Exam Machine Learning for the Quantified Self

18. 08. 2023
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

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 7 questions in total with 2 points). In total this means 37 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 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).

Good luck!

QUESTIONS

1. **Introduction (7 pt)**

   In daily life, many people experience stress. While some stress does not necessarily have to have a negative impact on people's lives or health, prolonged periods of high stress can have a severe impact, for instance resulting in a burn out. To prevent such burn outs and intervene early on, an app has been developed called *burn out buster*. The app monitors vital signs such as heart rate and respiration and in an initial training phase regularly asks the user to register perceived stress levels (a number between 1 and 10). It then learns to relate the vital sign data to the users perceived stress levels and provides warnings in case of prolonged periods of high stress.

   (a) **(1 pt)** If we were to use the 'Five Factor Framework of Self-Tracking Motivations', which factor would best describes the goal of *burn out buster*?

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

   As mentioned, one of the components of the app is to predict the perceived stress level based on a training set that is collected.

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

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

   (c) **(1 pt)** To perform the task of predicting stress levels, a training set is collected per user. Hereby vital signs are collected (to be specific the heart rate and the respiration) as well as the target (the stress level entered by the user). Which example below is certainly **not** correct?

   (A) $x_1 = [90]$ → ***correct answer***
   (B) $X_1 = [heart\_rate]$
   (C) $\mathbf{X} = \begin{bmatrix} 90 & 20 \\ 80 & 15 \\ 70 & 17 \end{bmatrix}$
   (D) $y_1 = [2.5]$ → ***also accepted as correct answer***

   (d) **(1 pt)** We want to build specific models for each user to predict the stress level. As the time component is important we want to take the order into account. According to the terminology in the book, which learning setup would match this scenario best?

   (A) Individual level temporal. → ***correct answer***
   (B) Individual level non-temporal.

(C) Population level with unknown users.

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

(e) **(1 pt)** Let us move to learning theory. To predict the stress level we use the numerical attributes *heart_rate* and *respiration* and we apply a decision tree learning algorithm. What can we say about the number of hypotheses?

(A) This is infinite. → ***correct answer***

(B) This is finite.

(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.

(f) **(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.

(g) **(1 pt)** Which learning algorithm does **not** by definition have an infinite hypothesis space?

(A) linear regression

(B) multi-layer perceptron

(C) support vector machine

(D) decision tree → ***correct answer***

2. Outlier Detection (8 pt)

This part concerns outlier detection and removal of noise.

Table 1: Example dataset

Time point	*Accelerometer x-axis*
0	10
1	9
2	8
3	9
4	1
5	8
6	0

(a) (**2 pt**) Consider the measurements for the x-axis of an accelerometer as 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 10 and 8 is 2). We set $d_{min} = 7$ and $f_{min} = 0.5$. How many points would be considered outliers?

(A) 1 → ***correct answer***
(B) 2
(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 whereby we take the current values for the observations into account?

(A) $\hat{s}_{t|t-1}$
(B) $s_{t|t-1}$
(C) $\hat{s}_{t|t}$ → ***correct answer***
(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

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

(A) 20 Hz
(B) 0.1 Hz
(C) 0.5 Hz → ***correct answer***
(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***

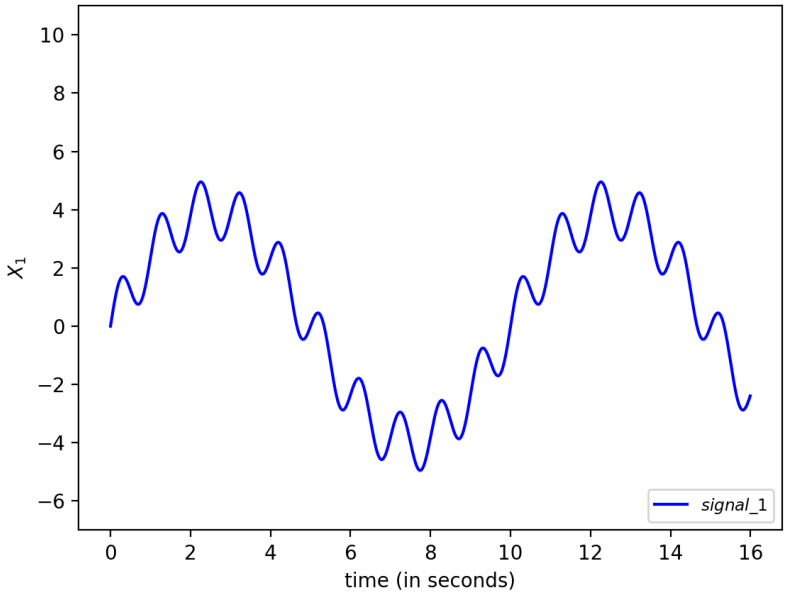

Figure 1: Lowpass filter example

(D) Local Outlier Factor

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

(A) 85 → ***correct answer***
(B) $86\frac{2}{3}$
(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.* decreasing the window size λ will ..... the number of found patterns.

Table 2: Example dataset

Time point	*Heart rate*	*Temporal feature*
0	80	
1	90	
2	100	
3	70	?
4	60	

(A) most likely increase

(B) most likely decrease → ***correct answer***

(C) certainly not influence

(D) certainly increase

(d) (**2 pt**) Consider Figure 1 again. We will now apply a Fourier transformation to the data with $\lambda + 1$=16 seconds. As feature, we use the highest amplitude frequency. What would be the value of this feature?

(A) 0.1 Hz → ***correct answer***

(B) 1 Hz

(C) 0 Hz

(D) None of the other answers

(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 fewer 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 → ***correct answer***

(B) only (i) is correct

(C) only (ii) is correct

(D) both are not correct

4. **Clustering (6 pt)**

This part concerns clustering approaches.

(a) **(1 pt)** We want to compare two datasets on a person-level using a temporal approach and we want to focus on a distance metric that can accommodate a shift in time (and nothing else). Which approach would be **most** suitable?

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

Table 3: Two datasets

Time point	*Value*
Bob	
1	0
2	0
3	1
4	1
5	0
Mark	
1	0
2	1
3	1
4	0
5	0

(b) **(2 pt)** Consider two datasets (of different individuals) shown in Table 3. We want to use the Cross Correlation Coefficient (CCC) to compute the correlation between the two time series. Assume we can shift one series with at most $\tau = 1$. What would be the value for the CCC in case we would use the best shift possible to maximize the value?

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

(c) **(2 pt)** We now apply Dynamic Time Warping to the same dataset shown in Table 3. 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 → ***correct answer***
(B) 1
(C) 2

(D) None of the other answers

(d) **(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. **Predictive Modeling with the Notion of Time (3 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 → ***correct answer***
(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 are randomly set and not trained in this network?

(A) $l \times n$
(B) $l \times n + n \times n$
(C) $n \times n + n \times p \rightarrow$ ***correct answer***
(D) $n \times n$

(c) **(1 pt)** In TCN's we have so-called dilations, indicated by the factor d. Assume we specify the dilation factor per layer as $d = 2^{m*i}$ with i the layer number and m a hyperparameter. Complete the following sentence: the lower we set the value of m the .....

(A) less deep we need to make the network to capture the complete history.
(B) deeper we need to make the network to capture the complete history. → ***correct answer***
(C) more weights we will have to learn in the network.
(D) None of the other answers.

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

We are now going to focus on Reinforcement Learning.

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

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha \cdot (R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Which algorithm does this equation belong to?

(A) Q-learning with eligibility traces
(B) SARSA with eligibility traces
(C) Q-learning without eligibility traces
(D) SARSA without eligibility traces → ***correct answer***

(b) **(1 pt)** Complete the following sentence. Q-learning is ...

(A) an on-policy Reinforcement Learning algorithm.
(B) an off-policy Reinforcement Learning algorithm. → ***correct answer***
(C) the same as SARSA.
(D) not a Reinforcement Learning algorithm.

(c) **(2 pt)** Consider the MDP shown in Figure 2 which focuses on reducing high stress levels by means of two actions. What is the value that is eventually learned for *Q(highly_stressed, go_for_a_walk)* given a value $\gamma = 1$?

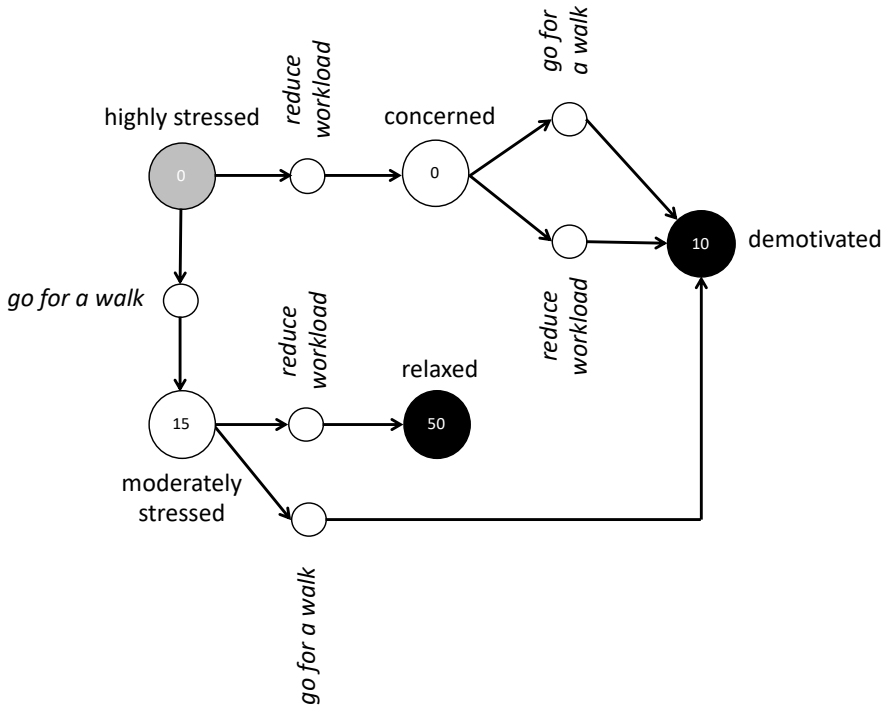

Figure 2: MDP for stress reduction

(A) 15
(B) 25
(C) 65 → ***correct answer***
(D) None of the other answers

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

(A) focus on long term rewards only.

(B) focus only on instant rewards.
(C) weigh all rewards in the future equally. → ***correct answer***
(D) end up with a SARSA algorithm.