Machine learning

Test annals

1 Asian cuisine (9 points)

A factory produces two kinds of spring rolls (nems), beef nems and pork nems. The two kinds can be distinguished from their length l and their diameter d.

When I buy a bag of frozen nems, I get a mix of the two kinds. Let us denote by x the ratio of pork nems (and thus $(1-x)$ the ratio of beef ones) that the factory produces, and that I have in my bag.

The distribution of (l, d) pairs differ according to the kind of nems (pork or beef), but these distributions are unknown to me. Figure 1 shows the distribution. Distributions are such that for each kind, dimensions (l, d) are distributed *uniformly* within a rectangular area. The rectangles for both kinds differ, as the figure shows.

I have a brand new outstanding supervised machine learning algorithm at my disposal, named smurtz. I want to use it in order to predict the taste of a nem from its shape. I am able to determine that taste... when I eat the nem. This is how I set up a learning database, from the nems in my bag, because I can measure (l, d) and get the taste (pork, beef) for each one.

Question 1.1 (.5 point) : How is the performance of schmurtz defined ? Give the definition of a numerical criterion, that is weak when the algorithm performs well. Name that criterion.

Question 1.2 (2 points) : I only put the lengths l and the tastes of the nems in the database. Give a theoretical lower bound $p_l(x)$ for the performance criterion, according to x, using the knowledge represented in figure 1.

Hint : Consider the three ranges denoted by A, B, C in the figure. For each of them, compute the probability to be in that range, and then estimate the best performance on that range.

Question 1.3 (2 points) : Same question for the bound $p_{l,d}(x)$, when both l and d are used in the database. Plot $p_l(x)$ and $p_{l,d}(x)$. Is it worth considering the diameter of the nems ?

To each measure (l, d) obtained from some nem, we decide to associate a vector $\phi(l, d)$ defined as :

$$\phi(l,d) = \begin{pmatrix} a & = & d \\ b & = & l^2 - 3d \\ c & = & d + l^2 \\ d & = & l \end{pmatrix}$$

Question 1.4 (1 point) : If we put now the values $(a, b, c, d) \in \mathbb{R}^4$ in the database, rather than the values $(l, d) \in \mathbb{R}^2$, determine the performance bound $p_{a,b,c,d}$ for algorithm schmurtz.

I get a 100-nem bag from my freezer, and I am not aware of the distribution (from the factory) depicted in figure 1. I observe, after having eaten a small piece of all nems, that there are 47 pork nems and 53 beef nems in the bag.

Question 1.5 (.5 point) : Is it reasonable to assess that $x = .5$?

In the following, let us consider $x = .5$, whatever the answer to the previous question.

Question 1.6 (1 point) : Instead of schmurtz, I use a C-SVC [1] with a Gaussian kernel (σ is the parameter). As you know the statistics in the figure 1, do you think I may get a null empirical risk with my bag ? If you think so, tell in what case. Otherwise, explain why it is impossible.

Question 1.7 (1 point) : I use now the basic dot product of $\mathbb{R}^2$ as a kernel. Re-draw the figure 1 and add on the drawing a 0.5 performance separator, as well as a minimal perfromance separator (i.e. the best).

An advertisement tells me that, on next fool's day, the nem producer will add a nem filled with tooth paste in the bags. That nem has the shape of a cylinder, as regular ones, but the (l, d) values for that nem differs from the pork as well as the beef nems.

[1] The bi-class SVM presented during the lecture.

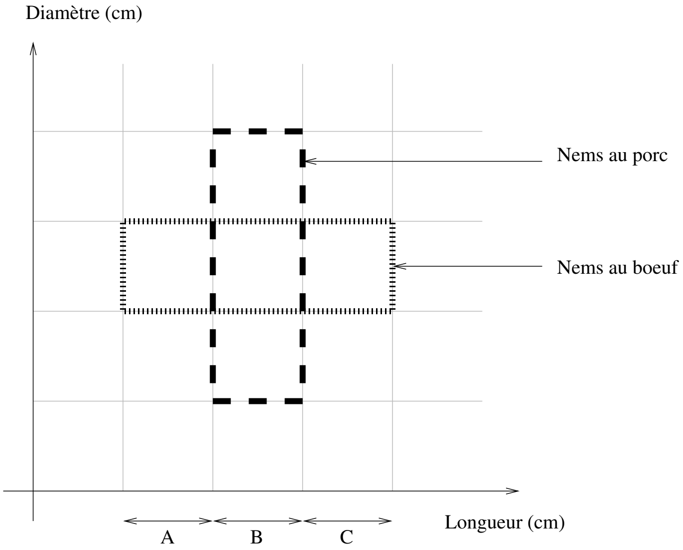

Figure 1 – Nem distribution. Pork nems are distributed uniformly over the vertical rectangle surface, and beef nems are distributed uniformly over the horizontal rectangle surface.

Question 1.8 (1 point) : From the bag I have today, and without knowing the statistics in figure 1, can I set up a tooth paste nem detector ? If you think I can, tell how. Otherwise, justify.

2 Dead leaves (10 points)

In a waste upgrading factory, there are two dead leaves containers. An operator uses a steam shovel to extract the leaves and put them in a incinerator (cf. figure 2). Let us use the numerical machine learning methods addressed in the lecture in order to investigate the relation between the commands of the shovel and the position of its end.

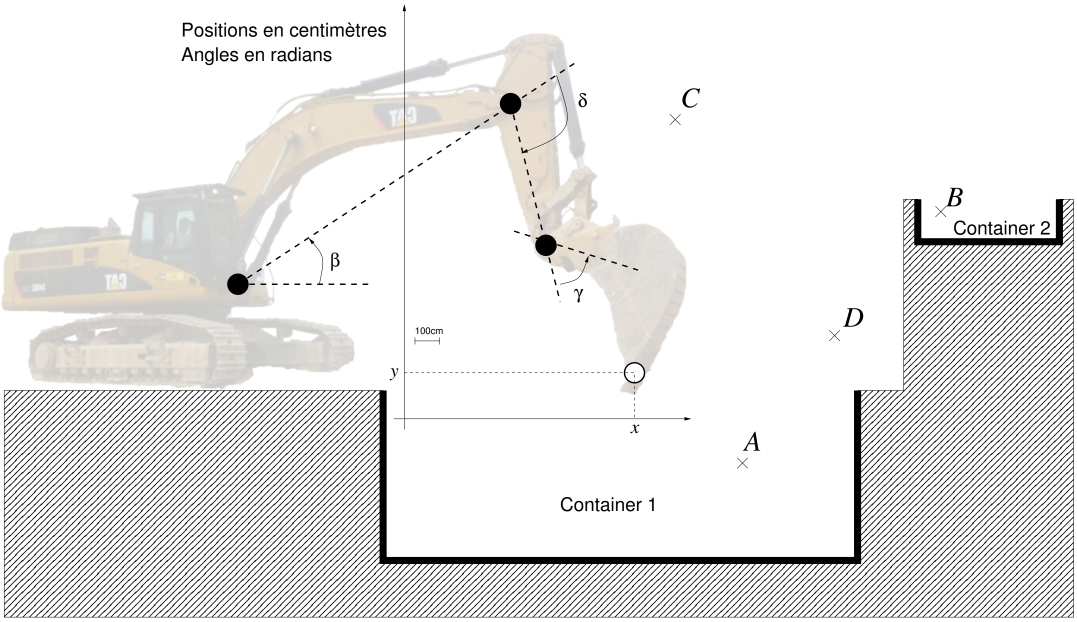

Figure 2 – The steam shovel and the containers. See text for the notations. We consider that **the inner space of both containers is reachable by the shovel end**.

2.1 Spin (1.5 points)

The steam shovel can spin. The operator uses a lever that he pushes with an angle $\bar{\alpha}$. There is a mapping between each $\bar{\alpha}$ angle of the lever and the spin angle α of the steam shovel. This mapping is unknown.

Let us learn the mapping with an ϵ-SVR, with $\epsilon = 10^{-5}$, using a basic dot product as a kernel. The idea consists in predicting the rotation angle α from the command $\bar{\alpha}$. Angles are expressed *in radians*. We observe the operator working, and we set up a database S with 5000 examples, for which the operator has tried random lever positions $\bar{\alpha}$. Each sample i in the database is a pair $(\bar{\alpha}_i, \alpha_i)$. Once the SVM has converged, we get a predictor h. Let us then compute :

$$E = \frac{1}{5000} \sum_{i=1}^{5000} (h(\bar{\alpha}_i) - \alpha_i)^2$$

Question 2.9 (0.5 point) : What is the notion introduced in the lecture that corresponds to E ?

Question 2.10 (1 point) : We obtain $E = 10^{-10}$, can we conclude that h is a good predictor for the value α ?

2.2 Position of the end of the shovel (5.5 points)

Let us consider now a steam shovel that do not spin (α constant), and let us consider the relation between angles $(\bar{\beta}, \bar{\delta}, \bar{\gamma})$ of 3 control levers and the position (x, y) de la shovel end. These levers control respectively the angles (β, δ, γ) or the mechanical arm. **Angles are expressed in radians and positions in centimeters**, from a fixed origin (cf. figure 2).

2.2.1 Mechanical model (1 point)

We have the documentation of the shovel at our disposal, and thus we have access to an accurate mechanical model $\mathcal{M}$ of both the commands and the articulated elements of the arm. Let us use this model as a predictor : $\mathcal{M}(\bar{\beta}, \bar{\delta}, \bar{\gamma}) \rightarrow (x, y)$. In real use case, the motors generate vibrations of the shovel. Let us build a learning database S by sampling every 10 seconds the levers and the shovel end position, while the operator is working.

$$S = \{(\bar{\beta}, \bar{\delta}, \bar{\gamma}, x, y)_1, (\bar{\beta}, \bar{\delta}, \bar{\gamma}, x, y)_2, (\bar{\beta}, \bar{\delta}, \bar{\gamma}, x, y)_3, \cdots (\bar{\beta}, \bar{\delta}, \bar{\gamma}, x, y)_N\}$$

Question 2.11 (0.5 point) : Does $\mathcal{M}$ have a null real risk ? Why ?

Question 2.12 (0.5 point) : Does $\mathcal{M}$ have a null empirical risk ? Why ?

2.2.2 Prediction using an SVM (4.5 points)

Let us consider that we do not know the mechanical model. We want to use a SVM with a Gaussian kernel (parameter is σ) in order to set up a predictor similar to the mechanical model. The operator controls the shovel while maintaining the end opened ($\gamma = \gamma_{\max}$), and then he closes it to grab the dead leaves. Let us record in the database S the angles $(\bar{\beta}, \bar{\delta})$ and the positions (x, y) just before the closing[2]. The database thus contains the commands $(\bar{\beta}, \bar{\delta})$ corresponding to the positions desired by the operator.

$$S = \{(\bar{\beta}, \bar{\delta}, x, y)_1, (\bar{\beta}, \bar{\delta}, x, y)_2, (\bar{\beta}, \bar{\delta}, x, y)_3, \cdots (\bar{\beta}, \bar{\delta}, x, y)_N\}$$

Question 2.13 (1 point) : What kind of SVM would you choose ? What are the inputs, the outputs ?

Question 2.14 (1 point) : Suggest, with justifications, a suitable value for σ.

Question 2.15 (0.5 point) : How can you determine the C or ν parameter of the SVM ?

[2] The lever angle $\bar{\gamma}$ is not registered in the database, since its value is always $\gamma_{\max}$ when the samples are acquired.

Question 2.16 (1 point) : From a cross-validation, the real risk is estimated to be 4. In the actual case of our problem, what is the signification of that number ?

Question 2.17 (1 point) : The database S has been set up while the operator was working on container 1. What could you say about the performances of the predictor we have got when applied to positions in container 2 ?

2.3 Inverse model with Kohonen maps (3 points)

Let us analyse the distribution of the element of S defined in paragraph 2.2.2 with a Kohonen self-organizing map. Let us suppose that S was obtained from observations of the operator *when he was working on both containers*. Let us recall that the elements of S belong to $\mathbb{R}^4$. We will use a 10×10 grid-shaped Kohonen map.

Question 2.18 (1 point) : Can we use the Euclidian distance of $\mathbb{R}^4$ to compare the samples with the prototypes ? If you think so, justify it, otherwise, propose a suitable distance.

Question 2.19 (1 point) : The operator wants to position the shovel end at point A (cf. figure 2). How could we suggest him a suitable command $(\bar{\beta}, \bar{\delta})$ from the Kohonen map (once it has converged, of course) ?

Question 2.20 (1 point) : Will the method of the previous question give good results for points B, C, D in the figure 2 ?

3 Machine Learning test prediction

During the academic year, the students of an university have to take n exams. Let $m_i^e \in [0, 100]$ the mark obtained by student e for the exam i. In addition to these exams, there is a final test for accessing the next year program. The result of that final test for a student e is denoted by $f^e \in \{\texttt{pass}, \texttt{fail}\}$. We would like to determine from the exams' marks wether the student will pass the final test. We have at our disposal the data set $S = \left\{(m_1^e, m_2^e, \cdots, m_n^e, f^e)\right\}_e$ corresponding to the 500 students of the previous year. Let us denote by $|X|$ the number of elements in a set X : $|S| = 500$. Let us assume that the students' statistics are the same over the years, and that the exams are the same too.

3.1 Failure prediction from the mean (7 points)

Let us use the average marks of a student in order to know if s/he will pass the test. Let $\mathcal{H}$ be the set of functions h_θ defined as

$$h_\theta\left(m_1, m_2, \cdots, m_n\right) = \left\{\begin{array}{ll} \texttt{pass} & \text{if } \frac{1}{n}\sum_i^n m_i \geq \theta \\ \texttt{fail} & \text{otherwise} \end{array}\right. \tag{1}$$

Let us also use the following notations [3] :

$$\begin{array}{rcl} S^+ & = & \left\{(m_1, m_2, \cdots, m_n, f) \in S : \ f = \texttt{pass}\right\} \\ S^- & = & \left\{(m_1, m_2, \cdots, m_n, f) \in S : \ f = \texttt{fail}\right\} \\ S_h^{\mathrm{FP}} & = & \left\{(m_1, m_2, \cdots, m_n, f) \in S : \ h\left(m_1, m_2, \cdots, m_n\right) = \texttt{pass} \text{ and } f = \texttt{fail}\right\} \\ S_h^{\mathrm{FN}} & = & \left\{(m_1, m_2, \cdots, m_n, f) \in S : \ h\left(m_1, m_2, \cdots, m_n\right) = \texttt{fail} \text{ and } f = \texttt{pass}\right\} \end{array} \tag{2}$$

Question 3.21 (1 point) : By using the above set definitions and the $|X|$ notation, write some mathematical expression of the empirical risk of h_θ.

Let us consider the following algorithm, called the $\mathcal{M}$ method, for the setting of a $\theta^\star$ parameter such as $h_{\theta^\star}$ is a good predictor.

- Compute $A = \left\{\frac{1}{n}\sum_i^n m_i\right\}_{(m_1, m_2, \cdots, m_n, f) \in S^+}$ and $B = \left\{\frac{1}{n}\sum_i^n m_i\right\}_{(m_1, m_2, \cdots, m_n, f) \in S^-}$.
- Compute $\mu_A = \frac{1}{|A|}\sum_{a \in A} a$ and $\mu_B = \frac{1}{|B|}\sum_{b \in B} b$.
- $\theta^\star = \frac{\mu_A + \mu_B}{2}$.

[3] The notation $\{A : B\}$ means "the set of As such as B".

Question 3.22 (2 point) : Do you think that $\mathcal{M}$ implements a minimization of the empirical risk as an induction principle ? If your answer is yes, justify it, otherwise, exhibit a case for which you can find a better θ (in terms of empirical risk) than the one returned by $\mathcal{M}$.

Question 3.23 (2 points) : Propose a method, different from $\mathcal{M}$, that relies on the empirical risk minimization (ERM) on $\mathcal{H}$.

Question 3.24 (2 point) : Do you think that there are some methods which compute some h_θ with good performances while being subject to overfitting ? If you answer yes, give an example, otherwise, justify.

3.2 Prediction from all the marks, linear SVM (4 points)

Let us use a ν-SVC with the standard dot product in $\mathbb{R}^n$ as a kernel and $\nu = 0.1$. We work with the dataset S defined previously. We measure an 0.01 empirical risk.

Question 3.25 (2 points) : Can I affirm that the predictor produced by the SVM will have good performances ? Justify. Is your answer dependent on n ? on ν ?

I apply a cross-validation and find a 0.012 estimated real risk.

I had the final test yesterday and today I got the marks $(m_1, m_2, \cdots, m_n)$ for my exams this year. When fed with my marks, the predictor given by the SVM returns `pass`.

Question 3.26 (2 points) : Can I conclude that it is very probable that I will pass the final test ?

3.3 Prediction from all the marks, non-linear SVM (6 points)

Let us assume that, a opposed to the previous question, with the usual dot product for kernel (i.e. the SVM is still linear), the empirical risk is 0.5.

Question 3.27 (1 point) : What do you think about the real risk of this SVM ?

Let us continue with a ν-SVC as previously, with $\nu = 0.1$, but the kernel is now a Gaussian kernel with a σ parameter.

Question 3.28 (1 point) : Using $\sigma = 1$, we compute the empirical risk of the SVM and find 0.01. Is there a chance that the SVM overfit ?

Question 3.29 (1 point) : Using $n = 4$ exams, give an approximate value for a suitable σ.

Let us suppose that the final test only evaluates skills in physics and mathematics. The subjects related to the exams are listed in table 1.

Mark	Subject
m_1	Mathematics
m_2	Literature
m_3	Sports
m_4	Physics
m_5	Art

Table 1 – Exams

It is reasonable to assume that marks m_2, m_3 et m_5 are not related to the student's ability to pass the final test.

Question 3.30 (1 point) : Would we get better performances if we restrict the dataset to $\mathbb{R}^2$, i.e. if we work with $S' = \{(m_1, m_4, f)\}_{(m_1,m_2,m_3,m_4,m_5,f)\in S}$? Justify.

Let us now suppose that I do not know the subjects related to the marks $(m_1, m_2, m_3, m_4, m_5)$. I consider that some of them may have no relation with the ability to pass the final test.

Question 3.31 (2 point) : Describe a procedure for the selection, among the 5 marks, of those which are actually relevant for predicting the final test result.

3.4 In the land of good students (3 points)

Let us now consider that the final test evaluates a skill level that, normally, all the students can reach if they are serious. The last year students were all serious, so *they all pass the final test.* Let us work with the predictors in the set $\mathcal{H}$ defined at the beginning of this text.

Question 3.32 (1 point) : Give a value for the θ parameter such as the real risk, estimated by cross-validation on s, is as low as possible.

We would like to detect students who, this year, may not be serious, and thus may fail in the final test.

Question 3.33 (2 point) : Can we set up a detector from S? If the answer is no, justify, otherwise propose a method.