Exam Machine Learning for the Quantified Self
12. 07. 2018
12:00 - 14:45

NOTES:

1. YOUR NAME MUST BE WRITTEN ON EACH SHEET IN CAPITALS.
2. Answer the questions in Dutch or English.
3. Points to be collected: 90, free gift: 10 points, maximum total: 100 points.
4. Grade: total number of points divided by 10.
5. This is a closed book exam (no materials are allowed).
6. You are allowed to use a SIMPLE calculator.

QUESTIONS

1. **Introduction and Removal of Sensory Noise (20 pt)**

   Cristiano is a professional athlete and normally has a very positive attitude and a lot of self-confidence. However due to some setbacks in his professional life he has ended up in a depressed state. His therapist advised him to use a dedicated app to manage his condition. This app uses the sensor information of the phone to track activities and based on the observed sensory information provides suggestions on activities that might be best for Cristiano. To collect data on the impact of activities, the app asks for self-rating every now and then (e.g. "Cristiano, what is your mood at the moment?").

   (a) **(3 pt)** Provide the definition of the quantified self we have discussed in the book and argue whether Cristiano complies to this definition.

   "The quantified self is any individual engaged in the self-tracking of any kind of biological, physical, behavioral, or environmental information. The self- tracking is driven by a certain goal of the individual with a desire to act upon the collected information.". Cristiano definitely complies to this definition, as he uses his sensory information (or in fact the app does) to improve his state of depression.

   (b) **(3 pt)** Identify a supervised learning task for the case of Cristiano.

   Many answers are accepted here, for example, predicting the activity based on the sensory data, or predicting the mood based on the sensory data.

   (c) **(3 pt)** What would be an appropriate step size for the dataset to create a machine learning dataset for the supervised machine learning task you have just identified? Argue why.

   Given that activities are important (and we know 1-1.5 Hz might be relevant for walking for example), a step size of around 250 ms seems appropriate.

(d) **(4 pt)** We want to apply an outlier detection on our data. We are in doubt between using Chauvenet's criterion and a Simple Distance-Based Approach. List two advantages of using the Simple Distance-Based Approach over Chauvenet's criterion.
Simple Distance-Based approach does not have an assumption of a normal distribution. Furthermore, it can take multiple attributes into account at the same time.

(e) **(7 pt)** Consider Figure 1, showing a part of the sensor data of Cristiano. We want to use a lowpass filter to filter out the high frequency noise we observe in the signal. Argue what would be a suitable frequency and show what figure results after applying the lowpass filter.

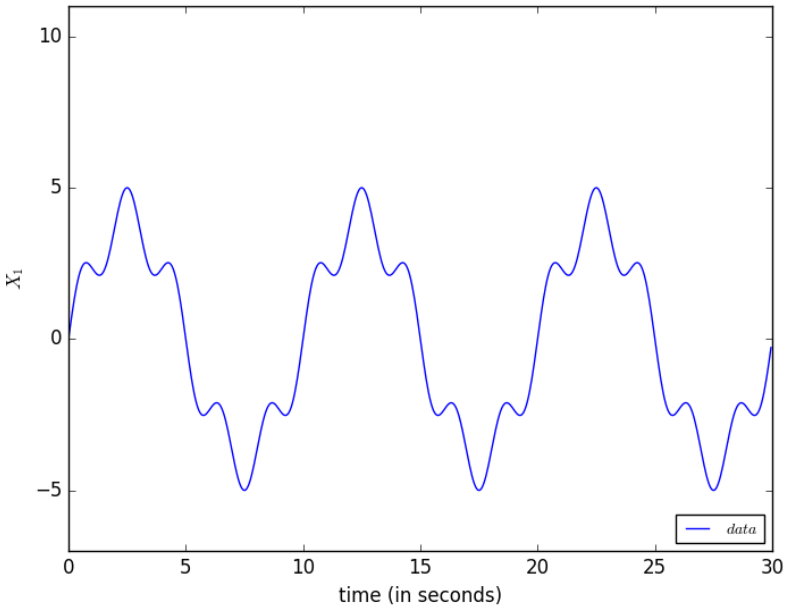

Figure 1: Example dataset

The low frequency signal seems to finish a complete cycle within around 10 seconds (slightly more). Which is around 0.1 Hz. To be on the safe side we could the f_c to 0.15 Hz. The resulting figure would be the same as the original except that the small variations around the low frequency data would no longer be present.

2. **Feature Engineering (15 pt)**

(a) **(3 pt)** Explain the difference between the frequency and the time domain.
"The time domain summarizes values values within a historical window by considering the measure values and applying some aggregation function over it (e.g. mean, slope). The frequency domain considers the periodicity in the historical values and decomposes the signal into sinusoid functions with different frequencies, each having their own amplitude."

(b) **(3 pt)** Imagine we know that only the most dominant frequency is relevant for making predictions in a dataset. Which of the three aggregation functions for the frequency domain would be best to apply for this case? Argue why.

The frequency with the highest amplitude since this would give us the most dominant frequency at the different time points.

Table 1: Example dataset 1

Time point	*Respiration*	*Tired*
0	30	no
1	20	no
2	50	yes
3	30	yes
4	30	no

(c) **(5 pt)** Consider Table 1. Extend the table with a feature in the time domain for the respiration attribute in which you take the mean with a value of $\lambda = 1$. Explain how you came to your answer.

We consider the current and previous time time point (windows size of $\lambda = 1$***) and take the mean of those two:***

Table 2: Example dataset 1: answer

Time point	*Respiration*	*Resp_temp*	*Tired*
0	30	-	no
1	20	25	no
2	50	35	yes
3	30	40	yes
4	30	30	no

(d) **(4 pt)** Provide and explain the NLP pipeline that has been explained during the lecture which we use before we can start identifying attributes in text data.

- ***tokenization: identify sentences and words.***
- ***lower case: change upper case into lower case.***
- ***stemming: map each word to its stem.***
- ***stop word removal: remove stop words from the resulting words.***

3. **Clustering (20 pt)**

Consider the data shown in Table 3, involving multiple people.

(a) **(5 pt)** Given the table in the data, we want to apply a feature based distance metric to compute the distance between the two time-series. Assume we use the mean as feature. Compute the distance between the time series of person 1 and person 2.

Table 3: Example dataset 2

Person 1	
Time point	*Acc. X*
1	8
2	-8
3	8
4	-8

Person 2	
Time point	*Acc. X*
1	-11
2	12
3	-11
4	12

We take the mean of the first series (0) and of the second (0.5) and compute the distance (let's say Euclidean), which ends up being $\sqrt{(0-0.5)^2} = 0.5$

(b) **(6 pt)** Compute the distance between the two time series using the cross correlation coefficient. Show how you came to your answer. In this explanation, also show what the optimal shift τ is.

For the distance we should multiple the numbers of the two series, given a shift τ ***that we make (we actually minimize one divided by this number, which is the same as maximizing this number). With shifts:***

- τ ***= 0: 8*-11 + -8 *12 + 8 *-11 + -8*12 = -88-96-88-96 = -368***
- τ ***= 1 (we shift person 1): -8*-11 + 8*12 + -8*-11 = 88 + 96 + 88 = 272***
- τ ***= 1 (we shift person 2): 8*12 + -8*-11 + 12*8 = 96 + 88 + 96 = 280 We cannot make any other shifts, so the last option is best, with a value of 280.***

(c) **(4 pt)** Explain the difference between k-means clustering and agglomerative clustering.

In k-means clustering you find k clusters. Using agglomerative clustering, you start with each point in each own cluster and combine clusters until you end up with all points in a single cluster. Hence, you have different numbers of clusters at different stages of this process.

(d) **(5 pt)** Explain the subspace clustering algorithm in words. What is the advantage of the clustering method compared to the other clustering approaches that have been discussed during the course?

You define ϵ ***distinct intervals for each attribute. You start with single attributes and in each interval for that attribute look for dense units (that contain more than a certain number of points). These can be combined into clusters if they have a common face or when they are selected. After looking at the single attributes, you look at multiple attributes, find clusters in there, etc. This advantage of the algorithm is that it can handle a large number of features, while the other clustering algorithms cannot.***

4. **Predictive Modeling (15 pt)**

Imagine that we have collected datasets around three people, identified by qs_1, qs_2, and qs_3. We identify their data instances as $(x_{qs_1,1}, \ldots, x_{qs_1,N_{qs_1}})$ for qs_1 (where N_{qs_1} is the

number of data points available for qs_1), and similarly we have $(x_{qs_2,1}, \ldots, x_{qs_2,N_{qs_2}})$ and $(x_{qs_3,1}, \ldots, x_{qs_3,N_{qs_3}})$.

(a) **(4 pt)** Imagine we want to apply predictive modeling on a population level for unseen data of known users, and we assume a temporal ordering in the dataset. Specify what data would go into our training set and what data would go into our test set (you can assume a 60/40 split). Argue how you came to your answer.

We are interested in predicting unseen data of known users. This means that we will train on part of the data of each user. Given that we have a dataset with temporal ordering, we take the first 60% of each user as training set (so we combine these chunks of 60% data over all users) and the remaining 40% as test data.

(b) **(4 pt)** Explain the concept of PAC learnability and how it relates to the VC dimension.

PAC learnability stands for Probably Approximately Correct learnability. A hypothesis set is said to be PAC learnable when it can be shown that given any value of δ, ϵ there is an n where with probability $1 - \delta$ the difference between the in-sample and out-of-sample error is less than ϵ. The VC dimension relates to hypothesis sets and the number of input vectors that can be shattered. It can be shown that any hypothesis set with a finite VC dimension is also PAC learnable.

(c) **(3 pt)** We want to tune the parameters of our predictive model, explain the best way to setup this process to perform parameter tuning. Include the parts of the data you use in your explanation.

Given that we have a temporal ordering in our dataset, it would be best to split our data into a training, validation, and test set in a time ordered fashion. You define values you want to try for each relevant parameter of the learning algorithms under consideration. Then, you do a grid search over those values. For each combination, you train the model on the training set and evaluate the performance on the validation set. In the end, you select the parameter combination that performs best on the validation set. Using those parameter settings, you train the model on the combination of the training and validation set.

(d) **(4 pt)** When we do parameter tuning on recurrent neural networks and echo state networks, which of the two networks would you expect to require the largest number of neurons? Argue why.

The echo state network would require more neurons as the connections between the neurons in the reservoir are randomly initialized (rather than learned like in the recurrent neural networks). This means that more neurons are needed to make sure that the right signal is produced by the reservoir to learn the problem properly.

5. **Reinforcement Learning (20 pt)**

We are going to focus on a reinforcement learning case. We focus on an app we have developed to motivate students to study better and make them pass their exams. We

assume a student can be in five states: *partying*, *studying*, *passed exam*, *failed exam*, and *just woke up*. The initial state of the student is assumed to be *just woke up*. We assume that we can take two actions: (1) send a message to take it easy and take time for *reflection*, or (2) send a message to *start studying*. When we send a message to start studying when the student has just woken up, the student will always go to the studying state. If we however send a message to reflect when the student just woke up he will certainly go to a partying state. Independent of the action performed in the partying state, the student will always end up in the failed the exam state. In the studying state however, when we send a message to start studying again, the student will stress out too much and ends up failing the exam. When a message is sent to reflect in the studying state the student will pass the exam. The rewards per state are shown in Table 4.

Table 4: Rewards

State	*Reward*
partying	5
studying	10
passed exam	100
failed exam	0
just woke up	5

(a) **(5 pt)** Provide a graphical representation of the Markov Decision Process that has just been described.

(b) **(4 pt)** Explain the difference between the two reinforcement learning algorithms *Q-learning* and *SARSA*.

In Q-learning, an action is selected using a certain action selection approach. To determine the value of a selected action, the value of the resulting state is taken, plus the value of the action with the highest value from that resulting state (while our action selection approach does not necessarily take that action with the highest reward). This is called off-policy. For SARSA, the same action selection selection mechanism is used, also to compute the action to select in the next state. This is called on-policy.

(c) **(5 pt)** Compute the final Q-values for the state-action pairs for the case that has been described provided that the Q-learning algorithm is used. Show how you come to your answer.

For each state-value pair we compute the value of the action, meaning we sum up all rewards we obtain in the future, assuming that we take the action with the highest future reward. We assume $\lambda = 1$. We can then see that:

- ***Q(just_woke_up, reflection) = 5***
- ***Q(just_woke_up, start_studying) = 110***
- ***Q(partying, reflection) = 0***
- ***Q(partying, start_studying) = 0***

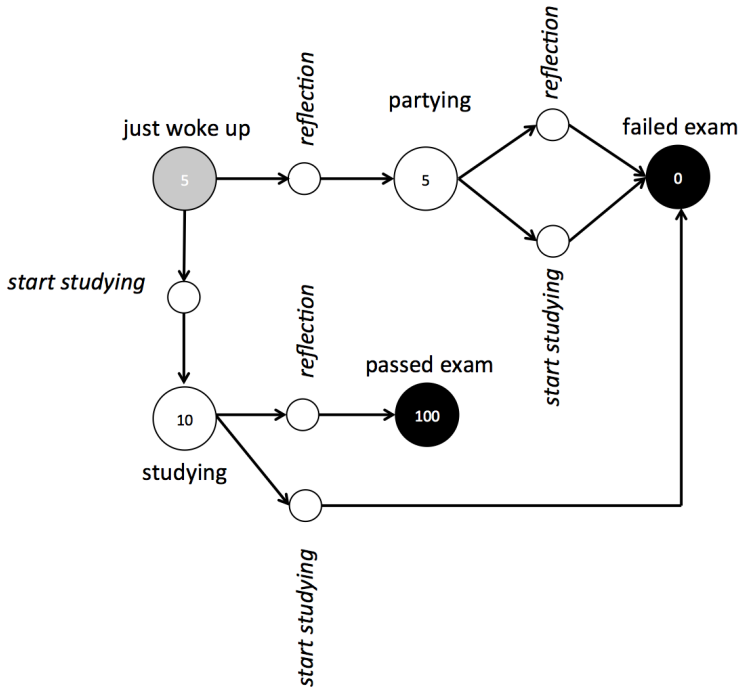

Figure 2: MDP answer

- ***Q(studying, reflection) = 100***
- ***Q(studying, start_studying) = 0***

(d) **(3 pt)** Explain the concept of eligibility traces.

Eligibility traces can be used to put credit on states and actions that have been seen more often in the past. Using these concepts, the estimated values of state-actions pairs are updated more rigorously when these are seen more often in the past.

(e) **(3 pt)** Normally, we assume all state-action pairs are stored in a table. Explain why this is not always feasible and what would be an appropriate approach to remedy this.

It might be that the number of states and actions is very large, making it impossible to store such a huge table. It is possible that estimate the Q-values based on a model (e.g. using a set of weight $\mathbf{w}$***) that maps a state-action pair to a Q-value.***