IFT-6390 Fundamentals of Machine Learning
Professor: Ioannis Mitliagkas

Homework 1 - Theoretical part

Devoir 1 - Partie Théorique

- This homework must be done and submitted to Gradescope individually. You are welcome to discuss with other students but the solution you submit must be your own. Note that we will use Gradescope's plagiarism detection feature. All suspected cases of plagiarism will be recorded and shared with university officials for further handling.
  Ce devoir doit être fait et envoyé sur Gradescope individuellement. Vous pouvez discuter avec d'autres étudiants mais les réponses que vous soumettez doivent être les vôtres. A noter que nous utiliserons l'outil de détection de plagiat de Gradescope. Tous les cas suspectés de plagiat seront enregistrés et transmis à l'Université pour vérification.

- You need to submit your solution as a pdf file on Gradescope using the homework titled `(6390: GRAD) Theoretical Homework 1`.
  Vous devez soumettre vos solutions au format pdf sur Gradescope en utilisant le devoir intitulé `(6390: GRAD) Theoretical Homework 1`.

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

   Rappels de probabilités: probabilité conditionnelle et règle de Bayes

   (a) Give the definition of the conditional probability of a discrete random variable X given a discrete random variable Y.
   Donnez la définition de la probabilité conditionnelle de la variable aléatoire discrète X sachant la variable aléatoire discrète Y

   (b) Consider a biased coin with probability 2/3 of landing on heads and 1/3 on tails. This coin is tossed three times. What is the probability that exactly two heads occur (out of the three tosses)

given that the first outcome was a head?
Soit une pièce déséquilibrée dont la probabilité d'obtenir face est 2/3 et la probabilité d'obtenir pile est 1/3. Cette pièce est lancée à trois reprises. Quelle est la probabilité d'obtenir exactement deux faces (parmis les trois lancers), sachant que le premier lancer a fait face ?

(c) Give two equivalent expressions of $P(X, Y)$:

(i) as a function of $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$

(ii) as a function of $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$

Donnez deux expressions équivalentes de $P(X, Y)$:

(i) en fonction de $\mathbb{P}(X)$ et $\mathbb{P}(Y|X)$

(ii) en fonction de $\mathbb{P}(Y)$ et $\mathbb{P}(X|Y)$

(d) Prove Bayes theorem:
Prouvez le théorème de Bayes:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

(e) A survey of certain Montreal students is done, where 55% of the surveyed students are affiliated with UdeM while the others are affiliated with McGill. A student is drawn randomly from this surveyed group.
Un sondage des étudiants Montréalais est fait, où 55% des élèves sondés sont affiliés à l'UdeM alors que les autres sont affiliés à McGill. Un étudiant est choisi aléatoirement parmis ce groupe.

i. What is the probability that the student is affiliated with McGill?
Quelle est la probabilité que l'étudiant soit affilié à McGill?

ii. Now let's say that this student is bilingual, and you know that 80% of UdeM students are bilingual while 50% of McGill students are. Given this information, what is the probability that this student is affiliated with McGill ?
Considérons maintenant que l'étudiant est bilingue, et que 80% des étudiants de l'UdeM sont bilingues alors que seulement 50% des étudiants de McGill le sont. Étant donné cette information, quelle est la probabilité que cet étudiant soit affilé à McGill ?

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

Bag of words (sac de mots) et modèle de sujet unique

We consider a classification problem where we want to predict the topic of a document from a given corpus (collection of documents). The topic of each document can either be *sports* or *politics*. 2/3 of the documents in the corpus are about *sports* and 1/3 are about *politics*.

We will use a very simple model where we ignore the order of the words appearing in a document and we assume that words in a document are independent from one another given the topic of the document.

On s'intéresse à un problème de classification où l'on veut prédire le sujet d'un document d'un certain corpus (ensemble de documents). Le sujet de chaque document peut être soit *sport*, soit *politique*. 2/3 des documents du corpus sont sur le *sport*, et 1/3 sont sur la *politique*.

On va utiliser un modèle très simple où on ignore l'ordre des mots apparaissant dans le document et l'on considère que les mots dans un document sont indépendants les uns des autres, étant donné le sujet du document.

In addition, we will use very simple statistics of each document as features: the probabilities that a word chosen randomly in the document is either "goal", "kick", "congress", "vote", or any another word (denoted by *other*). We will call these five categories the vocabulary or dictionary for the documents: $V = \{"goal", "kick", "congress", "vote", other\}$.

De plus, nous allons utiliser des statistiques très simples des documents: les probabilités qu'un mot choisi au hasard dans un document soit "goal", "kick", "congress", "vote", ou n'importe quel autre mot (dénoté par *other*). Nous appelons ces cinq catégories le vocabulaire ou dictionnaire pour les documents: $V = \{"goal", "kick", "congress", "vote", other\}$.

Consider the following distributions over words in the vocabulary given a particular topic:

Soit les distributions suivantes des mots du vocabulaire, par sujet:

This table tells us for example that the probability that a word chosen at random in a document is "vote" is only 5/1000 if the topic of the document is *sport*, but it is 1/100 if the topic is *politics*.

	$\mathbb{P}$(word | topic = *sports*)	$\mathbb{P}$(word | topic = *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:

Cette table nous dit par exemple que la probabilité qu'un mot choisi aléatoirement dans un document soit "vote" n'est que de 5/1000 si le sujet du document est le *sport*, mais est de 1/100 si le sujet est la *politique*.

(a) What is the probability that a random word in a document is "goal" given that the topic is *politics*?
Quelle est la probabilité qu'un mot aléatoire dans un document soit "goal" étant donné que le sujet est la *politique* ?

(b) In expectation, how many times will the word "goal" appear in a document containing 200 words whose topic is *sports*?
Quelle est l'espérance du nombre de fois où le mot "goal" apparait dans un document de 200 mots dont le sujet est le *sport*?

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "goal"?
On tire aléatoirement un document du corpus. Quelle est la probabilité qu'un mot aléatoire de ce document soit "goal"?

(d) Suppose that we draw a random word from a document and this word is "kick". What is the probability that the topic of the document is *sports*?
Supposons que l'on tire aléatoirement un mot d'un document et que ce mot est "kick". Quelle est la probabilité que le sujet du document soit le *sport*?

(e) Suppose that we randomly draw two words from a document and the first one is "kick". What is the probability that the second word is "goal"?
Supposons que l'on tire aléatoirement deux mots d'un document et que le premier soit "kick". Quelle est la probabilité que le second mot soit "goal"?

(f) Going back to learning, suppose that you do not know the conditional probabilities given a topic or the probability of each topic (i.e. you don't have access to the information in table 1 or the topic distribution), but you have a dataset of N documents where each document is labeled with one of the topics *sports* and *politics*. How would you estimate the conditional probabilities (e.g., $\mathbb{P}(\text{word} = \text{"}goal\text{"} \mid \text{topic} = politics)$) and topic probabilities (e.g., $\mathbb{P}(\text{topic} = politics)$) from this dataset?
Pour en revenir à l'apprentissage, supposons que l'on ne connaisse pas les probabilités conditionnelles étant donné chaque sujet ni les probabilités de chaque sujet (i.e. nous n'avons pas accès aux informations de la table 1 où aux proportions de chaque sujet), mais nous avons un jeu de données de N documents où chaque document est annoté avec un des sujet *sport* ou *politique*. Comment estimeriez vous les probabilitiés conditionelles (e.g., $\mathbb{P}(\text{mot} = \text{"}goal\text{"} \mid \text{sujet} = politique)$) et les probabilités des sujets (e.g., $\mathbb{P}(\text{sujet} = politique)$) à partir de ce jeu de données ?

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

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

Soit $x \in \mathbb{R}$ distribué uniformément dans l'interval $[0, \theta]$ où θ est un paramètre. C'est à dire que la fonction de densité de probabilité de x est donnée par :

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

Suppose that n samples $D = \{x_1, \ldots, x_n\}$ are drawn independently according to $f_\theta(x)$.
Supposons que n points $D = \{x_1, \ldots, x_n\}$ sont tirés aléatoirement indépendemment selon $f_\theta(x)$.

(a) Let $f_\theta(x_1, x_2, \ldots, x_n)$ denote the joint pdf of n independent and identically distributed (i.i.d.) samples drawn according to $f_\theta(x)$.

Express $f_\theta(x_1, x_2, \ldots, x_n)$ as a function of $f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)$
Soit $f_\theta(x_1, x_2, \ldots, x_n)$ la fonction de densité de probabilité jointe de n points indépendemment et identiquement distribué (i.i.d) selon $f_\theta(x)$. Exprimez $f_\theta(x_1, x_2, \ldots, x_n)$ en fonction de $f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)$

(b) We define the maximum likelihood estimate by the value of θ which maximizes the likelihood of having generated the dataset D from the distribution $f_\theta(x)$. Formally,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n),$$

Show that the maximum likelihood estimate of θ is $max(x_1, \ldots, x_n)$

On définie l'estimateur du maximum de vraisemblance comme la valeur de θ qui maximise la vraisemblance de générer le jeu de donnée D à partir de la distribution $f_\theta(x)$. Formellement,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n),$$

Montrez que l'estimateur du maximum de vraisemblance de θ est $max(x_1, \ldots, x_n)$.

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.
Soit la fonction de densité de probabilité suivante:

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

où θ est un paramètre et x est un nombre réel positif.

Using the same notation as in exercise 3, compute the maximum likelihood estimate of θ.
(hint: you may simplify computations by proving that the maximizer of $f_\theta(x_1, x_2, \ldots, x_n)$ is also the maximizer of $log[f_\theta(x_1, x_2, \ldots, x_n)]$)
En utilisant les mêmes notations que dans dans l'exercice 3, calculez l'estimateur du maximum de vraisemblance de θ.
(indice: vous pouvez simplifier les calculs en prouvant que le θ maximisant $f_\theta(x_1, x_2, \ldots, x_n)$ corresponds aussi au maximum de $log[f_\theta(x_1, x_2, \ldots, x_n)]$)

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.
Soit $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ un ensemble de n points labelés indépendants, tirés aléatoirement suivant la procédure suivante:

- le label de chaque $\mathbf{x}_i$ est tiré aléatoirement avec une probabilité de 50% pour chacune des deux classes
- x_i est tiré uniformément dans S^+ si son label est positif, et uniformément dans S^- sinon

Où S^+ et S^- sont deux hypersphères **unitaires** dont les centres sont espacés de 10 unités.

(a) Show that if k is odd the average probability of error of the k-NN classifier is given by

$$P_n(e) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j}.$$

Montrez que si k est impair, la probabilité d'erreur moyenne du classifieur des k plus proches voisins est donné par

$$P_n(e) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j}.$$

(b) Show that in this case the single-nearest neighbor classifier ($k = 1$) has a lower error rate than the k-NN classifier for $k > 1$.
Montrez que dans ce cas le classifieur du plus proche voisin ($k = 1$) a un plus faible taux d'erreur que le classifieur des k plus proches voisins pour $k > 1$.

(c) If k is allowed to increase with n but is restricted by $k \leq a\sqrt{n}$ (for some constant a), show that $P_n(e) \to 0$ as $n \to \infty$.
Si k peut augmenter avec n mais est limité par $k \leq a\sqrt{n}$ (pour une constante a), montrez que $P_n(e) \to 0$ lorsque $n \to \infty$.

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 $\mathbb{R}^2$:
Soit $\mu_1, \mu_2 \in \mathbb{R}^2$, et soit Σ_1, Σ_2 deux matrices 2x2 positives définies (i.e. symmétriques avec des valeurs propres strictement positives).
On définie maintenant les deux fonctions de densités de probabilités suivantes sur $\mathbb{R}^2$:

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

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

These pdf correspond to the multivariate Gaussian distribution of mean μ_1 and covariance Σ_1, denoted $\mathcal{N}_2(\mu_1, \Sigma_1)$, and the multivariate Gaussian distribution of mean μ_2 and covariance Σ_2, denoted $\mathcal{N}_2(\mu_2, \Sigma_2)$.
Ces fonctions de densité de probabilités correspondent à la distribution gaussienne multivariée de centre μ_1 et covariance Σ_1, notée $\mathcal{N}_2(\mu_1, \Sigma_1)$, et à la gaussienne multivariée de centre μ_2 et covariance Σ_2, notée $\mathcal{N}_2(\mu_2, \Sigma_2)$.

We now toss a balanced coin Y, and draw a random variable X in $\mathbb{R}^2$, following this process : if the coin lands on tails ($Y = 0$) we draw X from $\mathcal{N}_2(\mu_1, \Sigma_1)$, and if the coin lands on heads ($Y = 1$) we draw X from $\mathcal{N}_2(\mu_2, \Sigma_2)$.
On lance maintenant une pièce équilibrée Y, et on tire une variable aléatoire X dans $\mathbb{R}^2$, en suivant le procédé suivant : si la pièce atterit sur pile ($Y = 0$), on tire X selon $\mathcal{N}_2(\mu_1, \Sigma_1)$, et si la pièce atterit sur face ($Y = 1$), on tire X selon $\mathcal{N}_2(\mu_2, \Sigma_2)$.

Calculate $\mathbb{P}(Y = 0|X = \mathbf{x})$, the probability that the coin landed on tails given $X = \mathbf{x} \in \mathbb{R}^2$, as a function of μ_1, μ_2, Σ_1, Σ_2, and $\mathbf{x}$. Show all the steps of the derivation.

Calculez $\mathbb{P}(Y = 0|X = \mathbf{x})$, la probabilité que la pièce atterisse sur pile sachant $X = \mathbf{x} \in \mathbb{R}^2$, en fonction de μ_1, μ_2, Σ_1, Σ_2, et $\mathbf{x}$. Montrez toutes les étapes du calcul.