Exam Machine Learning for the Quantified Self
28. 06. 2019
12:00 - 14:45

NOTES:

1. YOUR NAME MUST BE WRITTEN ON EACH SHEET IN CAPITALS.
2. Answer the questions in Dutch or English (English is preferred).
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 (20 pt)**

   Sylvester is a social guy (despite his rather limited skill to articulate words properly) who likes to challenge his friends. Being quite wealthy he tends to buy the latest gadgets that can make any contribution to understanding and working on his physical shape such as smart watches, chest streps that measure respiration and heart rate, etc.. He then shares the results of such measurements in his network of friends and challenges them to see who is in best shape.

   (a) **(5 pt)** Gimpel *et al.* identify the five factor framework of self tracking motivations. List the five factors and argue which purpose best fits Sylvester.
   (1) self healing; (2) self-discipline; (3) self-design; (4) self-association, and (5) self-entertainment (3 pts, 1 point deducted per missing factor). Any answer could be given for which one is applicable for Sylvester's situation as long as the purposes match (2 pts). For example, it is self-associated as Sylvester wants to relate himself to the community of the quantified self.

   (b) **(3 pt)** To gain insight into the differences between Sylvester and his friends we are going to apply clustering. Explain what learning setup for clustering would be appropriate (in terms of data) to identify different groups among the network of friends of Sylvester.
   Learning would take on the person level, so datasets of the individual friends would be the datapoints for the clustering (1 pt for the setup and 2 points for the explanation).

   While Sylvester is still a healthy guy, he suffers from an addiction to substances that are not good for his health. His doctor recommended Sylvester to use an app which assists him in battling the addiction. This app will take the measurements from the

smart phone and all Sylvester's gadgets and will send interventions when appropriate. Assume the app uses a Reinforcement Learning algorithm to determine when to send an interventions.

(c) **(3 pt)** Specify in words what a reward function could look like for the specific case described above.

A reward function would give a positive value in case Sylvester is not using the aforementioned substances while it would give a highly negative reward in case the substance was taken. More sophisticated variants can also be thought of (1pt for showing the understanding of the reward function, 2pts for a suitable definition for this case).

(d) **(3 pt)** Define a possible state space that would be relevant for the task at hand. Argue why the state space is appropriate.

Any answer could go here, provided that it matches the problem and the concept of a state space. A possible state space could include all values measured for relevant sensors, such as heart rate, accelerometer, GPS and also the proximity of Sylvester to the substance Sylverster is addicted to. This combination would give a good insight into Sylvesters behavior and whereabouts as well as the risk and potential usage of the substance. These can be helpful in determining when to intervene (1 pt for the definition of the state space and 2 pts for explanation)

(e) **(3 pt)** Let us assume that the state space that we have defined is continuous. Would we be able to apply the off-the-shelf Q-learning approach we have discussed during the lecture? Explain why (not).

No, the Q-learning we have considered in the lecture assumes a complete table is made for state action pairs (i.e. a value is assigned to each state action pair). In case of a continuous state space the number of states in infinite, and hence, this approach is no longer feasible (1 pt for mentioning a value is assigned to each unique state-action pair, 2 pts for explaining the state space is infinite).

(f) **(3 pt)** Explain the difference between on policy and off policy Reinforcement Learning.

In on policy learning, the estimates of the value of an action or state are updated by considering the same action selection mechanism in the next state, while in off-policy evaluation a different selection mechanism is assumed. In Q-learning for example, in updating the value of an action it is assumed the action with the highest Q-value is always selected in the next state (1 pt for showing understanding that this is used to select actions, 2 pts for explaining the difference clearly).

2. **Outlier Detection (20 pt)**

(a) (**3 pt**) Explain the difference between distance based outlier detection and distribution based outlier detection.

In distribution-based outlier detection a certain distribution of the data is assumed (e.g. a normal distribution) and outliers are defined based on their position on this distribution (e.g. on the far end of the tail of the normal distribution). In distance based outlier detection no such distribution is assumed, merely a distance metric is assumed, and outliers are detected based on their distance from other data points. In distance-based outlier detection also multiple features can be used together to detect outliers (1 pt for showing understanding of the two different approaches, 2 pts for making the difference clear)

Consider the data shown in Figure 1.

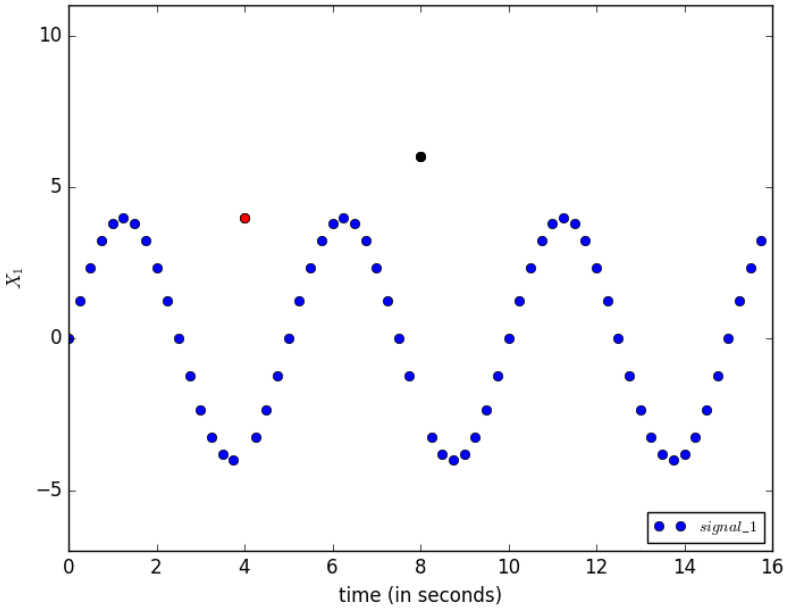

Figure 1: Example dataset - outlier

(b) (**5 pt**) Let us assume we use Chauvenet's criterion to detect outliers with respect to attribute X_1. When we focus on the red measurement in Figure 1 (at 4 seconds, with value 4), do you think this point would be considered an outlier according to this algorithm? And the black measurement at 8 seconds with value 6? Argue for both cases why (not).

The point at 4 seconds will not be an outlier because Chauvent's criterion does not look at the temporal dimension, just the values which it compares to a fitted normal distribution. Given that there are ample values around 4 this point will not be on the very end of the tail of the distribution and hence, not be considered an outlier. The point at 8 seconds however most likely will be as this is the only point in the dataset with such an extreme value and is highly likely to be on the far tail of the distribution (1 pt for

showing understanding of Chauvenet's criterion, and for each of the two cases 1 pt for the correct answer to the point being an outlier or not and 1 pt per explanation).

(c) (**4 pt**) We will apply a very simple version of the Kalman filter (with just identity matrices) to this data to detect outliers and impute the data. Explain what the dataset after the application of the Kalman filter to detect outliers and impute values would look like.

The Kalman filter would build up a model over time (i.e. move from left to right in the graph) and predict the next value based on the previously observed values. It is highly likely that both the red and the black point are far beyond the expected value at that time point, and hence, a new value would be imputed which would be more in line with the previous data. Therefore, the resulting data after imputation would be more in line with the sinusoid wave (2 pts for explaining what the Kalman filter would do, 2 pts for the conclusion drawn).

(d) (**4 pt**) We see that the measurements in this dataset are relatively coarse grained. We want to impute their values using either a mean value imputation or using interpolation. Argue which approach would be most suitable for the case at hand.

Given the fact that this is a temporal dataset, previous and next values are much more natural to use for imputation compared to the overall mean of the data, therefore interpolation is preferred (1 pt. for the answer, 3 pts for the explanation).

(e) (**4 pt**) Someone argues that we should apply a lowpass filter to get rid of potential noise in our data. This person suggests to use a cut-off frequency of $f_c = 0.1Hz$ with a very high order of the filter (i.e. value for n). What would the result of the application of the filter be? Describe your result in words or draw a graph. Explain how you came to your answer.

The lowpass filter filters out frequencies higher than the cut-off frequency (at least to a certain extend), certainly with a very high order of the filter these frequencies will be nearly completely removed. Looking at the figure we see periodic behavior of about one complete cycle in 5 seconds, meaning 0.2Hz. This is above the cut-off frequency. Therefore the current data (certainly the blue points) will be filtered out, resulting in all values around zero (2 pts for showing understanding of the lowpass filter and 2 pts for the correct answer).

3. **Feature Engineering (20 pt)**

Consider the data shown in Figure 2.

(a) **(5 pt)** We apply a Fourier transformation to both signals ($signal\_1$ and $signal\_2$) and use the frequency with the highest amplitude as a feature. Which of the two series would have the highest value for this feature? Argue your choice by explaining how you infer the value of this feature for both series.

The Fourier transformation will assign the high amplitudes to frequencies that occur most prominent in the data. When considering $signal\_1$ ***the***

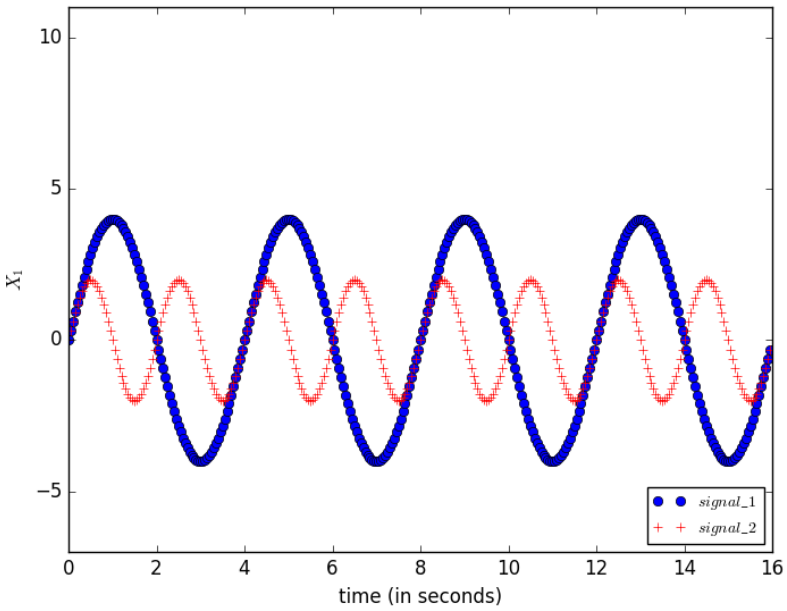

Figure 2: Example dataset - temporal

dominant frequency has a period of 0.25Hz (1 period completed in 4 seconds) while for $signal\_2$ ***it is around 2 seconds per complete period (i.e. 0.5 Hz). These would be the frequencies with the highest amplitude. Hence,*** $signal\_2$ ***would score highest (1 pt for explaining the Fourier transformation, 2 pts for computing the frequencies with the highest amplitudes per signal, 2 pts for the conclusion on the value of the feature and correct answer).***

(b) **(4 pt)** From now on, assume we set our step size for the dataset to be $\Delta t = 1$ *second*. We apply an approach from the time domain, namely we take the mean value over a window size of $\lambda = 3$. Compute what the value of the feature in the time domain would be for the different time points we have available and do so for both signals. Explain how you came to your results.

The window size is set to $\lambda = 3$ ***which means that we consider the current plus the last three time points, Given that we consider a step size of one second it means that for both*** $signal\_1$ ***and*** $signal\_2$ ***respectively one complete period and two complete periods of the sinus wave fit in. This also means that if we summarize these values it will always result in 0. Hence, across the entire set the value for the feature will be zero (2 pts for showing understanding of the aggregation approach and 2 pts for drawing and arguing the conclusion).***

(c) **(4 pt)** In feature engineering we can also consider other types of data next to the numerical data we have discussed so far. Explain the difference between a bag-of-words approach to engineer features from text compared to a topic modeling approach.

In a bag of words approach we consider all different words (all the stem of different words) as individual features. Sometimes we use combinations of these words as features (i.e. $n-grams$). This means a huge number of features. For the topic modeling, a more abstract perspective towards the text is used, namely studying the high level topic the text is about (and finding the topics in the data based on machine learning). This means a greatly reduced number of features (1 pt for explaining bag of words and 1 pt for explaining topic modeling, 2 pts for the comparison)

Table 1: Example dataset - temporal with categorical data

Time point	*Mood*	*Activity level*	*Depressed*
0	positive	high	no
1	positive	high	no
2	positive	low	no
3	positive	high	yes
4	negative	high	yes

(d) (**7 pt**) Apply the algorithm as proposed by Batal *et al.* to extract temporal features in the time domain on the combination of the categorical features *Mood* and *Activity level* on the data shown in Table 1. Consider a window size $\lambda = 1$ and a support threshold of $\Theta = 4/4$ (i.e., 1, this is the minimum support needed). Explain what features result. Explain how you came to these features.

We start with considering the 1-patterns and consider the current time point and the previous time point:

- ***mood = negative, support = 1/4***
- ***mood = positive, support = 4/4***
- ***activity_level = low, support = 2/4***
- ***activity_level = high, support = 4/4***

Hence, we proceed with only the positive mood and high activity level patterns as only these meet the minimum support threshold. We now try to make 2-patterns and consider all different combinations for both "co-occurs" and "before" given our set of 1-patterns (for co-occurs we only consider combinations of the two different attributes as we only have a single value per attribute per time step):

- ***mood = positive (b) mood = positive, support = 3/4***
- ***mood = positive (b) activity_level = high, support = 3/4***
- ***activity_level = high (b) activity_level = high, support = 2/4***
- ***activity_level = high (b) mood = positive, support = 2/4***
- ***mood = positive (c) activity_level = high, support = 4/4***

Hence, we have one 2-pattern and the rest 1-patterns. 3-patterns do not have enough support (3 pts for explaining the correct steps in the algorithms, 4 pts for the calculations with a deduction of 1 pt per wrong pattern).

4. **Clustering (15 pt)**

In this question, we are going to focus on comparing dataset of different quantified selves using clustering.

(a) (**4 pt**) Name the two conditions we have for matching data points in dynamic time warping and explain what they entail.

The two conditions are the monotonicity and boundary condition. The latter expresses that the first and last time points of time series should be matched. For the monotonicity condition it means that when making matches you can only move forward in time or remain at the same time point and not go back in time (2 pts for naming the two conditions, 2 pts for the explanation).

(b) (**3 pt**) When we use dynamic time warping as a distance metric between different datasets, what clustering algorithm is preferred: k-mean clustering or k-medoids clustering? Explain why.

k-medoids is preferred as this takes real datapoints rather than creating an artificial mean as the centre of the cluster (which k-means does). Given that we are considering dataset, the concept of a mean of datasets does not make sense (1 pt for choice, 2 pts for rationale).

(c) (**4 pt**) When we have a large number of features, clustering might not provide us with very insightful results. Explain how the subspace clustering algorithm tries to mitigate this problem.

Rather than looking at the entire set of features, subspace clustering tries to find interesting clusters in subsets of features based on defining ϵ intervals per feature and only splitting up further based on more features when there are intervals that are sufficiently dense (2 pts for answer, 2 pts for explanation).

(d) (**4 pt**) Name and explain one metric that can be used to evaluate the quality of clustering that has been treated during the course.

The silhouette score. It computes the average distance of data points to points within its own cluster and computes the average distance to points in the closest other cluster. The closer you are to points in your own cluster compared to the points in the nearby cluster, the higher the silhouette score will be (1 being ideal, -1 being worst where the points would have been better off in another cluster) (2 pts for the name, and 2 pts for the explanation).

5. **Supervised Learning (15 pt)**

Assume we have a time series available covering four different measurements (attributes) and one target we want to predict, which is numerical. We have 100 time points of data available where we both have values of the attributes and have target information available.

(a) **(4 pt)** In trying to learn based on this problem, one person suggests a recurrent neural network with four hidden layers, each including 1000 neurons, to make sure

the model can potentially capture all interesting relationships. Explain by means of the concepts from the learning theory we have discussed (e.g PAC learnability) whether this would be a suitable choice given the described dataset.

PAC learnability provides guarantees on the difference between the error on the training set and the test set. Only in case the difference is reasonably small will our algorithm be able to learn the task at hand and provide generalizable outcomes. PAC learnability combined with the VC dimension (in case of an infinite number of hypotheses) says something about the amount of data needed give a certain complexity of the hypothesis space. In this case, the amount of data is very limited while the network is highly complex, hence, most likely there will be a large difference between the error on the training and test set, and therefore the complex neural network is probably not suitable (2 pts for showing understanding of the concepts of learnability, 2 pts for the answer (which could also be yes provided that the right arguments are given)).

(b) **(3 pt)** Explain the difference in training procedure between a regular recurrent neural network and an echo state network.

In regular recurrent neural networks, backpropagation through time is applied, updating all the weights of the network by first unrolling the network and propagating the error back into the network, adjusting weight most that contributed most to the mistake made. In echo state networks, only the weights to the output are learned, while the other weights are fixed at random values. These weight are optimized based on the activation of the reservoir and the desired output (1 pt for the explanation of each training procedure, 1 pt for the comparison).

(c) **(2 pt)** We are going to apply an echo state network to this problem. What would be the range you would try for setting the number of neurons inside the reservoir?

In the algorithm values are mentioned between $N/2$ ***to*** $N/10$***, hence, trying between 10 and 50 neurons would do the trick (1 pt for reproducing the range, 1 pt for computing the number of neurons, other answer than this, provided a good argumentation, are also accepted).***

(d) **(3 pt)** We have a dynamical systems model that predicts the next vale of our target based on the current measurement and our attributes. The model has several parameters we can set, which influence the way in which the predictions are made. Give three approaches that can be used to select the best values of these parameters and have been discussed during the course.

The three approaches are:

- ***Simulated Annealing***
- ***Genetic Algorithm***
- ***NSGA-II (no need to provide the full name, but fine if it is done of course)***

(1 pt per approach)

(e) **(3 pt)** We are going to learn on this dataset (of one individual) using an algorithm, how can we best split our dataset into a training, validation and test set? Explain

why.

Given that this is a temporal dataset, it would be good to split it up based on time. For example, the first 60% of the data is used as training set (from the start till we reach 60% of the data), the next 20% as validation set and the final part (ending at the last time point) as test set (1 pt for explaining it should be split based on time, 1 pt for explaining how and 1 pt for mentioning percentages that make sense).