Exam Machine Learning for the Quantified Self
29. 06. 2023
18:45 - 21:30

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 and Reinforcement Learning (7 pt)**

   (a) **(1 pt)** During the lecture, we have discussed various definitions of the quantified self. We have adopted one definition, which captures the concept best within the scope of our course. Fill in the blanks in this definition: *The quantified self is any individual engaged in the .... of any kind of biological, physical, behavioral, or environmental information. The .... is driven by .... of the individual with a desire to act upon the collected information.*

   (A) collection (2x), a certain goal
   (B) self-tracking (2x), a certain goal → ***correct answer***
   (C) collection (2x), a certain personality
   (D) self-tracking (2x), a certain personality

   (b) **(1 pt)** A diabetic is using an app in combination with a digital glucose meter to better manage the glucose level. In this app, information on diet can also be recorded. The app provides up-to-date insights into glucose levels and also analyzes the influence of certain food and drink products (provided by the user) on glucose levels. If we were to take the "Five-Factor-Framework of Self-Tracking-Motivation", which one would describe the purpose of the diabetic to use the app best?

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

   (c) **(1 pt)** When we start to read and process a dataset of sensory data, a choice has to be made on the level of granularity (i.e. Δt). Consider the following statements.

   i. A higher value for Δt is more likely to give us a lower number of missing values in the dataset.
   ii. A higher value for Δt gives us less instances.

   Which of these statements is correct?

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

   (d) **(1 pt)** We want to identify the i^{th} feature of a dataset. Using what mathematical term do we identify this feature?

   (A) $\mathbf{X_i}$
   (B) x_i
   (C) X_i → ***correct answer***
   (D) None of the the other answers.

(e) **(1 pt)** In Reinforcement Learning, the Markov property is an essential assumption for many algorithms. Let us focus on sports games where we assume the state consists of a single observation (e.g. a single board for chess, a single video frame of a car race). Which one of the following games satisfies the Markov property?

(A) Tennis

(B) Car racing

(C) Chess → ***correct answer***

(D) None of the other answers.

(f) **(1 pt)** Fill in the missing part of this equation for Q-learning:
$Q(S_t, A_t) = Q(S_t, A_t) + \alpha(R_{t+1} + \gamma ...... - Q(S_t, A_t))$

(A) $\max_{A_{t+1} \in \mathcal{A}} Q(S_{t+1}, A_{t+1})$ → ***correct answer***

(B) $Q(S_{t+1}, A_{t+1})$

(C) $Q(S_t, A_t)$

(D) None of the other answers.

(g) **(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.

2. **Outlier Detection (5 pt)**

This part contains the removal of sensory noise.

(a) (**1 pt**) When we consider the number of instances N in our dataset, how does that impact Chauvenet's criterion?

(A) The higher N the less strict Chauvenet's criterion is. → ***correct answer***

(B) The higher N the more strict Chauvenet's criterion is.

(C) N does not impact the strictness of Chauvenet's criterion.

(D) None of the other answers.

(b) (**1 pt**) In the lowpass filter, $|G(f)|^2$ indicates the magnitude of the filter. The order of the filer n influences this magnitude. Complete the following sentence. The higher the order of the filter ......

(A) The less severe will frequencies just above the cut-off frequency be filtered out, and the higher the magnitude of the filter for those frequencies.

(B) The more severe will frequencies just above the cut-off frequency be filtered out, and the higher the magnitude of the filter for those frequencies.

(C) The less severe will frequencies just above the cut-off frequency be filtered out, and the lower the magnitude of the filter for those frequencies.

(D) The more severe will frequencies just above the cut-off frequency be filtered out, and the lower the magnitude of the filter for those frequencies. → ***correct answer***

(c) (**2 pt**) Consider Table 1 where we consider two features X_1 and X_2. Let us consider the Local Outlier Factor algorithm and assume we set $k = 2$. As a distance metric we use the Manhattan distance. What would the *local reachability distance* of the second datapoint (i.e. with $X_1 = 2$ and $X_2 = 2$) be?

Table 1: Example dataset

Data point	X_1	X_2
1	1	1
2	2	2
3	2	3
4	3	4

(A) $k_{lrd} = \frac{2}{3}$
(B) $k_{lrd} = 0.64$
(C) $k_{lrd} = \frac{2}{5} \rightarrow$ ***correct answer***
(D) None of the other answers.

(d) (**1 pt**) We apply the Kalman filter to find outliers and impute missing values. We notice that our prediction of the observed state compared to the actual observation seems to be off. In which part of the filter will this error **not** be reflected?

(A) $\hat{e}_t$
(B) $\hat{P}_{t+1|t}$
(C) $\hat{s}_{t|t}$
(D) $\hat{s}_{t|t-1} \rightarrow$ ***correct answer***

3. **Feature Engineering (9 pt)**

This part concerns feature engineering.

Table 2: Example dataset

Time point	*Heart_rate*	*Activity_level*	*Tired*
0	60	Low	No
1	80	High	No
2	90	High	Yes
3	80	Low	Yes
4	60	Low	No

(a) **(2 pt)** Consider the dataset shown in Table 2. Let us focus on the algorithm of Batal *et al.* as explained during the course. Assume we apply a window size $\lambda = 1$. A minimum threshold for the support of $\Theta = 0.5$ and generate patterns with a maximum length of 2 and we use *Activity_level* and *Tired* as features. How many patterns of size 2 result? Note that we count co-occurs patterns independent of the order.

(A) 6

(B) 2

(C) 0

(D) None of the other answers. → ***correct answer***

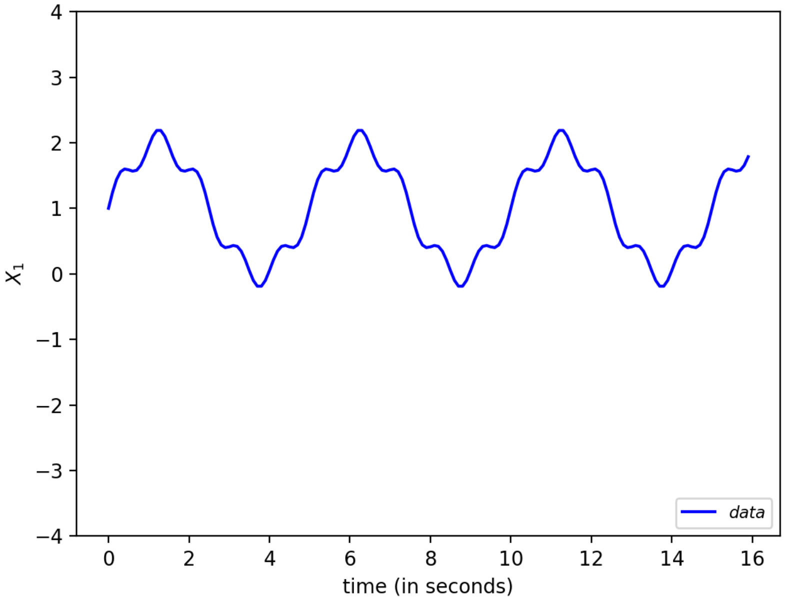

Figure 1: Time series

(b) (**2 pt**) Consider Figure 1. We will now apply a Fourier transformation to the data with $\lambda + 1 = 15$ seconds. How many frequencies would you expect to have a non-zero amplitude?

(A) 1

(B) 2

(C) 3 → ***correct answer***

(D) None of the other answers.

(c) **(2 pt)** Consider the dataset shown in Table 2 again. We focus on the time domain now and intend to summarize the *Heart_rate* feature using a window size λ by means of the maximum value observed in the window to predict whether the value for the feature *Tired* is *yes* or *no*. What would be the window size that results in values for the temporal feature of *Heart_rate* that can be directly used to predict the target using a simple decision tree without any errors?

(A) $\lambda = 0$

(B) $\lambda = 1$ → ***correct answer***

(C) $\lambda = 2$

(D) None of the other answers.

(d) **(2 pt)** We want to apply a bag of words approach with unigrams where we assign a TF-IDF score as value for each feature. Imaging that we have the following two instance of text:

- *I really like sensory data*
- *I really like the nice weather*

What would be the value of the feature "like" for the first instance?

(A) 0 → ***correct answer***
(B) 1
(C) $\frac{1}{2}$
(D) None of the other answers.

(e) **(1 pt)** In topic modeling, how are topics represented in the algorithm?

(A) By means of a single keyword that describes the topic best.
(B) Through the assignment of weights to each word occurring in the corpus of text. → ***correct answer***
(C) Using the TF-IDF score.
(D) None of the other answers.

4. **Clustering (7 pt)**

This part concerns clustering approaches.

(a) (**1 pt**) Consider the following statements about the learning setup for clustering and the associated distance metrics:

i. The cross correlation coefficient is only used for person-level clustering.
ii. The Euclidean distance is never relevant for person-level clustering.

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.

Table 3: DTW dataset

Time point	*Arnold*	*Bruce*
1	1	1
2	1	5
3	1	1
4	5	1
5	1	1

(b) **(2 pt)** Look at Table 3, where we see measurements of two quantified selves, namely Arnold and Bruce. We intend to use Dynamic Time Warping to compute the distance. To compute this, you need to create a table as shown in Table 4. Assume we use the Manhattan difference for comparing two values (e.g. distance between 1 and 3 is 2). What would be the distance between the two time series when applying Dynamic Time Warping?

Table 4: 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				

(A) 0 → ***correct answer***

(B) 4

(C) 8

(D) None of the other answers.

(c) **(1 pt)** We now decide to limit the maximum difference in time points for matches in Dynamic Time Warping to 1, meaning that for instance $t = 2$ for Arnold can be matched with $t = 1$, $t = 2$, and $t = 3$ of Bruce, but not with $t = 4$. How many cells in our table (of the 25) are still relevant?

(A) 25

(B) 13 → ***correct answer***

(C) 19

(D) None of the other answers.

(d) **(2 pt)** Provided the constraint on the maximum difference of 1 between time points in Dynamic Time Warping and following the same Manhattan distance to compute the distance between time points, what would be the new distance between the two time series when applying Dynamic Time Warping?

(A) 0

(B) 4

(C) 8 → ***correct answer***

(D) None of the other answers.

(e) **(1 pt)** In the Silhouette score to evaluate clustering, for an instance x_i which value do you want to be as low as possible?

(A) $a(x_i)$ → ***correct answer***

(B) $b(x_i)$

(C) $c(x_i)$

(D) None of the other answers.

5. **Learning Theory and Predictive Modeling without 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)** During the lecture, we have discussed Figure 2. What does ϵ provide guarantees on in PAC learnability?

(A) The difference between $E_{out}(h)$ and $E_{in}(h)$. → ***correct answer***

(B) The height of $E_{in}(h)$.

(C) The height of $E_{out}(h)$.

(D) None of the other answers.

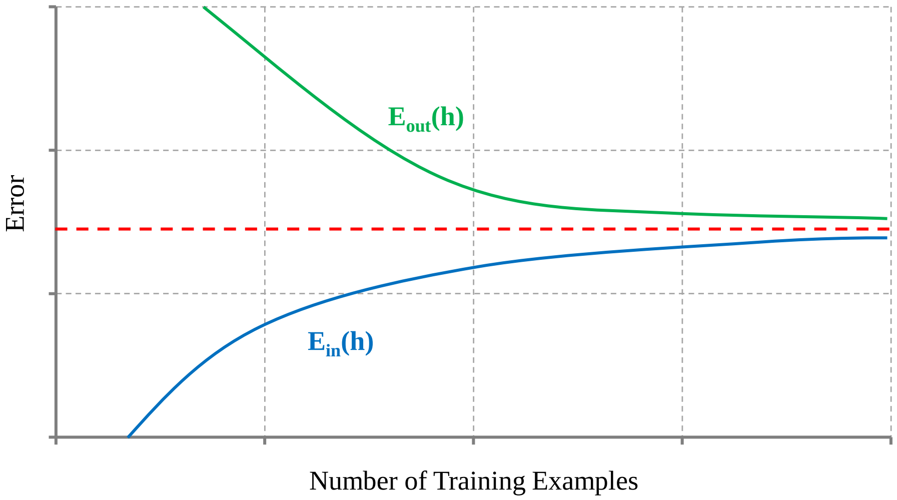

Figure 2: Error as Function of N

(b) **(1 pt)** We consider a simple perceptron (i.e. a model that is able to create a hyperplane using a linear combination of feature values). What is the VC dimension of such a perceptron?

(A) Infinite

(B) 3 → ***correct answer***

(C) 4

(D) None of the other answers.

(c) **(2 pt)** We have 3000 measurements (i.e. instances) in our dataset and want to create temporal features using a window size of $\lambda + 1 = 20$. To avoid instances from being too similar, we limit the amount of overlap allowed to 50%. How many datapoint will we have left after we exclude all instances with too much overlap?

(A) 1500

(B) 300

(C) 298

(D) None of the other answers. → ***correct answer***

(d) (**1 pt**) We have collected a dataset on a single human performing different activities during 2 hours. We decide to split the data in a training and test set based on random sampling and allow for a maximum overlap between instances (we use temporal features) of 90%. We get scores of close to 100% accuracy on the test set. Consider the following statements:

i. Due to the substantial overlap between instances, the training set and test set might contain very similar instances, resulting in a potential overestimation of generalizability.

ii. The dataset has been collected during a very brief period, making the data more uniform compared to data collected over a longer period.

Which of these statements is correct?

(A) both are correct. → ***correct answer***

(B) only (i) is correct.

(C) only (ii) is correct.

(D) both are incorrect.

(e) **(1 pt)** Some machine learning approaches have an explicit regularization parameter, while for others this is left more implicit. For a decision tree, which hyperparameter would be most obvious to play the role of a regularization parameter?

(A) The split criterion.

(B) The maximum tree depth. → ***correct answer***

(C) The type of decision tree algorithm.

(D) None of the other answers.

6. **Predictive Modeling with the Notion of Time (4 pt)**

We are now going to focus on predictive modeling with the notion of time.

(a) **(1 pt)** Order the following networks in long-term memory performance (worst to best): LSTM, RNN, and TCN:

(A) RNN, LSTM, TCN. → ***correct answer***

(B) RNN, TCN, LSTM.

(C) TCN, LSTM, RNN.

(D) None of the other answers.

(b) **(1 pt)** What does the echo state property guarantee?

(A) That activations of neurons in the reservoir of an echo state network gradually die out. → ***correct answer***

(B) That a reservoir of an echo state network can store some information.

(C) That a reservoir of an echo state network has enough neurons compared to the number of data points.

(D) None of the other answers.

(c) **(1 pt)** Which one of the following equations represents the forget gate in an LSTM?

(A) $h_t = u_t \cdot tanh(C_t)$

(B) $C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$

(C) $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ → ***correct answer***

(D) None of the other answers.

(d) **(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 higher we set the value of m the .....

(A) deeper we need to make the network to capture the complete history.

(B) less deep 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.