

## Homework 1 - Theoretical part (Matteo Esposito)

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 ?

(a)

$$P(X = x|Y = y) = \frac{P(X = x) \cap P(Y = y)}{P(Y = y)}$$

(b) Let  $T_n$  be the result of the nth toss.

$$\begin{aligned} P((T_1 = H, T_2 = T) \text{ or } (T_1 = T, T_2 = H) | T_0 = H) &= \frac{P(HHT) + P(HTH)}{P(H)} \\ &= \frac{2 * (1/3)^1 * (2/3)^2}{(2/3)} \\ &= \boxed{4/9} \end{aligned}$$

(c) (i)

$$P(X, Y) = P(Y, X) = P(Y|X)P(X)$$

(ii)

$$P(X, Y) = P(X|Y)P(Y)$$

(d) Since  $P(Y, X) = P(Y|X)P(X)$  and  $P(X, Y) = P(X|Y)P(Y)$  and  $P(X, Y) = P(Y, X)$  then,

$$\begin{aligned} P(X|Y)P(Y) &= P(Y|X)P(X) \\ P(X|Y) &= \frac{P(Y|X)P(X)}{P(Y)} \quad \square \end{aligned}$$

(e) (i)

$$P(McGill) = 1 - P(UdeM) = \boxed{0.45}$$

(ii)

$$\begin{aligned} P(McGill|bilingual) &= \frac{P(McGill \cap bilingual)}{P(bilingual)} \\ &= \frac{P(McGill \cap bilingual)P(McGill)}{P(bilingual)} \\ &= \frac{P(McGill \cap bilingual)P(McGill)}{P(McGill \cap bilingual)P(McGill) + P(UdeM \cap bilingual)P(UdeM)} \\ &= \frac{(0.50)(0.45)}{(0.50)(0.45) + (0.80)(0.55)} \\ &= \boxed{45/133} \end{aligned}$$

## 2. Bag of words and single topic model [12 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:  $V = \{\text{"goal"}, \text{"kick"}, \text{"congress"}, \text{"vote"}, \text{other}\}$ .

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"	3/200	8/1000
word = "kick"	1/200	2/1000
word = "congress"	0	1/50
word = "vote"	5/1000	2/100
word = <i>other</i>	960/1000	950/1000

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 2/100 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.,  $\mathbb{P}(\text{word} = \text{"goal"} \mid \text{topic} = \text{"politics"})$ ) and topic probabilities (e.g.,  $\mathbb{P}(\text{topic} = \text{"politics"})$ ) from this dataset?

(a)

$$P(\text{"goal"}|\text{politics}) = \frac{8}{1000}$$

(b)

$$200 * P(\text{"goal"}|\text{sports}) = 200 \left( \frac{3}{200} \right) = 3$$

(c)

$$\begin{aligned} P(\text{"goal"}) &= P(\text{"goal"}|\text{sports})P(\text{sports}) + P(\text{"goal"}|\text{politics})P(\text{politics}) \\ &= \left( \frac{3}{200} \right) \left( \frac{2}{3} \right) + \left( \frac{8}{1000} \right) \left( \frac{1}{3} \right) \\ &= 0.0127 \end{aligned}$$

(d)

$$\begin{aligned} P(\text{sports}|\text{"kick"}) &= \frac{P(\text{"kick"}|\text{sports})P(\text{sports})}{P(\text{"kick"}|\text{sports})P(\text{sports}) + P(\text{"kick"}|\text{politics})P(\text{politics})} \\ &= \frac{\frac{1}{200} * \frac{2}{3}}{\frac{1}{200} * \frac{2}{3} + \frac{2}{1000} * \frac{1}{3}} \\ &= \frac{5}{6} \end{aligned}$$

- (e) Seeing as we calculated  $P(sports|"kick")$  in the previous question, and knowing that the first word selected was "kick", we have that;

$$\begin{aligned} P("goal") &= \left(\frac{5}{6}\right) \left(\frac{3}{200}\right) + \left(\frac{1}{6}\right) \left(\frac{8}{1000}\right) \\ &= 0.0138 \end{aligned}$$

- (f) For the conditional probabilities, we would take a count of all words in each of the 2 document classes then divide that by the total number of words found in all documents per document class. i.e.

$$P("goal"|politics) = \frac{\# \text{ instances of "goal" in all politics documents}}{\# \text{ words in all politics documents}}$$

For the topic probabilities, it would suffice to get a count of the number of documents classified as politics and sports and take the quotient of those totals and N (where N is total number of documents). i.e.

$$P("politics") = \frac{\# \text{ labels} = \text{politics in dataset}}{N}$$

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

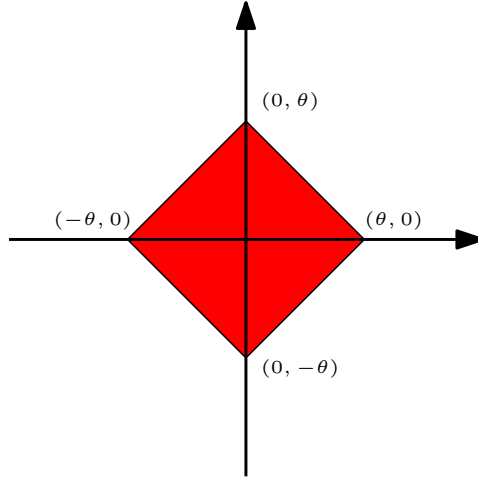
Let  $\mathbf{x} \in \mathbb{R}^2$  be uniformly distributed over a diamond area with diagonals  $2\theta$  where  $\theta$  is a parameter as shown in the figure. That is, the pdf of  $\mathbf{x}$  is given by

$$f_{\theta}(\mathbf{x}) = \begin{cases} 1/2\theta^2 & \text{if } \|\mathbf{x}\|_1 \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where  $\|\mathbf{x}\|_1 = |x_1| + |x_2|$  is the L1 norm.

Suppose that  $n$  samples  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  are drawn independently according to  $f_{\theta}(\mathbf{x})$ .

- (a) Let  $f_{\theta}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$  denote the joint pdf of  $n$  independent and identically distributed (i.i.d.) samples drawn according to  $f_{\theta}(\mathbf{x})$ . Express  $f_{\theta}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$  as a function of  $f_{\theta}(\mathbf{x}_1), f_{\theta}(\mathbf{x}_2), \dots, f_{\theta}(\mathbf{x}_n)$



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

$$\theta_{MLE} = \arg \max_{\theta \in \mathbb{R}^+} f_\theta(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n),$$

Find the maximum likelihood estimate of  $\theta$ .

(a)

$$\begin{aligned} L(\theta) &= f_\theta(x_1, x_2, \dots, x_n) \\ &= \prod_{i=1}^n f_\theta(x_i) \\ &= \left(\frac{1}{2\theta^2}\right) \left(\frac{1}{2\theta^2}\right) \cdots \left(\frac{1}{2\theta^2}\right) \\ L(\theta) &= \begin{cases} \frac{1}{(2\theta^2)^n} & \|x_i\|_1 \leq \theta \quad i = (1, \dots, n) \\ 0 & \text{Otherwise} \end{cases} \end{aligned}$$

- (b) We can see that the value of  $\theta$  that will maximize  $L(\theta)$  will have to be the smallest value of  $\theta$  such that  $\theta \geq \|x_i\|_1$ . Therefore,

$$\hat{\theta} = \max(\|x_1\|_1, \|x_2\|_1, \dots, \|x_n\|_1)$$

#### 4. Maximum likelihood meets histograms [10 points]

Let  $X_1, X_2, \dots, X_n$  be  $n$  i.i.d data points drawn from a piece-wise constant probability density function over  $N$  equal size bins between 0 and 1 ( $B_1, B_2, \dots, B_N$ ), where the constants are  $\theta_1, \theta_2, \dots, \theta_N$ .

$$p(x; \theta_1, \dots, \theta_N) = \begin{cases} \theta_j & \frac{j-1}{N} \leq x < \frac{j}{N} \text{ for } j \in \{1, 2, \dots, N\} \\ 0 & \text{otherwise} \end{cases}$$

We define  $\mu_j$  for  $j \in \{1, 2, \dots, N\}$  as  $\mu_j := \sum_{i=1}^n \mathbb{1}(X_i \in B_j)$ .

- (a) Using the fact that the total area underneath a probability density function is 1, express  $\theta_N$  in terms of the other constants.
- (b) Write down the log-likelihood of the data in terms of  $\theta_1, \theta_2, \dots, \theta_{N-1}$  and  $\mu_1, \mu_2, \dots, \mu_{N-1}$ .
- (c) Find the maximum likelihood estimate of  $\theta_j$  for  $j \in \{1, 2, \dots, N\}$ .

(a)

$$\begin{aligned} \sum_{j=1}^N \frac{\theta_j}{N} &= 1 \\ \sum_{j=1}^{N-1} \frac{\theta_j}{N} + \frac{\theta_N}{N} &= 1 \\ \theta_N &= N - \sum_{j=1}^{N-1} \theta_j \end{aligned}$$

(b)

$$\begin{aligned} \mathcal{L}(\theta) &= \log(L(\theta)) = \log \left( \prod_{j=1}^N \left( \frac{\theta_j}{N} \right)^{\mu_j} \right) = \log \left( \left( \frac{\theta_N}{N} \right)^{\mu_N} \prod_{j=1}^{N-1} \left( \frac{\theta_j}{N} \right)^{\mu_j} \right) \\ &= \log \left( \left( 1 - \sum_{j=1}^{N-1} \frac{\theta_j}{N} \right)^{n - \sum_{j=1}^{N-1} \mu_j} \prod_{j=1}^{N-1} \left( \frac{\theta_j}{N} \right)^{\mu_j} \right) \\ &= \left( n - \sum_{j=1}^{N-1} \mu_j \right) \log \left( 1 - \sum_{j=1}^{N-1} \frac{\theta_j}{N} \right) + \left( \sum_{j=1}^{N-1} \mu_j \log \left( \frac{\theta_j}{N} \right) \right) \end{aligned}$$

Staying aligned with the problem statement, the above expression should be our final answer. We can however simplify it some more.

$$\mathcal{L}(\theta) = \mu_N \log \left( \frac{\theta_N}{N} \right) + \sum_{j=1}^{N-1} \mu_j \log \left( \frac{\theta_j}{N} \right)$$

(c) Assume  $k \in \{1, \dots, N\}$ , then

$$\frac{\partial}{\partial \theta_k}(\mathcal{L}(\theta_k)) = -\frac{\mu_N}{1 - \sum_{j=1}^{N-1} \theta_j} + \frac{\mu_k}{\theta_k} = 0 \iff \frac{\mu_N}{1 - \sum_{j=1}^{N-1} \theta_j} = \frac{\mu_k}{\theta_k}$$

Rerranging our terms, we have that,

$$\theta_k = \frac{\mu_k \theta_N}{\mu_N}$$

To maximize this we assume  $\mu_N = n$  and  $\theta_N = N$ , in the case where all points fall into one bin and therefore,

$$\hat{\theta}_k = \frac{\mu_k N}{n} \quad \forall k \in \{1, \dots, N\}$$

## 5. Histogram methods [10 points]

Consider a dataset  $\{x_j\}_{j=1}^n$  where each point  $x \in [0, 1]^d$ . Let  $f(x)$  be the true unknown data distribution. You decide to use a histogram method to estimate the density  $f(x)$  and divide each dimension into  $m$  bins.

- Show that for a measurable set  $S$ ,  $\mathbb{E}_{x \sim f}[\mathbb{1}_{\{x \in S\}}] = \mathbb{P}_{x \sim f}(x \in S)$ , where  $\mathbb{1}_{\{x \in S\}} = 1$  if  $x \in S$  and 0 otherwise.
- Combining the result of the previous question with the Law of Large Numbers, show that the estimated probability of falling in bin  $i$ , as given by the histogram method, tends to  $\mathbb{P}_{x \sim f}(x \in V_i) = \int_{V_i} f(x) dx$ , the true probability of falling in bin  $i$ , as  $n \rightarrow \infty$ .  $V_i$  denotes the volume occupied by bin  $i$ .
- Consider the MNIST dataset with 784 dimensions (i.e.  $x \in [0, 1]^{784}$ ). We divide each dimension into 2 bins. How many digits (base 10) does the total number of bins have?
- Assuming a uniform distribution over all bins, how many data points would you need to get  $k$  points per bin on average?
- Assuming a uniform distribution over all bins, what is the probability that a particular bin is empty, as a function of  $d$ ,  $m$  and  $n$ ?
- We use the definition of expectation of a continuous rv and the fact that all points  $\notin S$  will not contribute to the expectation to yield the following:

$$\mathbb{E}_{x \sim f}[\mathbb{1}_{\{x \in S\}}] = \sum_{x \sim f} f(x) \mathbb{1}_{\{x \in S\}} dx = \sum_S f(x) dx = \mathbb{P}_{x \sim f}(x \in S)$$



- (b) Given a region  $V_i$ , we can leverage the LLN and the previous expression to assert that,

$$\mathbb{P}_{x \sim f}(x \in V_i) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{x \in V_i} f(x) = \mathbb{E}_{x \sim f}[\mathbb{1}_{\{x \in V_i\}}]$$

- (c)

$$\# \text{ bins} = 2^{784} = 10^x$$

where  $x$  is going to be the  $\#digits - 1$ . We have,

$$x = 784 * \log_{10} 2 \sim 236$$

Therefore, we have  $10^{236}$  bins and 237 digits.

- (d) If we assume a uniform distribution, to be able to get  $k$  points per bin on average, we would need to satisfy the following expression:

$$k = \frac{\# \text{ points}}{\# \text{ bins}} = \frac{n}{m}$$

Therefore, given our previous result, we would need

$$\frac{n}{2^{784}} = k \iff n = k * 2^{784} \text{ points}$$

- (e) The probability of a point falling into a bin is given by  $\frac{1}{m^d}$  therefore, the probability of not falling into a bin is  $1 - \frac{1}{m^d}$ . If we extend this to the case where a bin will never receive a point we have that,

$$P(\text{Empty bin}) = \left(1 - \frac{1}{m^d}\right)^n$$

## 6. Gaussian Mixture [10 points]

Let  $\mu_0, \mu_1 \in \mathbb{R}^d$ , and let  $\Sigma_0, \Sigma_1$  be two  $d \times d$  positive definite matrices (i.e. symmetric with positive eigenvalues).

We now introduce the two following pdf over  $\mathbb{R}^d$  :

$$f_{\mu_0, \Sigma_0}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\Sigma_0)}} e^{-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma_0^{-1} (\mathbf{x} - \mu_0)}$$

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

These pdf correspond to the multivariate Gaussian distribution of mean  $\mu_0$  and covariance  $\Sigma_0$ , denoted  $\mathcal{N}_d(\mu_0, \Sigma_0)$ , and the multivariate Gaussian distribution of mean  $\mu_1$  and covariance  $\Sigma_1$ , denoted  $\mathcal{N}_d(\mu_1, \Sigma_1)$ .

We now toss a balanced coin  $Y$ , and draw a random variable  $X$  in  $\mathbb{R}^d$ , following this process : if the coin lands on tails ( $Y = 0$ ) we draw  $X$  from  $\mathcal{N}_d(\mu_0, \Sigma_0)$ , and if the coin lands on heads ( $Y = 1$ ) we draw  $X$  from  $\mathcal{N}_d(\mu_1, \Sigma_1)$ .

- (a) Calculate  $\mathbb{P}(Y = 0|X = \mathbf{x})$ , the probability that the coin landed on tails given  $X = \mathbf{x} \in \mathbb{R}^d$ , as a function of  $\mu_0$ ,  $\mu_1$ ,  $\Sigma_0$ ,  $\Sigma_1$ , and  $\mathbf{x}$ . Show all the steps of the derivation.
- (b) Recall that the Bayes optimal classifier is  $h_{Bayes}(\mathbf{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \mathbb{P}(Y = y|X = \mathbf{x})$ . Show that in this setting if  $\Sigma_0 = \Sigma_1$  the Bayes optimal classifier is linear in  $\mathbf{x}$ .

(a) We can use Bayes Rule here,

$$\begin{aligned}
\mathbb{P}(Y = 0|X) &= \frac{f_{\mu_0, \Sigma_0}(\mathbf{x}) * P(Y = 0)}{f_{\mu_0, \Sigma_0}(\mathbf{x}) * P(Y = 0) + f_{\mu_1, \Sigma_1}(\mathbf{x}) * P(Y = 1)} \\
&= \frac{f_{\mu_0, \Sigma_0}(\mathbf{x}) * \frac{1}{2}}{f_{\mu_0, \Sigma_0}(\mathbf{x}) * \frac{1}{2} + f_{\mu_1, \Sigma_1}(\mathbf{x}) * \frac{1}{2}} \\
&= \frac{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\Sigma_0)}} e^{-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma_0^{-1}(\mathbf{x} - \mu_0)}}{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\Sigma_0)}} e^{-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma_0^{-1}(\mathbf{x} - \mu_0)} + \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1}(\mathbf{x} - \mu_1)}} \\
&= \frac{\sqrt{\det(\Sigma_1)} e^{-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma_0^{-1}(\mathbf{x} - \mu_0)}}{\sqrt{\det(\Sigma_1)} e^{-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma_0^{-1}(\mathbf{x} - \mu_0)} + \sqrt{\det(\Sigma_0)} e^{-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1}(\mathbf{x} - \mu_1)}}
\end{aligned}$$

(b) If we let  $\Sigma_0 = \Sigma_1 = \Sigma$ , we have,

$$\begin{aligned}
\mathbb{P}(X|Y = 1) &= \mathbb{P}(X|Y = 0) \\
\log(P(Y = 1)) - \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) &= \log(P(Y = 0)) - \frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)
\end{aligned}$$

Representing the constant values by  $c$ , we have,

$$\begin{aligned}
c + x^T \Sigma^{-1} x - 2\mu_1^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} \mu_1 &= x^T \Sigma^{-1} x - 2\mu_0^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0 \\
(2(\mu_0 - \mu_1)^T \Sigma^{-1}) x - (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1) &= c
\end{aligned}$$

Which can be represented alternatively as:

$$a^T x - b = 0 \quad \square$$

which is obviously linear.