IFT-6390 Fundamentals of Machine Learning
Professor: Ioannis Mitliagkas

Homework 1 - Theoretical part

Devoir 1 - Partie Théorique

Student: Gustavo Alonso Patino Ramirez

1. **Probability warm-up: conditional probabilities and Bayes rule** [5 points]

   Rappels de probabilités: probabilité conditionnelle et règle de Bayes

   (a) Assuming the probability of an event B, $P(B) \neq 0$, the conditional probability of an event A given that B have occurred is defined as

   $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

   (b) Our sample space is $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. We are interested in

   $$P(2heads \mid 1\,toss\,is\,head) = \frac{P(2heads, 1toss\,is\,head)}{P(1toss\,is\,head)} \tag{2}$$

   where

   $$P(1toss\,is\,head) = P(\{HHH, HHT, HTH, HTT\})$$

   using the total theorem

   $$\begin{aligned} P(1toss\,is\,head) &= P(\{HHH\}) + P(\{HHT\}) + P(\{HTH\}) + P(\{HTT\}) \\ &= \frac{2}{3}\frac{2}{3}\frac{2}{3} + \frac{2}{3}\frac{2}{3}\frac{1}{3} + \frac{2}{3}\frac{1}{3}\frac{2}{3} + \frac{2}{3}\frac{1}{3}\frac{1}{3} = \frac{18}{27} \end{aligned} \tag{3}$$

   in the same way

   $$\begin{aligned} P(2heads, 1toss\,is\,head) &= P(\{HHT, HTH\}) = \\ &\frac{2}{3}\frac{2}{3}\frac{1}{3} + \frac{2}{3}\frac{1}{3}\frac{2}{3} = \frac{8}{27} \end{aligned} \tag{4}$$

substituting in Eq.2

$$P(2heads \mid 1\,toss\,is\,head) = \frac{8}{27}\frac{27}{18} = \frac{4}{9}$$

(c) Equivalent expressions for $P(X, Y)$:

(i) $P(X, Y) = P(Y \mid X)P(X)$

(ii) $P(X, Y) = P(X \mid Y)P(Y)$

(d) Prove Bayes theorem:
Using the definition of conditional probability and the result from item (c) :

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)} = \frac{P(Y \mid X)P(X)}{P(Y)} \tag{5}$$

(e) We consider the event of a student being affiliated to UdeM as AM, and the event of being affiliated to McGill as AMC

i. In the first leaf of the decision tree it holds that $P(AM) + P(AMC) = 1$, so $P(AMC) = 0.45$

ii. Using the Bayes theorem, and defining the event bilingual as bil, we know that

$$P(AMC \mid bil) = \frac{P(bil \mid AMC)P(AMC)}{P(bil)} \tag{6}$$

using the total theorem and the 2 leaves where the event happens

$$P(bil) = 0.55 * 0.8 + 0.45 * 0.5 = 0.665$$

Substituting in Eq.6

$$P(AMC \mid bil) = \frac{0.5 * 0.45}{0.665} = 0.34$$

2. **Bag of words and single topic model** [10 points]

Bag of words (sac de mots) et modèle de sujet unique Consider the following distributions over words in the vocabulary given a particular topic:

	$\mathbb{P}(\text{word} \mid \text{topic} = \textit{sports})$	$\mathbb{P}(\text{word} \mid \text{topic} = \textit{politics})$
word = "*goal*"	1/100	7/1000
word = "*kick*"	1/200	3/1000
word = "*congress*"	0	1/50
word = "*vote*"	5/1000	1/100
word = *other*	980/1000	960/1000

Table 1:

we consider the event of a document be politics as Pol and sports as S.

(a) $P(goal \mid Pol) = \frac{7}{1000}$ According to the table.1

(b) $P(goal \mid S) = \frac{1}{100}$ so in a text of sports with 200 words it must appear two times.

(c) Using the total theorem

$$\begin{aligned} P(goal) &= P(S)P(goal \mid S) + P(Pol)P(goal \mid Pol)) \\ &= \frac{2}{3}\frac{1}{100} + \frac{1}{3}\frac{7}{1000} = \frac{9}{1000} \end{aligned} \tag{7}$$

(d) Using the Bayes theorem

$$P(S \mid kick) = \frac{P(kick \mid S)P(S)}{P(kick)} \tag{8}$$

where the probability of having the word "kick" in all the documents is

$$\begin{aligned} P(kick) &= P(S)P(kick \mid S) + P(Pol)P(kick \mid Pol) \\ &= \frac{2}{3}\frac{1}{200} + \frac{1}{3}\frac{3}{1000} = \frac{13}{3000} \end{aligned} \tag{9}$$

Substituting in Eq.8

$$P(S \mid kick) = \frac{\frac{1}{200}\frac{2}{3}}{\frac{13}{3000}} = \frac{10}{13}$$

(e) Using the total theorem, we are seeking for

$$P(kick\,after\,goal) = P(S)P(kick\,after\,goal \mid S)+P(Pol)P(kick\,after\,goal \mid Pol)$$

as the words are independents they belong from different leaves of the decision tree so

$$\begin{aligned} P(kick\,after\,goal) =& P(S)[P(kick \mid S) + P(goal \mid S)]+ \\ & P(Pol)[P(kick \mid Pol) + P(goal \mid Pol)] = \\ & \frac{2}{3}\left[\frac{1}{200} + \frac{1}{100}\right] + \frac{1}{3}\left[\frac{3}{1000} + \frac{7}{1000}\right] = \frac{1}{75} \end{aligned} \tag{10}$$

(f) We pick up the N documents and we count how many documents are labeled with politics (Pol) and sports (S), in this way we can obtain

$$\begin{aligned} P(S) &= \frac{\#documents\,label\,S}{N} \\ P(Pol) &= \frac{\#documents\,label\,Pol}{N} \end{aligned} \tag{11}$$

After in a specific document (D) (sports or politics), we count the number of words in the document and the number of occurrences of a specif word w, so

$$P(w \mid D) = \frac{\#occurences\,of\,w}{\#words\,in\,D}$$

3. **Maximum likelihood estimation** [5 points]

Estimateur du maximum de vraisemblance

Let $x \in \mathbb{R}$ be uniformly distributed in the interval $[0, \theta]$ where θ is a parameter. That is, the pdf of x is given by

$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose that n samples $D = \{x_1, \ldots, x_n\}$ are drawn independently according to $f_\theta(x)$.

(a) Let $f_\theta(x_1, \ldots, x_n)$, as the variables x_n are independents we have

$$f_\theta(x_1, \ldots, x_n) = P(x \mid \theta) = f_\theta(x_1) \cdot f_\theta(x_1) \cdots f_\theta(x_n) \tag{12}$$

(b) the likelihood is defined as $l(\theta) = \prod_{k=1}^{n} P(x_k \mid \theta)$. Using the definition of the indicator function I

$$I(x) = \begin{cases} 1 & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

the likelihood of the *pdf* can be written as

$$\begin{aligned} l(\theta) = & \prod_{k=1}^{n} P(x_k \mid \theta) \\ & \frac{1}{\theta^n} I\left(\max_k x_k \leq \theta\right) I\left(\min_k x_k \geq 0\right) \end{aligned} \tag{13}$$

In the last equation θ^n decays monotonically with n so for seeking the maximum likelihood we must see the behavior of the indicator function,. $I\left(\max_k x_k \leq \theta\right)$ is null if $\max\{x_k\} > \theta$ so the maximum value of the function happens when

$$\theta = \max_k \{x_k\}$$

4. **Maximum likelihood estimation 2** [10 points]

Estimateur de maximum de vraisemblance 2

Consider the following probability density function:

$$f_\theta(x) = 2\theta x e^{-\theta x^2}$$

where θ is a parameter and x is positive real number.
The likelihood of the function is done by

$$\begin{aligned} l_\theta(x) = \sum_{i=1}^{n} \log[f(x_i \mid \theta)] &= \sum_{i=1}^{n} \log\left(2\theta x_i e^{-\theta x_i^2}\right) \\ &= \sum_{i=1}^{n} \log(2\theta x_i) - \theta x_i^2 \end{aligned} \tag{14}$$

so the derivative is done by

$$l'_\theta(x) = \frac{1}{\theta} - \sum_{i=1}^{n} x_i^2$$

and the function is maximized when

$$\sum_{i=1}^{n} x_i^2 = \frac{1}{\theta}$$

5. **k-nearest neighbors** [10 points]

k plus proches voisins

Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n independent labelled samples drawn using the following sampling process:

- the label of each $\mathbf{x}_i$ is drawn randomly with 50% probability for each of the two classes
- x_i is drawn uniformly in S^+ if its label is positive, and uniformly in S^- otherwise

Where S^+ and S^- are two **unit** hyperspheres whose centers are 10 units apart.

(a) By definition

$$P(j \; nearest \; neighbors \; in \; S^+ \mid S^+) = \frac{\#samples \; in \; S^+}{cardinality \; of \; \Omega} \tag{15}$$

$$P(j \; nearest \; neighbors \; in \; S^- \mid S^-) = \frac{\#samples \; in \; S^-}{cardinality \; of \; \Omega} \tag{16}$$

where Ω is our sample space.
We have a sample space with n elements, we are looking for how many $j-$element different subsets we can obtain from our sample space, which are organized inside the spheres S^+, S^-. By definition the number of j-element subsets of a given n-element set is just the combination

$$\binom{n}{j} = \#samples \; in \; S^+ (or \; S^-)$$

The cardinality of our sample space Ω is just the total number of subsets we can build in a set with n elements which is 2^n, so replacing in Eq.15, Eq.16

$$P(j \; nearest \; neighbors \; in \; S^+ \mid S^+) = \frac{\binom{n}{j}}{2^n} \tag{17}$$

$$P(j \; nearest \; neighbors \; in \; S^- \mid S^-) = \frac{\binom{n}{j}}{2^n} \tag{18}$$

hence

$$P(error) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \tag{19}$$

(b) if $k = 1$, $P(error) = \frac{1}{2^n}$, when $k > 1$,

$$P(error) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \frac{n!}{j!(n-j)!} \tag{20}$$

we define

$$\sum_{j=0}^{(k-1)/2} \frac{n!}{j!(n-j)!} \equiv m$$

where $m > 0$ so $P(error) \mid_{k>1} = \frac{m}{2^n}$, in this manner we can induce that

$$P(error) \mid_{k=1} \leq P(error) \mid_{k>1}$$

(c) $\forall j \geq 0$ and $j \leq \frac{k-1}{2}$ we can say that

$$\binom{n}{j} \leq \binom{n}{\frac{k-1}{2}}$$

so

$$\frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \leq \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{\frac{k-1}{2}} \tag{21}$$

but

$$\sum_{j=0}^{(k-1)/2} \binom{n}{j} = \frac{k+1}{2} \binom{n}{\frac{k-1}{2}} \tag{22}$$

substituting in Eq.21, we arrive to

$$\frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \leq \frac{1}{2^n} \frac{k+1}{2} \binom{n}{\frac{k-1}{2}} \tag{23}$$

Using the definition of binomial combination

$$\binom{n}{\frac{k-1}{2}} = \frac{n!}{\left(\frac{k-1}{2}\right)! \left(n - \frac{k-1}{2}\right)!} \tag{24}$$

Replacing in Eq.23

$$\frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \leq \frac{1}{2^n} \frac{k+1}{2} \frac{n!}{\left(\frac{k-1}{2}\right)! \left(n - \frac{k-1}{2}\right)!} \tag{25}$$

if k is a big number it follows that

$$\frac{k+1}{2}\frac{1}{\left(\frac{k-1}{2}\right)!} \leq 1 \; and \; n - \frac{k-1}{2} \to n-k \tag{26}$$

Replacing in Eq.25

$$\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} \leq \frac{1}{2^n}\frac{n!}{(n-k)!} = \frac{1}{2^n}n^k \tag{27}$$

so we found a bound for the sum as

$$\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} \leq \frac{1}{2^n}n^{a\sqrt{n}} = \left(\frac{n^a}{2\sqrt{n}}\right)^{\sqrt{n}} \tag{28}$$

but $\lim_{n\to\infty}\frac{n^a}{2\sqrt{n}} \to 0$ so we demonstrate that

$$\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} \to 0 \tag{29}$$

when $k \leq a\sqrt{n}$ and $n \to \infty$

6. **Gaussian Mixture** [10 points] Mélange de Gaussiennes

Let $\mu_1, \mu_2 \in \mathbb{R}^2$, and let Σ_1, Σ_2 be two 2x2 positive definite matrices (i.e. symmetric with positive eigenvalues).
We now introduce the two following pdf over $\mathbb{R}^2$:

$$f_{\mu_1,\Sigma_1}(\mathbf{x}) = \frac{1}{2\pi\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}$$

$$f_{\mu_2,\Sigma_2}(\mathbf{x}) = \frac{1}{2\pi\sqrt{det(\Sigma_2)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2)}$$

These pdf correspond to the multivariate Gaussian distribution of mean μ_1 and covariance Σ_1, denoted $\mathcal{N}_1(\mu_1, \Sigma_1)$, and the multivariate Gaussian distribution of mean μ_2 and covariance Σ_2, denoted $\mathcal{N}_2(\mu_2, \Sigma_2)$.
Using the bayes rule for a discrete distribution of Y and continuous distribution for X

$$\begin{aligned} P(Y=0 \mid X=x) =& \frac{P(Y=0)P(X=x \mid Y=0)}{P(X=x)} \\ =& \frac{P(Y=0)P(X=x \mid Y=0)}{P(Y=0)P(X=x \mid Y=0) + P(Y=1)P(X=x \mid Y=1)} \end{aligned} \tag{30}$$

but $P(Y=0) = P(Y=1) = 0.5$, therefore

$$\begin{aligned} P(Y=0 \mid X=x) &= \frac{0.5 f_{\mu_2,\sigma_2}(x)}{0.5 f_{\mu_1,\sigma_1}(x) + 0.5 f_{\mu_2,\sigma_2}(x)} \\ &= \frac{f_{\mu_2,\sigma_2}(x)}{f_{\mu_1,\sigma_1}(x) + f_{\mu_2,\sigma_2}(x)} \end{aligned} \tag{31}$$

5 Practical homework

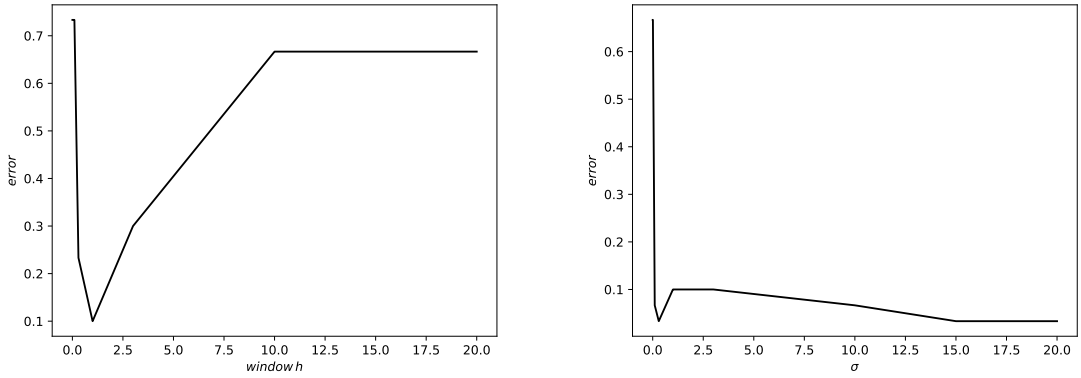

Figure 1: Classification error on the validation set of iris: Left: Hard parzen window. Right: RBF Parzen's

In Fig.1 (left) we can see the behavior of the classification error on the validation set in the case of hard parzen window. For very small windows the error is big given that we are considering a set of neighbors with very few elements. Additionally, for very small windows there is a great risk of not having neighbors close to the test point and the classification is given by a random function increasing the error. As h varies, we find the optimal size of the window (where the error has the minimum value) later the error grows again because our window is very large, allowing that very distant points to our test set have a vote with the same weight as the close neighbors.
In Fig.1 (right) we note the classification error for soft parzen windows. In this case the weight (vote) of each neighboring point is modeled using a Gaussian function with standard variation σ. For very small values of σ the distribution would be very sharp increasing the error, for values of σ above the optimal value, the Gaussian functions are very open allowing distant neighbors to have a considerable weight in the decision of the classification of the test point.

7 Practical homework

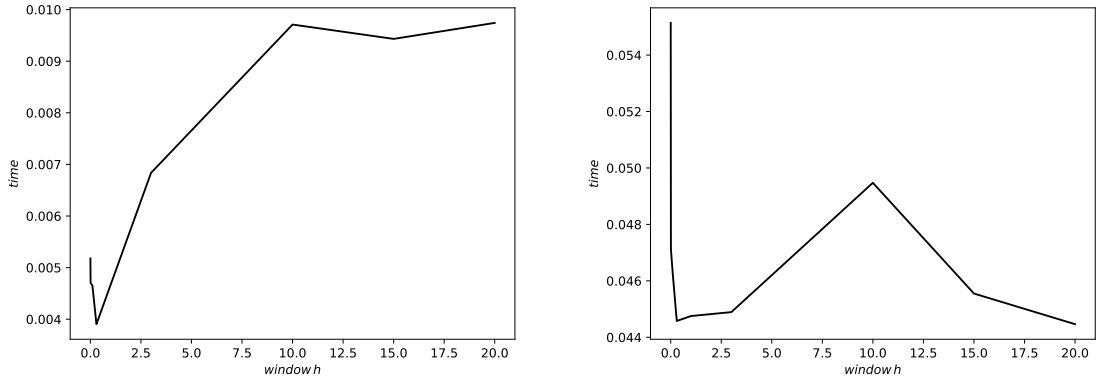

Figure 2: Running time complexity : Left: Hard parzen window. Right: RBF Parzen's

In Fig.2 we see the behavior of the running time complexity. This time is calculated from the computational time taken in the calculation of **compute predictions** in the algorithm. In the case of hard parzen windows, the algorithm evaluates whether a training point is within the window around the test point, subsequently it determines the classification of the test point according to the vote of each of the points within the window. Clearly, the computational time should be expected to increase as the window is longer because more training points are voting for the ranking. In the case of soft parzen windows we see less abrupt variations in the computational time, since in the algorithm of the RBZ parzens all the training points vote, what changes is the weight of their vote according to their distance to the testing point.

9 Practical homework

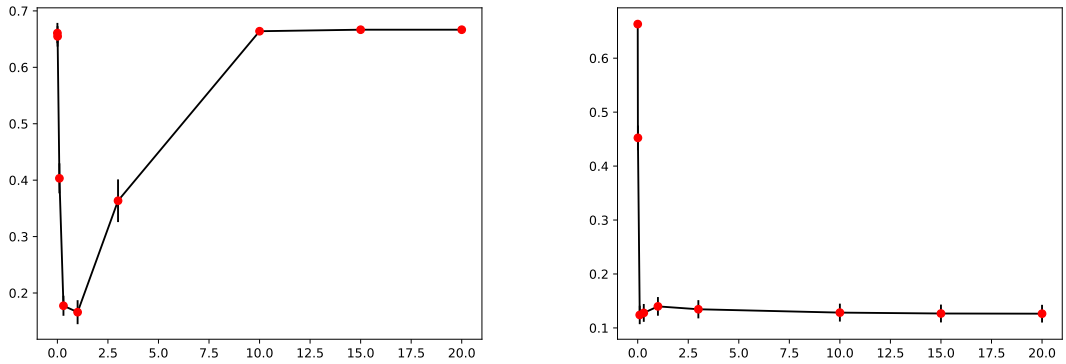

Figure 3: Classification error on the validation set of a projected training set: Left: Hard parzen window. Right: RBF Parzen's

In Fig.3 the red dots and the vertical lines correspond to the average value $\overline{\mu}$ and the standard deviation σ (actually 0.2σ) of 500 training points projected in a small space. Originally, the data set contained four features and it was projected in a reduced space with only two features. The quality of the projection logically depends on the projected space, whose elements were chosen randomly from a Gaussian distribution of mean $\mu = 0$ and standard deviation $\sigma = 1$, this distribution guarantees that the vectors of the projection matrix are almost orthogonal allowing the results in the reduced space not to be very different to those obtained in the complete data space. As we can see in Fig.1 and Fig.3, the overall behavior of the classification error is quite similar, although the mean minimum error in the reduced space is slightly greater than the minimum error of the entire space due to the projection.