Machine Learning

SDI/Metz, 2021-2022

written test, duration 2h

All paper documents allowed, electronic devices are prohibited. Never anwer by only "yes" or "no", you always have to justify your answers.

Answers can be written in French or English.

1 Milk production (15 point(s))

Let us consider a farm for milk production. The farmer has cows, which give milk approximately once a day (it can be slightly more or slightly less). Milk extraction [1] is automated. A cow, when it feels ready for that, goes by itself on a milking machine. The cow is identified by the machine thanks to a tag. The weight of the cow is measured at that time as well. Then the machine handles the milk extraction and releases the cow.

The farmer would like to analyse the data collected by this process. Indeed, the milking machine records milking events $e \stackrel{\text{def}}{=} (h, v, c, w, r, a)$, where h is the time of the day (e.g. 14h23), v the volume of milk extracted in litters (e.g. 30), c the identifier of the cow (e.g. 13782), w the weight of the cow expressed in tons (e.g. 0.8) and a the age of the cow, expressed in days (e.g. 572). The race of the cows in the farm are $r \in \{\text{limousine, normande, charolaise}\}$.

1.1 Preprocessing

Let us first *ignore* the time attribute h. It will be considered later.

▷ **Q1** : Propose a preprocessing of the data (without h) suitable for managing samples as vectors (for the purpose of vector quantization, SVMs, ...). *(2 point(s))*

You can use one-hot for the r argument. In order to avoid numerical issues when distances will be computed, we have to make the values homogeneous. E.g. volume can be divided by 25, the weight is ok when expressed in tons, the age could be a rescale of the ages where cow can give milk in the [0, 1] interval. Attribute c needs to be ignored since it hosts no information about the milking.

Now, let is consider the h attribute.

▷ **Q2** : What is the problem with the h attribute ? *(1 point(s))*

We can express it in minutes from 0h00, to avoid sexagesimal computation. With this, it cannot be manipulated as a vector, since 0h00 is very close to 23h59 while the numerical difference is high.

▷ **Q3** : Propose computing rules for that attribute to solve the issue *(1 point(s))*

We can use $e^{i\theta}$ complex numbers, with the arithmetic in $\mathbb{C}$ and a resetting of the result to the closest θ. Another solution would be to replace that attribute with the time elapsed from the previous event, in order to make the time relative.

In the following, we consider now that h attribute is handled in an appropriate way, and that the preprocessing is applied. Events e denotes now such preprocessed events.

1. "milking" in English, "traîte" in French

1.2 Analyze each cow

Let us consider a particular cow c by filtering all the events having its identifier as the c attribute. Let us denote this sample set by S_c.

▷ **Q4** : Do you think that S_c could be modelled as i.i.d [2] ? *(1 point(s))*

No, since once a cow is milked, it cannot be milked again immediately. So the amount of milk and the hour depend on the previus milking event. Events are therefore not independent.

▷ **Q5** : How would you implement a detector of health problems from S_c. *(2 point(s))*

It is an anomaly detection problem. Here, the anomaly is the amount of milk got, or weight loose, etc... Any unsupervized-learning method could be tested (1-class SVM, vector-quantization...).

1.3 Analizing the whole dataset

Let us define $S_a \stackrel{\text{def}}{=} \{(h, v, c, w, r, \alpha) \mid a - \epsilon < \alpha < a + \epsilon\}$ extracted from all the collected events. Let us consider the preprocessing in section 1.1 as correctly implemented, and let us consider the dataset as i.i.d, whatever your anwers to questions in section 1.2.

For a given a let us implement a 10×10 Kohonen Self-Organizing Map (SOM) fed by S_a.

▷ **Q6** : Describe the expected organization of the prototypes. *(1 point(s))*

On the surface of the map, close prototypes should be similar milking events (e.g same race, same hour, same volume). The variety of events should be spread on the whole map surface.

▷ **Q7** : Do you expect any difference with a map that is fed with all the events ? *(1 point(s))*

The map may be richer (if it can...) since the milking features, as volume, may strongly depend on the age of the cow.

Let us suppose that for each value a, S_a contains about 10000 events (i.e. $\forall a,\ |S_a| \simeq 10000$).

▷ **Q8** : What happens if you feed the SOM with all the events, but ordered by increasing a ? *(1 point(s))*

.

We generate a non stationary distribution, that can be tracked by successive and continuous adjustments of the arrangement of the prototypes (as we did for the digits, the map adapts to a change in the statistics of the input (passing to all digits to only odd once, for example). Tracking supposes that the h-radius and the learning rate are kept constant. If you considered that it was not the case, say that lastly presented examples may have less effect.

1.4 Prediction of milk production

Let us consider the preprocessing in section 1.1 as correctly implemented, and let us consider the dataset as i.i.d, whatever your anwers to questions in section 1.2.

We want to orient the cows to the milking machine, if we think that the cow will give a desired amount of milk. Weight sensors and tag readers are placed all around the farm, so we can detect the instantaneous weight of some cow at some time instant. The idea then is to predict the amount of milk v we could expect for that cow if we decide to milk it at that time.

▷ **Q9** : From the dataset that we have (i.e. all the collected events), which SVM would you use to build up such a predictor ? *(1 point(s))*

This is a regression problem, label is scalar. I recommend an ϵ-SVR.

▷ **Q10** : Discuss the choice of a suitable kernel. *(1 point(s))*

2. identically and independently distributed.

I expect the dependance of the labels as something smooth, where noise is present. Maybe a linear kernel does the job (inputs are in R^4 since we do not use c.)

▷ **Q11** : A cross-validation process returns 4. What does it mean ? Is it a good or a bad result ? *(1 point(s))*

If the risk is computed from the l_2 loss, it means the expected precision is 2, since CV estimates the real risk. As cows deliver quantities like 30 litters, this is quite good. Moreover, there is certainly an intrinsic noise in the milk production. Expecting a very small real risk, even with a good algorithm, seams unrealistic. So 2 is certainly a very good result.

Let us build up that predictor, and call p this labelling function, learnt/fitted from our dataset. We give p to *another* farmer, so that s/he can predict the amount of milk of his/her cows. That farmer uses p on 20000 milking events, and s/he measures an error that, in average, is 10.

▷ **Q12** : This measurement is a statistical risk about p... but of what process ? Is it a real or an empirical risk ? *(1 point(s))*

This empirical risk is measure on a dataset that had not influenced the building of p. The law of big numbers applies, so this is an estimation of the real risk of the predictor p, when used to predict the milking on the second farm, i.e. when used with samples provided according tothe statistics of the second farm.

▷ **Q13** : What would you conclude about the difference between the 4 and the 10 values observed in that story ? *(1 point(s))*

It means that the statistics of the second farm is not the same as the one for which p has been trained. For example, the amount of old cows could be very high while it was low on the previous farm.

2 Dimensionality reduction for topic classification (5 point(s))

In this problem we are interested in predicting the topic of a message posted on a forum. We identified 7 classes of interest (`politics`, `electronics`, `computer science`, `mathematics`, `religion`, `sport`, `cryptocurrency`) and scrapped messages posted on thematic forums. We scrapped 18846 posts which vary in size for 10 to 500 words. If we collect all the unique words from all the collected messages, we end up with approximately 400.000 words. Each message has been automatically labeled with the class corresponding to the theme of the forum from which the message originates. We want to use a predictor to automatically tag messages posted on arbitrary forums.

As an example, you are given below a post on the `sport` topic :

```
From: paul@csd4.csd.uwm.edu (Paul R Krueger)
Subject: Brewer bullpen rocked again...
Organization: Computing Services Division, University of Wisconsin - Milwaukee
Lines: 30
Distribution: world
NNTP-Posting-Host: 129.89.7.4
Originator: paul@csd4.csd.uwm.edu

For the second straight game, California scored a ton of late runs to crush
the Brewhas. It was six runs in the 8th for a 12-5 win Monday and five in
the 8th and six in the 9th for a 12-2 win yesterday. Jamie Navarro pitched
seven strong innings, but Orosco, Austin, Manzanillo and Lloyd all took part
in the mockery of a bullpen yesterday. How's this for numbers? Maldanado has
pitched three scoreless innings and Navarro's ERA is 0.75. The next lowest
on the staff is Wegman at 5.14. Ouch!

In other news, Nilsson and Doran were reactivated yesterday, while William
Suero was sent down and Tim McIntosh was picked up by Montreal. Today's game
with California was cancelled.

--salty
```

▷ **Q14** : Given the large dimensionality of the vocabulary, which preprocessings **not involving the target** do you suggest to reduce the size of the vocabulary ? *(1 point(s))*

We can remove some variability of the texts without impacting the classification performance : spell correction to remove useless variability, stemming as the derivation of the word may probably not be informative on the topic, removing the punctuation, ... The first few lines of the header do not seem to be very informative as well although the origin of the email as well as the organization can provide some hints.

▷ **Q15** : Describe **with sufficient details** a machine learning pipeline involving a filtering approach for the dimensionality reduction for solving our machine learning problem. You are expected to completely specify (formalize) the inputs, outputs and algorithms involved in every step of your pipeline. *(2 point(s))*

A univariate filter approach involves a heuristic for excluding the features the most independent from the target. A general pipeline is the following. The first step is to vectorize the document as a bag of word. There are multiple possibilities for this vectorization but we may consider projecting each document on a vector of size 400.000 *filled with either 0 or 1 to indicate if the i-th word is present or not in the document. Given that each feature is categorical (word present or absent) and the target is categorical as well, we can use a* χ^2 *test for selecting the input dimensions. We may select the* N *(*N *to be fixed) less independent features. Following that filter, we need a classifier which takes as input a vector of size* N *(the presence/absence of the* N *selected words) and outputs a class among* 7 *of the target classes. The classifier can be any predictor such as a SVC, a decision tree, etc.. The whole pipeline can be cross validated and the hyperparameters estimated with grid-search with a cross-validated score for scoring the hyperparameters.*

▷ **Q16** : Describe **with sufficient details** a machine learning pipeline involving a wrapper approach for the dimensionality reduction for solving our machine learning problem. You are expected to completely specify (formalize) the inputs, outputs and algorithms involved in every step of your pipeline. *(2 point(s))*

A multivariate wrapper involves a predictor for scoring a subset of features as well as an algorithm for generating the candidate subsests of features. For the classifier, we can use a vectorizer (which transforms a document as a vector of presence/absence of the selected words) followed by a predictor such as a SVC, decision tree classifier, logistic regression, ... Its real risk can be estimated by cross-validation withe accuracy (fraction of correctly classified samples) measure. The algorithm for generating the candidate subsets of features can be a sequential forward floating search.
The last steps of the pipeline are similiar to the filter case of the previous question and the comment about cross-validation and hyperparameter search also apply.