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

   (a) Give the definition of the conditional probability of a discrete random variable X given a discrete random variable 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?

   (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 $\mathbb{P}(X|Y)$

   (d) Prove Bayes theorem:
   $$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

   (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.

      i. What is the probability that the student is affiliated with 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 ?

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

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: $V = \{"goal", "kick", "congress", "vote", other\}$.

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

	$\mathbb{P}(\text{word} \mid \text{topic} = sports)$	$\mathbb{P}(\text{word} \mid \text{topic} = politics)$
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:

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*.

(a) What is the probability that a random word in a document is "goal" given that the topic is *politics*?

(b) In expectation, how many times will the word "goal" appear in a document containing 200 words whose topic is *sports*?

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "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*?

(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"?

(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?

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

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

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

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

(a) Let $f_\theta(x_1, x_2, \ldots, x_n)$ 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 $f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)$

(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 $max(x_1, \ldots, x_n)$

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

   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.
   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$[f_\theta(x_1, x_2, \ldots, x_n)]$)

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

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

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

   (b) Show that in this case the single-nearest neighbor classifier ($k = 1$) has a lower error rate than the k-NN classifier for $k > 1$.

   (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 $n \to \infty$.

6. **Gaussian Mixture** [10 points]

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 $\mathcal{N}_2(\mu_1, \Sigma_1)$, and the multivariate Gaussian distribution of mean μ_2 and covariance Σ_2, denoted $\mathcal{N}_2(\mu_2, \Sigma_2)$.

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 $\mathcal{N}_2(\mu_2, \Sigma_2)$.

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.