Exam Machine Learning for the Quantified Self
01. 07. 2022
15:30 - 18:15

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 8 questions in total with 2 points). In total this means 38 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 (6 pt)**

   Several quantified self enthusiasts have decided to develop a game-like app that aims to get people to record as much information about themselves as they can. The app is called *AllQuantified*. The more different measurements you collect, the more points you gather in the game. Rewards are given both for the different aspects that are being measured (e.g. heart rate, mood, steps) as well as how frequently they are measured. In the end, the game developers want to use this wealth of data to see what kind of measurements can be useful for which kind of situations.

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

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

   (b) **(1 pt)** Due to the diversity of measurements and how frequently they are captured the developers are confronted with a challenging problem, namely that a choice has to be made on the level of granularity (i.e. Δt) with which they process their datasets to further analyze the usefulness of the measurements. Consider the following statements.

   i. A higher value for Δt is more likely to give us a higher number of missing values in the dataset.
   ii. A higher 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 the value of the k^{th} feature of the n^{th} instance in a dataset of one person resulting from *AllQuantified*. Using what mathematical term do we identify this feature?

   (A) x_k^n
   (B) $\mathbf{X_n^k}$
   (C) x_n^k
   (D) None of the the other answers

   (d) **(1 pt)** The developers aim to predict the activity type (e.g walking, running) based on all the measurements that have been collected by the users. What kind of machine learning task is this?

(A) Reinforcement Learning
(B) regression
(C) classification
(D) clustering

(e) **(1 pt)** As a setup to create a predictive model for the activity type, the developers decide to exploit the order in which measurements are done, and they combine all datasets of the quantified selves that participate in the app. They test on data of users that are not in the training set. 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
(D) Population level with unseen data of known users.

(f) **(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.
(B) Do a cross validation on the training set to tune the hyperparameters.
(C) Not using forward selection.
(D) Using a very complex model.

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

This part concerns outlier detection and removal of noise.

(a) (**1 pt**) Consider Figure 1. Let us focus on the red plus and only $Feature_1$. We use a mixture model. With what value for K for the mixture model would it be very likely that this point would be **not** considered an outlier?

(A) $K = 1$
(B) $K = 2$
(C) $K = 3$
(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 1 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 point **not** being an outlier?

(A) $d_{min} = 10$ and $f_{min} = 0$
(B) $d_{min} = 10$ and $f_{min} = 1$
(C) $d_{min} = 1$ and $f_{min} = 0.5$
(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 when taking only information of the previous time point into account?

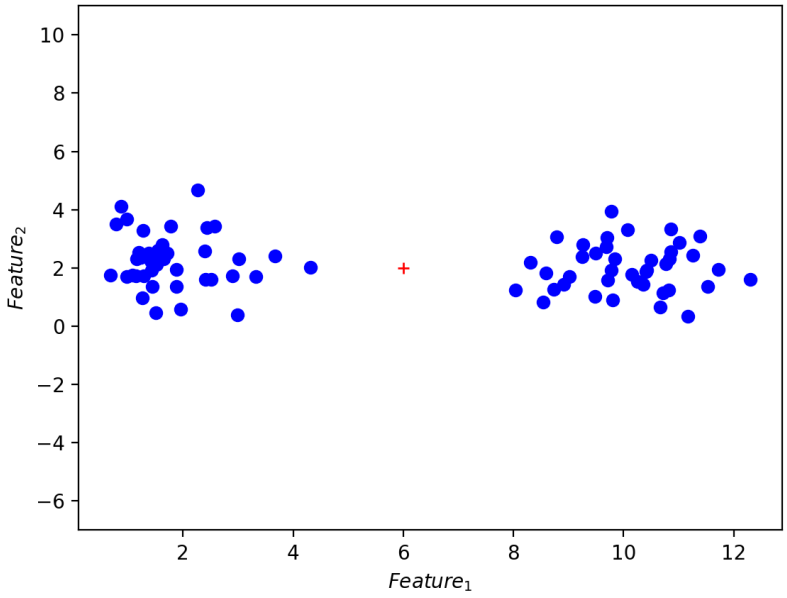

Figure 1: Example outliers

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

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

(e) (**2 pt**) Consider Figure 2. We see a signal with two different frequencies combined here, which two frequencies are these?

(A) 0.5 Hz and 0.025 Hz
(B) 0.25 Hz and 0.05 Hz
(C) 0.25 Hz and 0.025 Hz
(D) None of the other answers

(f) (**1 pt**) Which approach could be used to remove the highest frequency shown in Figure 2?

(A) Interpolation

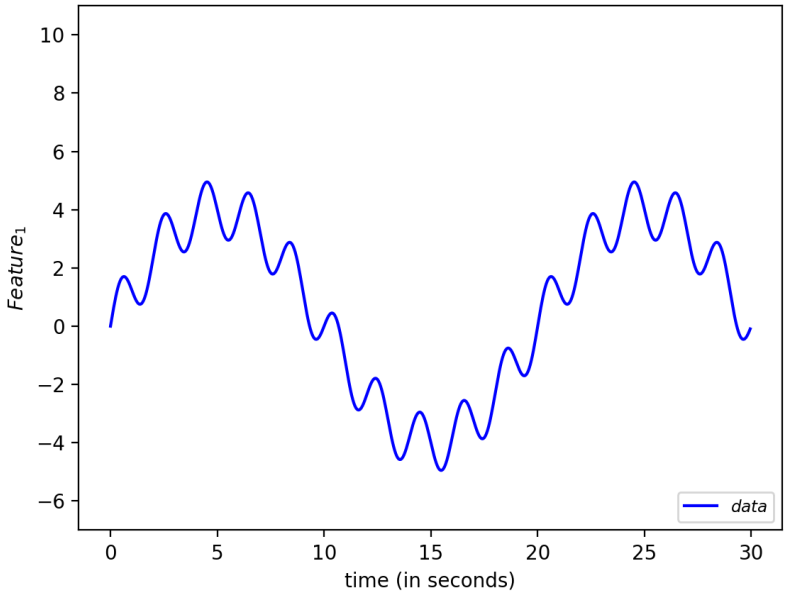

Figure 2: Example time series

(B) Principal Component Analysis
(C) Lowpass filter
(D) Local Outlier Factor

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

This part concerns feature engineering.

(a) (**2 pt**) Consider Figure 2 again. We will now apply a Fourier transformation to the data with $\lambda + 1$=30 seconds. As feature, we use the highest amplitude frequency. What would be the value of this feature?

(A) 0.05 Hz
(B) 0.5 Hz
(C) 0.25 Hz
(D) none of the other answers

(b) **(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.

Table 1: Example dataset

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

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

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) $83\frac{1}{3}$
(B) $82\frac{1}{2}$
(C) 90
(D) none of the other answers

(d) (**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.

(A) certainly not influence
(B) certainly increase
(C) most likely increase
(D) most likely decrease

4. **Clustering (7 pt)**

This part concerns clustering approaches.

(a) **(2 pt)** Look at Figure 3, where measurements of two quantified selves Eric and Mark for $Feature_1$ are shown with the measured values indicated by diamonds for Eric and crosses for Mark. We want to compute the distance using the cross correlation coefficient with a maximum leg of $\tau = 1$. What is the value that results from this computation over the entire time series?

(A) 11
(B) $\sqrt{8}$
(C) 0
(D) None of the other answers

(b) **(2 pt)** Rather than using the cross correlation coefficient 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 3 is 2). What distance should be filled in in Table 2 at the position of the "**?**"?

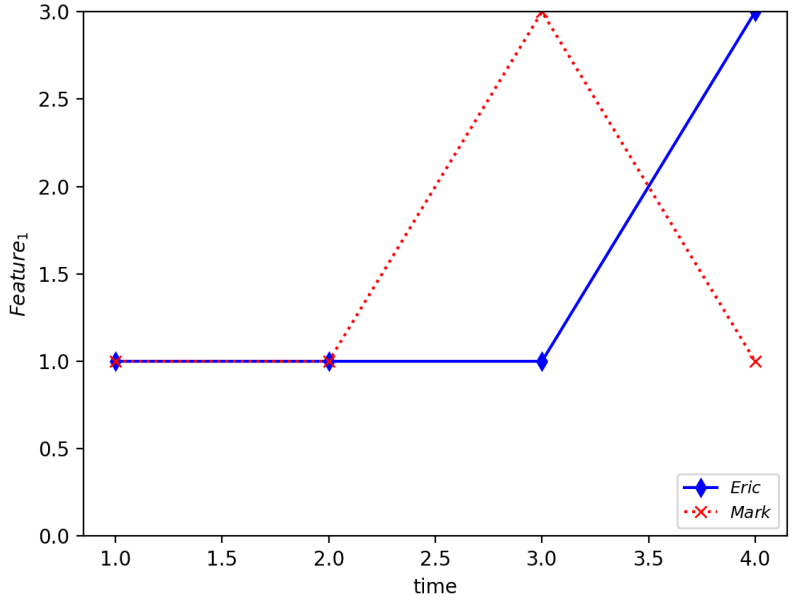

Figure 3: Time series

Table 2: DTW answer table

Eric	t=4				
	t=3				
	t=2		?		
	t=1				
		t=1	t=2	t=3	t=4
		Mark			

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

(c) **(2 pt)** What is the value of the shortest path (i.e the value in the upper right of Table 2) when making the full match with Dynamic Time Warping?

(A) 0
(B) 2
(C) 4
(D) None of the other answers

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

i. A higher value for ϵ will increase the number of dense units.

ii. Units only have a common face when the boundaries for all but one feature are the same, and for the feature where they are not the upper bound of one equals the lower bound of the other.

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 with the Notion of Time (6 pt)**

Below, we will focus on questions related to learning theory as well as predictive modeling with the notion of time.

(a) **(1 pt)** What does a hypothesis set with an infinite VC dimension mean?

(A) That we cannot provide guarantees on the different between the in sample and out of sample error according to the learning theory we have discussed during the lecture
(B) That the hypothesis set cannot shatter all possible restriction of our learning problem.
(C) That the hypothesis set is PAC learnable.
(D) None of the other answers.

(b) (**1 pt**) Fill in the right words in the following sentence: according to PAC learnability, if we have more data, then the difference between the in sample and out of sample error will .....

(A) become smaller
(B) become bigger
(C) stay the same
(D) only change in case the complexity of the model goes down

(c) **(1 pt)** Which parameter in the ARIMA model represents the order of differencing?

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

(d) (**1 pt**) Consider the following statements about Echo State Networks:

i. Only the last layer of weights is trained.
ii. The echo state property is satisfied when activation of a neuron does not vanish over time.

Which of these statements is correct?

(A) both are correct.
(B) only (i) is correct.

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

(e) **(1 pt)** Which algorithm uses a temperature parameter?
(A) NSGA-II
(B) Simple GA
(C) Simulated Annealing
(D) None of the other answers

(f) **(1 pt)** Consider that we apply the Simple GA and we get two parents: $[1, 0, 0, 1]$ and $[1, 1, 0, 1]$. What could be a child of these two parents after crossover?
(A) $[1, 1, 0, 1]$
(B) $[1, 0.5, 0, 1]$
(C) $[1, 0, 1, 0]$
(D) $[1, 1, 1, 1]$

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

We are now going to focus on Reinforcement Learning.

(a) **(1 pt)** Complete the following sentence. Q-learning is ...
(A) an on-policy Reinforcement Learning algorithm
(B) an off-policy Reinforcement Learning algorithm
(C) the same as SARSA
(D) not a Reinforcement Learning algorithm

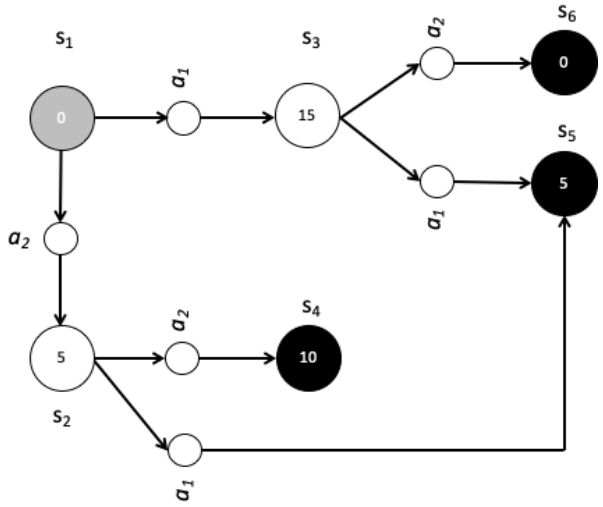

Figure 4: Example MDP for Reinforcement Learning problem

(b) **(2 pt)** Consider the MDP shown in Figure 4. What is the value that is eventually learned for $Q(s_1, a_1)$ given a value $\gamma = 1$?

(A) 5
(B) 15
(C) 20
(D) None of the other answers

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

(A) weigh all future rewards equally.
(B) focus only on instant rewards.
(C) will not be able to run a Reinforcement Learning algorithm.
(D) end up with a SARSA algorithm.

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

(A) To handle a continuous state space.
(B) To handle a continuous action space.
(C) To make a Reinforcement Learning problem satisfy the Markov property.
(D) None of the other answers.