Exam Machine Learning for the Quantified Self
22. 07. 2022
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 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).

Good luck!

QUESTIONS

1. **Introduction (7 pt)**

   Patients that have been diagnosed with cancer often suffer from mental health problems as a result of their disease. To support these patients best, an app has been developed that measures both the physical and mental wellbeing of these patients, and also allows tracking the cancer treatments the patients receive. Physical measurements include the heart rate, body temperature, and respiration rate. The mental wellbeing is measured using regular questions posed to the patients in which they are asked to rate their mood. The treatment schedule is extracted from the hospital records. In the end the goal of the app is to provide coaching to patients based on their observed state, provide mental health treatments, and optimize the cancer treatment schedule (within the bounds that are possible) so as to maximize not only the treatment effectiveness, but also the mental health of the patient.

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

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

   (b) **(1 pt)** The measurements take place at very different granularities. Therefore deciding on the right Δt for processing the dataset is challenging. 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 less missing values.

   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)** Given the data we have collected in the app, we want to predict the mood (a continuous value) based on the physical measurements and the treatment schedule. Which notation do we use for the set of targets according to the notation put forward in the book?

   (A) $\mathbf{X}$
   (B) $\mathbf{Y}$
   (C) $\mathbf{G}$
   (D) None of the the other answers

(d) **(1 pt)** Which learning algorithm would **not** be suitable for the task to predict the continuously valued mood?

(A) linear regression
(B) multi-layer perceptron
(C) decision tree
(D) naive bayes

(e) **(1 pt)** We would like to use Reinforcement Learning (RL) to learn when to best send messages to patients to coach them, and also what the precise content of the messages should be. We want to derive this based on the measurements of the patient at that moment. What is the term used in RL for the messages we intend to send?

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

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

(A) SARSA always uses eligibility traces.
(B) SARSA assumes future actions are selected in the same way as current actions while Q-learning does not.
(C) Q-learning is able to handle continuous values while SARSA is not.
(D) There is no difference between SARSA and Q-learning.

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

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

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

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 lower we set the value of c in Chauvenet's criterion, the fewer points we will identify as outlier.
ii. The lower the value for k in the local outlier factor 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.

Consider the data shown in Figure 1.

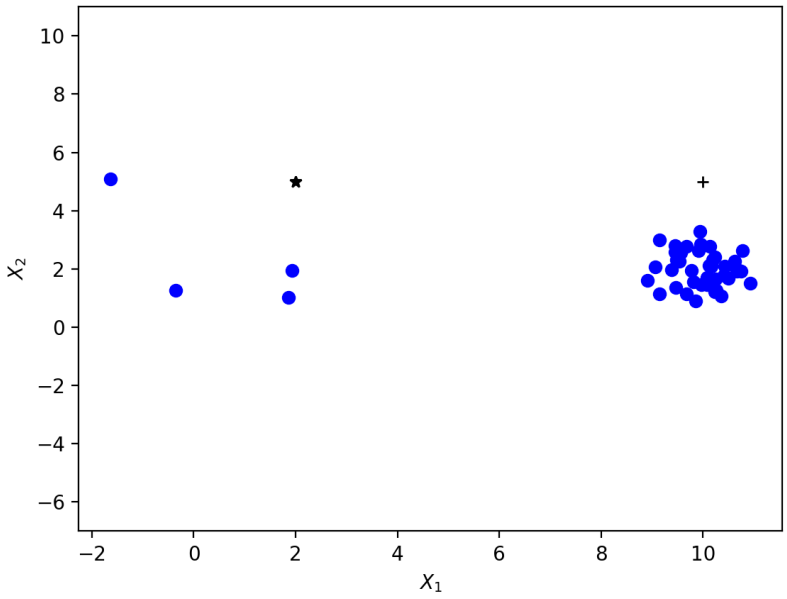

Figure 1: Example dataset - outlier

(b) (**2 pt**) Of the points represented by the star and plus in Figure 1, which would be likely to be considered outliers when we would apply the local outlier factor algorithm to the entire dataset shown?

(A) the point represented by the plus
(B) the point represented by the star
(C) the points represented by the plus and the star
(D) none of the points

(c) (**1 pt**) Which component in the Kalman filter determines how the latent state relates to an observation?

(A) $P_{t|t}$
(B) H_t
(C) B_t
(D) F_t

Consider the data shown in Figure 2.

(d) (**1 pt**) We want to apply a lowpass filter to Figure 2 and decide to select a cut-off frequency $f_c = 0.5Hz$. What would the data after application of the filter look like?

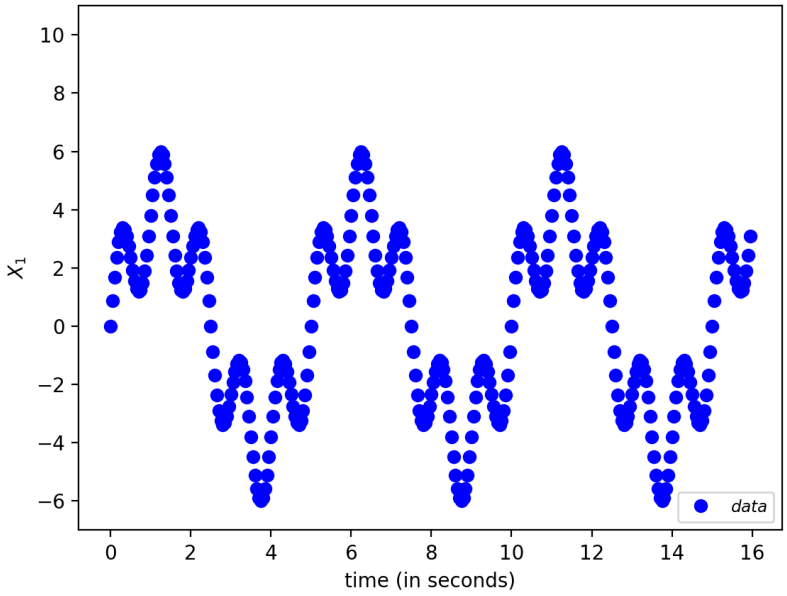

Figure 2: Example periodic dataset - outlier

(A) The signal that has a periodicity of approximately $1Hz$ would disappear.
(B) The signal that has a periodicity of approximately $0.2Hz$ would disappear.
(C) The line of the data would become flat (i.e. all periodic data is removed).
(D) The data would remain unchanged.

(e) (**1 pt**) When we consider the data shown in Figure 2 and we would like to impute missing values, which approach would be appropriate?

(A) Median imputation
(B) Mean imputation
(C) Kalman filter
(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 $\lambda = 1$ for a particular feature X_1, which approach would result in the most added features in our dataset?

(A) Time domain using the mean, median, and standard deviation
(B) Time domain using the mean
(C) Frequency domain using the amplitude of each frequency considered
(D) Frequency domain using the power spectral entropy

Table 1: Example dataset

Time point	*Activity_level*	*Mood*
0	Low	Good
1	Low	Good
2	Low	Bad
3	Low	Bad
4	Low	Bad

(b) **(2 pt)** Consider the dataset shown in Table 1. Let us focus on the algorithm of Batal *et al.* as explained during the course. Assume we apply a window size $\lambda = 1$, what is the support for the pattern $Mood = Good$?

(A) 2/5
(B) 2/4
(C) 3/5
(D) None of the other answers

(c) **(2 pt)** We continue with the algorithm of Batal *et al.*. Assume we apply a window size $\lambda = 1$ and select a minimum threshold for the support of $\Theta = 0.6$ and generate patterns with a maximum length of 3. How many patterns result? Note that we count co-occurs patterns independent of the order (e.g. $Mood = Good$ (c) $Activity\_level = Low$ and $Activity\_level = Low$ (c) $Mood = Good$ is the same pattern and counts only once).

(A) 2
(B) 4
(C) 5
(D) None of the other answers

(d) (**2 pt**) Consider the data shown in Figure 3.
We apply a Fourier transformation with a window size $\lambda + 1$ of 16 seconds. Which frequency would have the highest amplitude?

(A) 0.1 Hz
(B) 2 Hz
(C) both 0.1 Hz and 2 Hz
(D) None of the other answers

(e) (**1 pt**) We now consider text based data. More specifically, we consider the NLP pipeline that has been discussed during the lecture. We start with tokenization, and end with stopword removal. Which two steps are done in between and in which order?

(A) Lower case and then stemming
(B) Stemming and then lower case
(C) Sentence splitting and then stemming
(D) None of the other answers

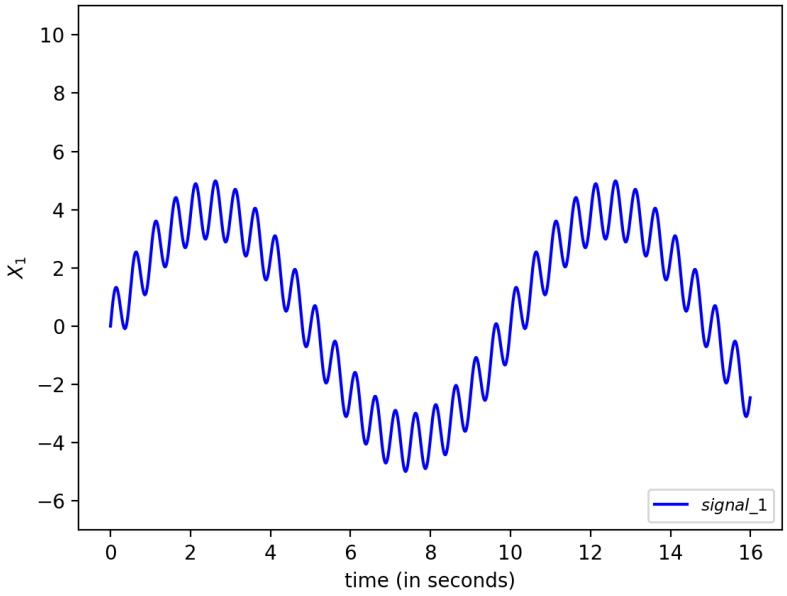

Figure 3: Example dataset - temporal

4. **Clustering (5 pt)**

This part concerns clustering approaches.

(a) (**1 pt**) When we apply clustering on an instance 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
(B) 100
(C) 1000
(D) None of the other answers

(b) **(2 pt)** Look at Figure 4, 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 Euclidean distance on raw data. What is the value that results from this computation over the entire time series?

(A) 4
(B) $\sqrt{8}$
(C) $\sqrt{20}$
(D) None of the other answers

(c) **(1 pt)** In Dynamic Time Warping we have several so-called conditions. Which condition states that the first and last time points of two series should be matched?

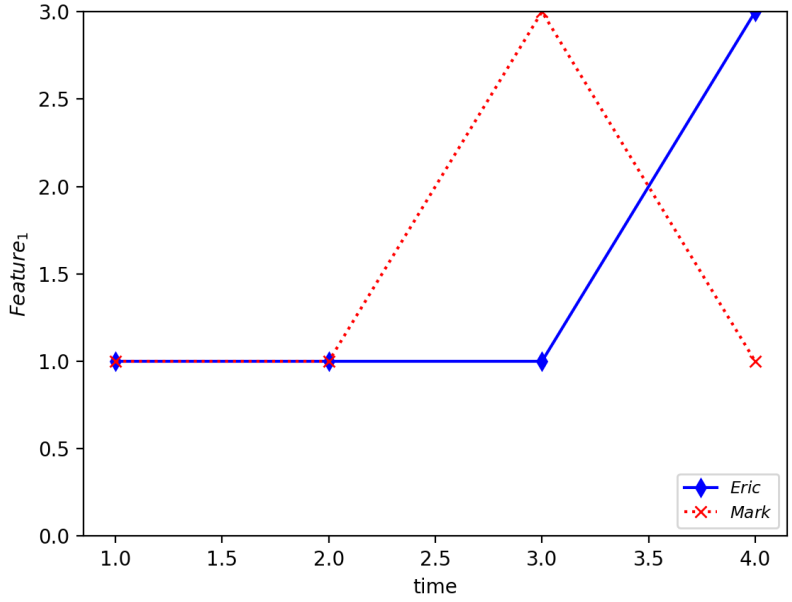

Figure 4: Time series

(A) The boundary condition
(B) The initial condition
(C) The monotonicity condition
(D) None of the other answers

(d) (**1 pt**) When are two units that use k attributes considered to be *connected* in subspace clustering?

(A) When the units share the same upper and lower limits for $k-1$ attributes and for the other attribute k the lower bound of one unit equals the upper bound of the other unit.
(B) When the units have a common face or when they share a unit they both have a common face with.
(C) When the specifications of the two units fully overlap (i.e. the upper and lower bound for all k attributes are identical).
(D) None of the other answers.

5. **Learning Theory and Predictive Modeling without 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)** When we consider PAC learnability, what does this theory say when we consider one datasets with N instances and two finite sets of hypotheses M_1 and M_2 where $|M_1| > |M_2|$ (i.e. M_1 contains more hypotheses than M_2)?

(A) The difference between the in sample error and out of sample error is smaller in the case of M_1.

(B) The difference between the in sample error and out of sample error is smaller in the case of M_2.

(C) The VC dimension of M_1 is smaller.

(D) The VC dimension of M_2 is smaller.

(b) **(1 pt)** Assume we have a dataset with two instances and we have a binary classification problem. What is the number of elements in the restriction H_X on a hypothesis set H?

(A) 2

(B) 4

(C) 8

(D) Restrictions do not apply to binary classification problems.

(c) **(1 pt)** We are applying a machine learning algorithm which does not consider the temporal dimension explicitly. Hereby, we use temporal features. To avoid too much overlap between instances we decide to set a limit to the amount of overlap allowed. We try out two different values for the overlap we allow: 50% and 90%. We sample data randomly in our training-and test set and see performance is much better for the case of 90% overlap. Consider the following explanations:

i. There are a lot more instances that are highly similar in the case of the 50% overlap compared to the 90% overlap.

ii. For the case of 90% there are more instances to learn from, that could result in better performance.

Which of these explanations could be correct?

(A) both

(B) only (i)

(C) only (ii)

(D) both are not correct

(d) **(1 pt)** In case we focus on decision trees as a machine learning algorithm with a parameter expressing the maximum depth a tree can have. We train the decision tree on our training set. Complete the following sentence: the higher the value of the maximum depth is .....?

(A) the less likely it is that the algorithm overfits the training data.

(B) this will not influence the model.

(C) the shallower the tree will become.

(D) the more likely it is that the algorithm overfits the training data.

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

We are now going to focus on predictive modeling approaches which take the notion of time into account explicitly.

(a) **(1 pt)** Consider the following statements:

i. Differencing can be used to remove a trend from time series data.

ii. ARIMA models have two parameters that need to be set, namely p and q.

Which statement is correct?

(A) both

(B) only (i)

(C) only (ii)

(D) both are not correct

(b) **(1 pt)** What is the advantage of Echo State Networks over regular RNNs?

(A) They have less weights to train.

(B) They have less weights in total.

(C) They have less neurons.

(D) None of the other answers.

(c) (**2 pt**) Consider Figure 5 which shows the errors we obtain on multiple objectives when fitting the parameters of a dynamical systems model.

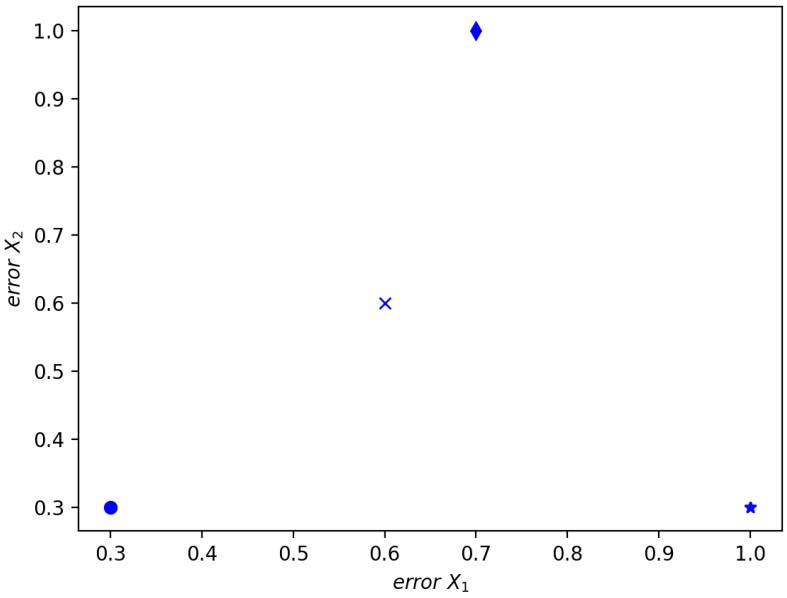

Figure 5: Example error rates on two objectives

Which points are/point is **not** dominated?

(A) The cross and the diamond

(B) The star and the circle

(C) The star

(D) The circle

(d) **(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 lower the chance of getting stuck in a local optimum.

(B) The higher the probability, the lower 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.

(e) **(1 pt)** Which one of the following statements on the temperature parameter in simulated annealing is correct?

(A) The lower the temperature, the less likely it is that a parameter setting with worse performance is accepted.

(B) The temperature stimulates exploration more in the end of the run.

(C) The value of the temperature parameter increases over the run.

(D) The temperature parameter is not part of simulated annealing.