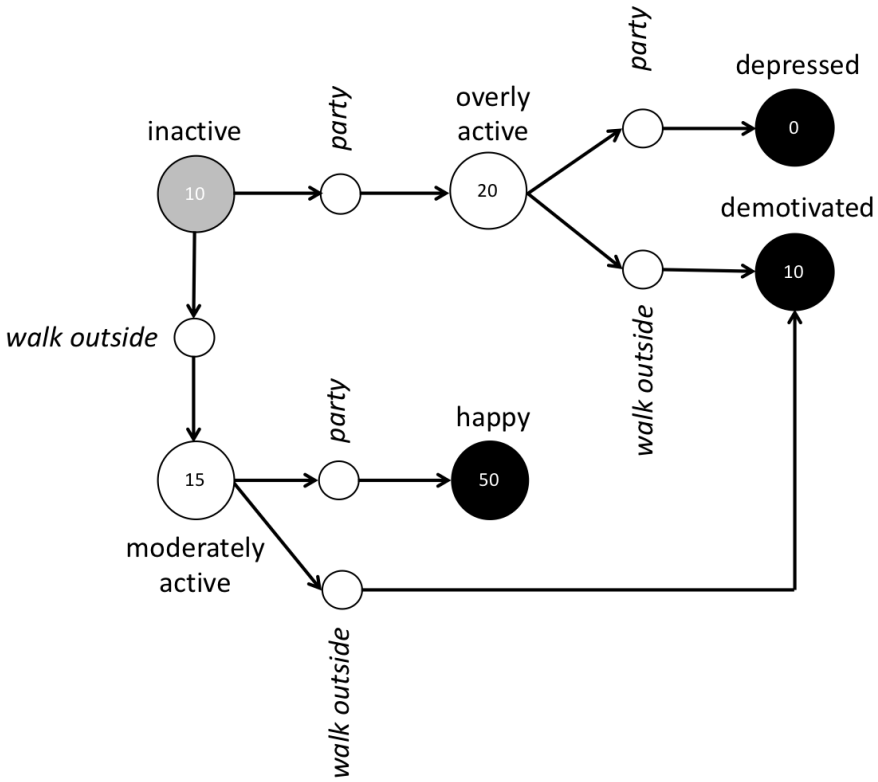

Figure 1: Example MDP for Reinforcement Learning problem

(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, walk outside)* given a value $\gamma = 1$?

(A) 15
(B) 65 → ***correct answer***
(C) 50
(D) None of the other answers

(f) **(1 pt)** What is the purpose of the U-tree algorithm?

(A) To handle a continuous state space. → ***correct answer***
(B) To handle a continuous action space.
(C) To make a reinforcement learning problem satisfy the Markov property.
(D) None of the other answers.

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

This part concerns outlier detection and removal of noise.

Table 1: Example dataset

Time point	*Heart rate*
0	60
1	62
2	65
3	82
4	100
5	102
6	105

(a) (**2 pt**) Consider the measurements for heart rate shown in Table 1. We want to apply the simple distance based approach. Assume we use the absolute distance as a distance metric (e.g. the distance between 100 and 80 is 20). We set $d_{min} = 15$ and $f_{min} = 0.7$. How many points would be considered outliers?

(A) 1 → ***correct answer***
(B) 3
(C) 7
(D) None of the other answers

(b) (**1 pt**) Instead of a distance-based approach we now move to a distribution based approach for the data shown in Table 1. Given the specifics of the data, a mixture model will be used. What would be a natural value for the parameter K for the mixture model given the data, assuming that we use normal distributions in the model?

(A) 1
(B) 2 → ***correct answer***
(C) 3
(D) None of the other answers

(c) (**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 without taking the current values for the observations into account?

(A) $\hat{s}_{t|t-1}$ → ***correct answer***
(B) $s_{t|t-1}$
(C) $\hat{s}_{t|t}$
(D) None of the other answers

(d) (**1 pt**) Under what category does a k-nearest neighbor imputation method fall?

(A) interpolation
(B) model-based → ***correct answer***

(C) mean
(D) median

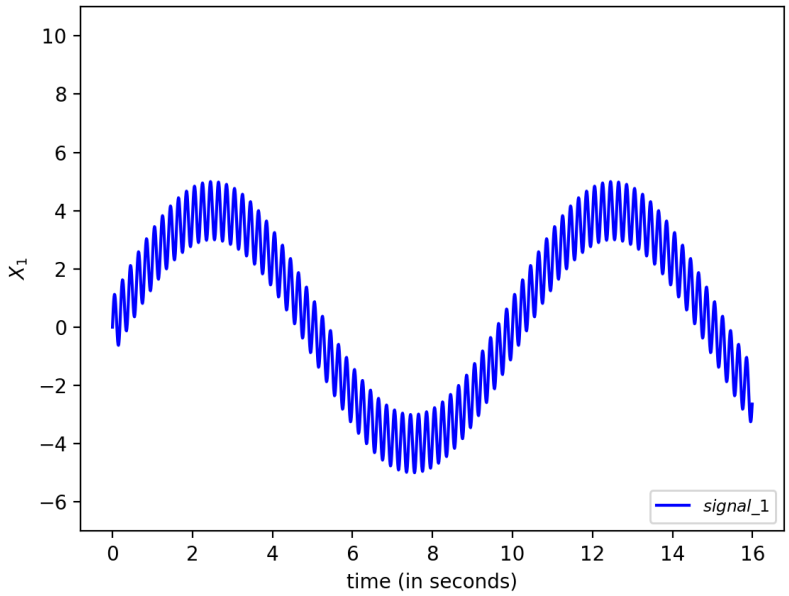

Figure 2: Lowpass filter example

(e) (**1 pt**) We want to get rid of the noise (meaning the high frequency periodic behavior) we see in Figure 2. We apply a lowpass filter. To which value should we set the cut-off frequency to filter that noise out?

(A) 20 Hz
(B) 0.3 Hz → ***correct answer***
(C) 0.05 Hz
(D) None of the above

(f) (**1 pt**) Which approach can be used to reduce the number of features of our dataset?

(A) Lowpass filter
(B) Interpolation
(C) Principal Component Analysis → ***correct answer***
(D) Local Outlier Factor

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

This part concerns feature engineering.

(a) **(1 pt)** Assume we have a dataset of size 100 without any missing values and we apply a window size $\lambda = 10$. We want to derive temporal features over one of our numerical features and use the mean. How many missing values will we have for this temporal feature that we derive?

(A) 1
(B) 9
(C) 10 → ***correct answer***
(D) 11

(b) **(1 pt)** We focus on the algorithm introduced by Batal *et al.* to generate temporal patterns for categorical features. Consider the following statements.

i. A higher value for the window size λ typically results in more patterns compared to a lower value.

ii. A higher value for the threshold Θ typically results in more patterns compared to a lower value.

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.

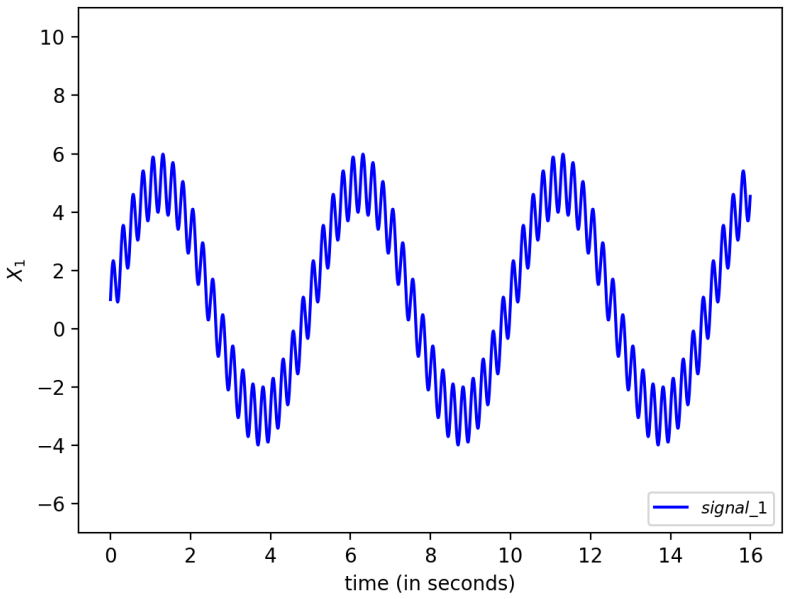

Figure 3: Fourier transformations data example

(c) **(2 pt)** Consider Figure 3. We apply a Fourier transformation to this data with a window which spans the complete data you see in the Figure (i.e. 16 seconds). We see several frequencies for which the amplitude is non-zero. Which frequencies will have a non-zero amplitude?

(A) 4 Hz
(B) 0.2 Hz and 4 Hz

(C) 0 Hz, 0.2 Hz, and 4 Hz → ***correct answer***
(D) None of the other answers

(d) **(1 pt)** We use various NLP approaches to generate features from a lot of texts we have. We consider the following approaches: (A) tokenization followed by a bag-of-words approaches with 1-grams; (B) tokenization, followed by stemming, followed by a bag-of-words approaches with 1-grams; (C) topic modeling. Put these approaches in order of the expected number of features given a representative corpus of texts and settings of parameters according to commonly used values. Start with the smallest number of features and end with the largest.

(A) CAB
(B) ABC
(C) BAC
(D) CBA → ***correct answer***

(e) **(1 pt)** We know that multiple frequencies are relevant in predicting a certain label, even within the windows we are considering per instance (i.e. of size λ). Given this setting, we consider multiple aggregate features for the values that result from the application of a Fourier transformation. Which one would **not** be suitable?

(A) Frequency with the highest amplitude → ***correct answer***
(B) Frequency weighted signal average
(C) Power spectral entropy
(D) All of the other answers are suitable

4. **Clustering (7 pt)**

This part concerns clustering approaches.

(a) **(1 pt)** Which approach should preferably **not** be used when considering person-level clustering?

(A) k-medoids
(B) agglomerative clustering
(C) divisive clustering
(D) k-means → ***correct answer***

(b) **(1 pt)** We want to compare two datasets on a person-level using a temporal approach and know that the time series that are present in the datasets are both shifted and can have a different speed. Which approach would be **most** suitable?

(A) Euclidean distance
(B) Cross Correlation Coefficient
(C) Dynamic Time Warping → ***correct answer***
(D) All the other answers are not suitable

(c) **(2 pt)** Consider two datasets (of different individuals) shown in Table 2. We want to use the Cross Correlation Coefficient (CCC) to compute the correlation between the two time series. Assume we can shift each series with at most $\tau = 2$. What would be the value for the CCC in case we would use the best shift possible to maximize the value?

Table 2: Two datasets

Time point	*Value*
Wenchen	
1	0
2	1
3	1
4	1
5	0
Daniel	
1	1
2	1
3	1
4	0
5	0

(A) 0
(B) 1
(C) 2
(D) None of the other answers → ***correct answer***

(d) **(2 pt)** We now apply Dynamic Time Warping to the same dataset shown in Table 2. What is the value of the shortest path when making the full match, thereby assuming we use the absolute difference as a distance metric between two values?

(A) 0
(B) 1 → ***correct answer***
(C) 2
(D) None of the above

(e) **(1 pt)** Which of the following algorithms can cluster high dimensional data best?

(A) Agglomerative clustering
(B) K-means clustering
(C) Divisive clustering
(D) Subspace clustering → ***correct answer***

5. **Learning Theory and Predictive Modeling with 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)** Consider the following statements about learning theory.

i. Even in case we have an infinite hypothesis set, we can always give certain guarantees on the difference between the in-sample and out-of-sample error because of the theory on VC-dimensions.

ii. Learning theory favors simpler hypothesis sets over more complex ones in terms of the out of sample error, even when there is lots of data.

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***

(b) **(1 pt)** Given a fixed finite hypothesis set of size M, how does the training set size influence the difference between the in-sample and out-of-sample error given the theory of PAC-learnability?

(A) The larger the training set size, the larger the difference between the in-sample and out-of-sample error.
(B) The larger the training set size, the smaller the difference between the in-sample and out-of-sample error. → ***correct answer***
(C) The training set size does not influence the difference between the in-sample and out-of-sample error.
(D) None of the other answers.

(c) **(1 pt)** Which one of the following statements about regularization is *incorrect*?

(A) Regularization avoids overfitting.
(B) Regularization punishes more complex models.
(C) Regularization can be applied to a variety of machine learning approaches.
(D) All of the other statements are correct. → ***correct answer***

(d) **(1 pt)** Which of the following approaches decreases the quality of our assessment on the generalizabilty of our model most?

(A) Do hyper parameter tuning on the test set. → ***correct answer***
(B) Do a cross validation on the training set to tune the hyperparameters.
(C) Not using forward selection.
(D) Not using backward selection.

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)** What does the parameter p in the ARIMA model represent?

(A) The window size for the moving average part.
(B) The window size for the autoregressive part. → ***correct answer***
(C) The degree of differencing.
(D) None of the other answers.

(b) **(1 pt)** Assume we use the Simple GA to find good parameter values for a dynamical systems model we have developed. We have decided not to apply crossover, just mutation. How does the setting for the mutation rate influence the chance of getting stuck in a local optimum?

(A) The lower the probability, the higher the chance of getting stuck in a local optimum. → ***correct answer***

(B) The higher the probability, the higher the chance of getting stuck in a local optimum.

(C) The probability does not influence the chance of getting stuck in a local optimum.

(D) The mutation probability is not part of the Simple GA.

(c) **(2 pt)** Consider the following dynamical systems model:

$$\hat{y}_{hr}(t+1) = y_{hr}(t) + \gamma_1 \times (y_{stress}(t) - \gamma_2) \quad (1)$$

$$\hat{y}_{stress}(t+1) = y_{stress}(t) + \gamma_3 \quad (2)$$

Furthermore, consider the dataset shown in Table 3, which includes measurements around these concepts.

Table 3: Example dataset

Time point	*Heart rate (hr)*	*Stress level (stress)*
0	40	0.2
1	45	0.4

Assume we want to compare two parameter value vectors for the dynamical systems model, namely v_1: $\gamma_1 = 25$, $\gamma_2 = 0$, and $\gamma_3 = 0.2$ and v_2: $\gamma_1 = 20$, $\gamma_2 = 0$, and $\gamma_3 = 0.1$. We compare them on both states (i.e. heart rate and stress level). To do so, we take the measured values at time point 0, feed them into the model instance (so the dynamical systems model with the parameter values), predict the values for time point 1, and compute the error for both states (absolute difference between the predicted and measured value at time point 1). When you compute these for v_1 and v_2, what can we say about the dominance of the two vectors on the Pareto Front?

(A) v_2 is dominated by v_1. → ***correct answer***

(B) v_1 is dominated by v_2.

(C) Both do not dominate each other.

(D) For this problem we cannot compute the dominance of one over the other.

(d) **(1 pt)** Which algorithm suffers most from problems with learning long term dependencies?

(A) Regular Recurrent Neural Networks → ***correct answer***

(B) LSTM networks

(C) Echo State Networks

(D) They all suffer from the problem in a similar fashion