Exam Machine Learning for the Quantified Self

with answers

21. 06. 2017

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 (15 pt)**

   Let us consider Harry. Harry is a typical Apple lover and he buys every single device Apple decides to put on the market. As a result, Harry wears an Apple Watch all the time and also carries his phone around with him at every imaginable moment. While he wasn't such a fan of tracking his health state at first, since he suffered from a stroke he decided to start living a healthier life and he is using his Apple products to assist him. His watch for example, provides him with a goal of an amount of physical activity per day and measures how close he is to his target for the day. He closely monitors this information and tries to reach the goals set for him.

   (a) **(3 pt)** Would Harry adhere to our definition of the Quantified Self? Explain why (not).

   The definition of the Quantified Self is: "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.". Harry is tracking his physical state, so he is engaged in the self-tracking of information. In addition, he uses the information to achieve a certain goals. Hence, he adheres to the definition.

   (b) **(4 pt)** Independent of whether you call Harry a Quantified Self, how would you categorize him in Five-Factor-Framework of Self-Tracking Motivations? Explain how you came to your answer.

   It would the "self-healing" option (or alternatives such as self-discipline, self-design, self-association or self-entertainment could also be accepted if the proper explanation is provided.) because Harry uses the self-tracking to become or stay healthy.

(c) **(4 pt)** Identify two machine learning tasks that can potentially result in insights to better assist Harry.

Pretty much anything goes here as long as it is relevant for the Harry case and explained in sufficient detail. Examples could be:

- ***predicting whether Harry will reach his target for the day (e.g. in the morning of that day)***
- ***predicting Harry's heart rate during given activities.***

(d) **(4 pt)** For each of the two machine learning tasks, argue what would be a proper step size (i.e. granularity for the dataset) for this task, assuming that we collect a temporal dataset.

For the example tasks we have identified these would be:

- ***For the first task identified we would be interested to see how much physical exercise Harry is performing. For this we need to be able to derive the type of activity. Given that walking and running result in periodic behavior or around 1-2 Hz a good step size would be 250ms.***
- ***With respect to the second task, this will probably be done on the same level as we want to relate activities (for which we can use the accelerometer to identify them) to predict the heart rate.***

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

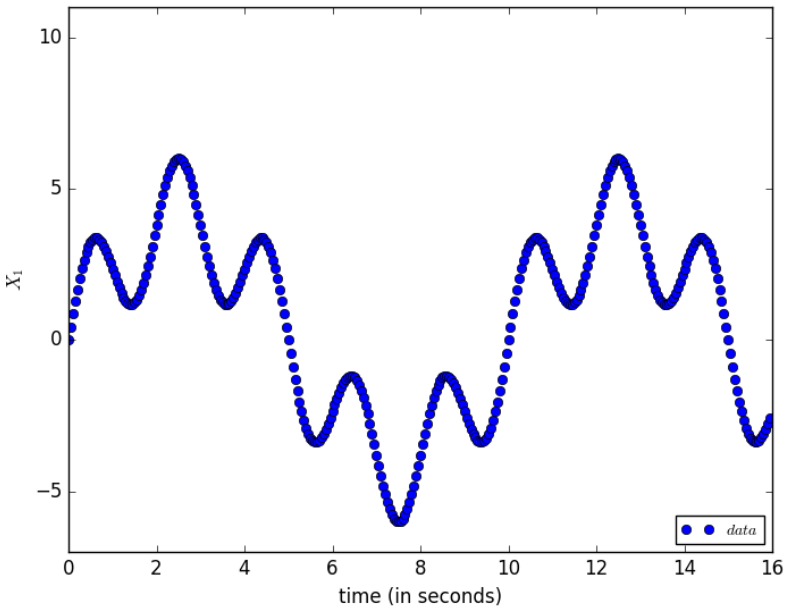

Figure 1: Example temporal dataset

Consider the dataset shown in Figure 1. The figure shows time on the x-axis and the value for attribute X_1 on the y-axis. This dataset shows some periodic behavior we want to exploit.

(a) **(5 pt)** Describe on a conceptual level what a Fourier transformation is.

A Fourier transformation tries to decompose a temporal sequence of measurements with some form of periodicity into a set of sinusoid functions of different frequencies. These frequencies are dependent upon the amount of time steps included in the sequence (this part is optional).

(b) **(6 pt)** If we would apply a Fourier transformation to the data shown above, which frequency/what frequencies would you expect to have a high amplitude. Explain why.

We see a periodic behavior in two forms: a low frequency sinusoid (with a period of about 10 seconds) and a high frequency one with a period around 2 seconds. These can be translated to 0.1 Hz and 0.5 Hz. These are the frequencies one would expect to obtain a high amplitude.

(c) **(3 pt)** List three features that summarize the frequencies and amplitudes into a single number and explain how they summarize the values.

Three examples are:

- ***highest amplitude frequency: perform a Fourier transformation and select the frequency which gets the highest amplitude in the decomposition.***
- ***frequency weighted average: take the product of each frequency amplitude pair, sum them, and divide them by the sum of the amplitudes.***
- ***power spectral entropy: calculate the power spectral density (squared amplitude divided by the number of frequencies, normalized by the total sum of the squared amplitudes) and compute the entropy of those values.)***

(d) **(6 pt)** In addition to the frequency domain, what other domain is available to derive useful features from temporal data? Explain how features can be derived in that domain.

The time domain. We derive values by identifying a historical window of the specific measurements (number of time points considered before for the current measurement plus the current measurement), select the values within that window and applying some aggregation function (e.g. the mean)

3. **Clustering (20 pt)**

Consider the data shown in Figure 2. We see two attributes, X_1 and X_2, both having values in the range [0,1].

(a) **(3 pt)** Explain the purpose of subspace clustering and provide a textual description of the algorithm.

Subspace clustering is meant for cases where there is a huge number of attributes, most common clustering approaches will then result in meaningless clusters. The subspace clustering tries to identify units in subsets of attributes that are sufficiently dense (i.e. contain enough points). It starts with single attributes, splits the range of the values up in regions and identifies intervals such that the unit enclosed by the interval is dense.

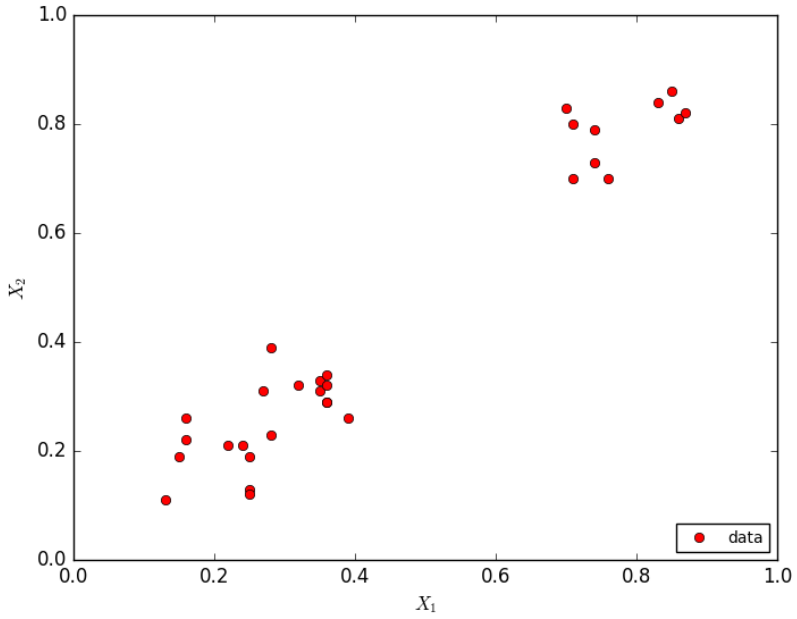

Figure 2: Example clustering dataset

These can then be further explored (seeing whether we can split them up using another attribute). In the end, the clusters are connected dense units.

(b) **(2 pt)** Why is subspace clustering appropriate for the setting of the Quantified Self?
Because we potentially have lots of attributes and traditional clustering approaches cannot cope with such large sets of attributes.

(c) **(6 pt)** Imagine we select a unit to be dense if it contains at least 5 data points, and we create 5 distinct intervals per attribute. Calculate what units would be found with subspace clustering given the dataset shown in Figure 2. Show your calculations.
We have five regions per attribute: [0,0.2], [0.2,0.4], [0.4,0.6], [0.6,0.8], and [0.8,1] for both. If we would start with X_1 we would get one dense unit ([0.2,0.4]). For X_2 we would get two ([0,0.2] and [0.2,0.4]). For the two combined you would get one dense unit (X_1=[0.2,0.4] and X_2=[0.2,0.4]). If only the combination of the two attributes is taken (so only the last dense unit is mentioned) a modest deduction will be applied.

(d) **(6 pt)** Name and explain two raw-based person level distance metrics that have been treated during the lectures.
For example:

- ***Euclidean distance: sequences are assumed to have the same length and are directly compared by taking the Euclidean distance between all measurements.***
- ***Lagged cross correlation coefficient: shift one sequence a bit in time and compute the sum of the product of the pairs of measurements.***

- ***Dynamic Time Warping: create pairs of measurements (originating from the two different sequences) and pair in such a way that the minimum distance between the sequences is reached, provided that: (a) the first pair includes the first point of both sequences; (b) the last pair is the last point of both sequences, and (c) you can not move backward in time when moving ahead in the pairing.***

(e) **(3 pt)** Would k-means clustering be suitable for person level clustering as well? Argue why (not).

Nope, because k-means clustering creates a center of the cluster that is the average over all datapoints in the cluster. In the person level case this datapoint is an entire dataset. Research has shown that this does not work well for this case, nor is it intuitive.

4. **Theory (15 pt)**

(a) **(4 pt)** Imagine we want to apply a perceptron and a multi-layer neural network to a dataset. Which of these two approaches would have a higher VC dimension? Argue why.

The multi-layer neural network, it has much more expressiveness (i.e. a larger hypothesis space) and can therefore shatter datasets with more datapoints.

(b) **(3 pt)** What does the VC dimension tell you about PAC learnability?

In case the VC dimension of a hypothesis set is finite, the hypothesis set is PAC learnable.

(c) **(4 pt)** Which one of the two (perceptron and multi-layer neural network) would need more data to generate a low (we are minimizing the error) out-of-sample error? Argue why.

The multi-layer perceptron as more complex hypothesis sets require more training examples to achieve the same difference between the in-sample and out-of-sample error.

(d) **(4 pt)** If we would have an infinite amount of data, which learning approach (i.e. perceptron or multi-layer neural network) would you expect to work best? Argue why.

The multi-layer neural network, as you would expect that a more complex set of hypotheses (which the neural network is compared to the perceptron) is able to reproduce the true target function better as long as there is sufficient data.

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

Imagine the following dynamical systems model of the relationship between stamina (how much endurance a person has) and the intensity of activities being conducted:

$$\hat{y}_{stamina}(t + \Delta t) = y_{stamina}(t) + \gamma \cdot (y_{activity\_level}(t) - y_{stamina}(t)) \cdot \Delta t \quad (1)$$

$$\hat{y}_{activity\_level}(t + \Delta t) = y_{activity\_level}(t) + \gamma \cdot \Delta t \quad (2)$$

The model basically says that *stamina* increases when the *activity level* is above the current *stamina*. The precise change depends on the parameter γ. The *stamina* decreases when the *activity level* is below the current *stamina*. Furthermore, the *activity level* increases with a fixed value γ. We assume a setting of $\Delta t = 1$. In addition, we have collected a dataset shown in Table 1 about the values for *stamina* and the *activity level*. Finally, we assume the absolute difference to be used as a distance metric (i.e. $E(target) = \sum_{t=0}^{N} |\hat{y}_{target}(t) - y_{target}(t)|$).

Table 1: Example dataset

Time point	*Stamina*	*Activity level*
0	0.5	0.1
1	0.4	0.2
2	0.3	0.3

(a) **(3 pt)** List three machine learning algorithms that have been treated during the lecture and can be used to optimize the parameter λ of the dynamical systems model on the data.

The three are:

- ***simulated annealing***
- ***genetic algorithms***
- ***NSGA-II***

(b) **(5 pt)** Assume we want to predict both *stamina* and *activity level* well. Give an example of two parameter settings for the dynamical systems model whereby one model instance clearly dominates the other model instance when considering the data in Table 1. Explain your reasoning.

An example would be $\lambda = 0$ ***and*** $\lambda = 0.5$ ***If we consider*** $\lambda = 0$ ***we would get the predicted sequence {0.5, 0.5, 0.5} for stamina and {0.1, 0.1, 0.1} for the activity level. This has an error of 0.3 for stamina and 0.3 for activity level. For*** $\lambda = 0.1$ ***we would get the predicted sequence {0.5, 0.46, 0.434} for stamina and {0.1, 0.2, 0.3} for the activity level. This has an error of 0.194 for stamina and 0 for activity level, i.e. better on both. Hence, it dominates*** $\lambda = 0$

(c) **(7 pt)** Give an example of three model instances (i.e three different parameter settings) that do not dominate each other but do have different scores in terms of the two targets. Explain your reasoning.

An example would be $\lambda = 0.1$ ***(resulting in a sequence of {0.5, 0.46, 0.434} and an error of 0.194 for stamina and {0.1, 0.2, 0.3} and an error of 0 for activity level),*** $\lambda = 0.15$ ***(error of 0.15 on both aspects with sequences {0.5, 0.44, 0.41} and {0.1, 0.25, 0.4} for stamina and activity level respectively) and*** $\lambda = 0.2$ ***({0.5, 0.42, 0.396} and an error of 0.116 for stamina and an error of 0.3 for activity level with a sequence of {0.1, 0.3, 0.5})***

(d) **(5 pt)** If we were to use a time series based approach, such as ARIMA, instead of the dynamical systems model, what parameters would we need to optimize to accurately describe the data? And what do these parameters represent?

We would need to optimize p, q, and d. p would reflect the windows for the autoregressive process, q the windows size for the moving average component and d the order of the differencing.