**IFT 6390 Fundamentals of Machine Learning**
**Guillaume Lam**

# Homework 1 - Theoretical part

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

   Solution:

   (a) $P(X|Y) = \frac{P(X,Y)}{P(Y)}$

   (b) $P(H,H|T) = \frac{P(H,H,T)}{P(T)} = \frac{\frac{4}{12}}{\frac{1}{3}} = \frac{4}{9}$

   (c) i. $P(X,Y) = P(Y|X)P(X)$
   ii. $P(X,Y) = P(X|Y)P(Y)$

   (d)

$$\begin{aligned} P(X|Y) &\stackrel{?}{=} \frac{P(Y|X)P(X)}{P(Y)} \\ P(X|Y) &= \frac{P(X,Y)}{P(Y)} \\ &= \frac{P(Y|X)P(X)}{P(Y)} \end{aligned}$$

   (e) i. $P(MG) = 0.45$
   ii.

$$\begin{aligned} P(b|MG) &= \frac{P(b|MG)P(MG)}{P(b)} \\ &= \frac{P(b|MG)P(MG)}{\sum P(b|C_i)P(C_i)} \\ &= \frac{0.5 * 0.45}{0.5 * 0.45 + 0.8 * 0.55} \\ &= 0.338 \end{aligned}$$

2. **Bag of words and single topic model** [12 points]

Solution:

(a) $P(w = goal|t = pol) = \frac{8}{1000}$

(b) $\mathbb{E}[w = goal|t = pol]_{n=200} = 200 * \frac{3}{200} = 3$

(c)

$$\begin{aligned} P(goal) &= P(w = goal|t = sp)P(t = sp) + P(w = goal|t = pol)P(t = pol) \\ &= \frac{3}{200} * \frac{2}{3} + \frac{8}{1000} * \frac{1}{3} \\ &= \frac{76}{6000} \\ &= \frac{19}{1500} \end{aligned}$$

(d)

$$\begin{aligned} P(t = sp|w = kick) &= \frac{P(w = kick|t = sp)P(t = sp)}{P(w = kick)} \\ &= \frac{P(w = kick|t = sp)P(t = sp)}{P(w = kick|t = sp)P(t = sp) + P(w = kick|t = pol)P(t = pol)} \\ &= \frac{\frac{1}{300}}{\frac{1}{300} + \frac{2}{3000}} \\ &= \frac{5}{6} \end{aligned}$$

(e)

$$\begin{aligned} P(w = goal|w = kick) &= P(w = goal)P(w = kick) \\ &= P(w = goal|t = sp)P(t = sp) + P(w = goal|t = pol)P(t = pol) \\ &\quad + \\ &\quad P(w = kick|t = sp)P(t = sp) + P(w = kick|t = pol)P(t = pol) \\ &= \frac{19}{1500} * \frac{12}{3000} \\ &= \frac{19}{375000} \end{aligned}$$

(f) To estimate the conditional probabilities and topic probabilities, you simply need to count the words for some number of documents N to get an estimate for the probabilities of the whole

dataset and the larger N gets to the size of the dataset, the more the estimate will be true to the dataset.

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

Solution:

(a) Since the dataset is iid, by definition, we can say:

$$f_\theta(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

(b)

$$\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta \in \mathbb{R}^+} f_\theta(x_1, ..., x_n|\theta) \\
&= \arg\max_{\theta \in \mathbb{R}^+} \prod_{i=1}^{n} f_\theta(x_i|\theta) \\
&= \sum_{i=1}^{n} log(f_\theta(x_i|\theta)) \\
&= \prod_{i=1}^{n} \frac{1}{2\theta^2} \mathbb{1}(||x||_1 \leq \theta) \\
&= \frac{1}{2^n \theta^{2n}} \mathbb{1}(||x||_1 \leq \theta)
\end{aligned}$$

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

Solution:

(a)

$$\begin{aligned}
\sum_{i=1}^{N} p(x; \theta_i) &= 1 \\
\sum_{i=1}^{N} \theta_i &= 1 \\
\sum_{i=1}^{N-1} \theta_i + \theta_N &= 1 \\
\theta_N &= 1 - \sum_{i=1}^{N-1} \theta_i
\end{aligned}$$

(b)

$$\begin{aligned}
\ell(\theta) &= log(\prod_{i=1}^{N} \theta_i^{\mu_i}) \\
&= \sum_{i=1}^{N} \mu_i log(\theta_i) \\
&= \sum_{i=1}^{N-1} \mu_i log(\theta_i) + \mu_N log(\theta_N) \\
&= \sum_{i=1}^{N-1} \mu_i log(\theta_i) + (n - \sum_{i=1}^{N-1} \mu_i)(1 - \sum_{i=1}^{N-1} \theta_i)
\end{aligned}$$

(c)

$$\theta_{MLE} = \arg\max_{\theta} \ell(\theta_j)$$

By starting with the definition of MLE, we can easily see that we need to find the derivative of the log-likelihood function in order to find the maximum.

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}[\sum_{i=1}^{N} \mu_i log(\theta_i)] \\
&= \sum_{i=1}^{N} \frac{\mu_i}{\theta_i}
\end{aligned}$$

Now we can set the derivative to 0 in order to find the maximum. Then by isolating for $\theta_j$, we can find the MLE for $\theta$.

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta} &= \sum_{i=1}^{N} \frac{\mu_i}{\theta_i} = 0 \\
0 &= \sum_{i=1, i\neq j}^{N} \frac{\mu_i}{\theta_i} + \frac{\mu_j}{\theta_j} \\
\frac{\mu_j}{\theta_j} &= -\sum_{i=1, i\neq j}^{N} \frac{\mu_i}{\theta_i} \\
\theta_j &= \frac{-\mu_j}{\sum_{i=1, i\neq j}^{N} \frac{\mu_i}{\theta_i}} \\
\theta_{MLE} &= \frac{-\mu_j}{\sum_{i=1, i\neq j}^{N} \frac{\mu_i}{\theta_i}}
\end{aligned}$$

5. **Histogram methods** [10 points]

Solution:

(a)

$$\begin{aligned}\mathbb{E}_{x\sim f}[\mathbb{1}x \in S] &= \sum \mathbb{1}_S(x)\mathbb{P}_{x\sim f}(\mathbb{1}_S(x)) \\ &= 0 \times \mathbb{P}(x \notin S) + 1 \times \mathbb{P}(x \in S) \\ &= \mathbb{P}_{x\sim f}(x \in S)\end{aligned}$$

(b) Since we are looking for the estimated probability of falling in bin i, this simply corresponds to $\mathbb{E}_{x\sim f}[\mathbb{1}_{V_i}(x)]$.

$$\begin{aligned}\mathbb{E}_{x\sim f}[\mathbb{1}_{V_i}(x)] &= \int_{V_i} 0 \cdot P_{x\sim f}(x \notin V_i)dx + \int_{V_i} 1 \cdot P_{x\sim f}(x \in V_i)dx \\ &= \int_{V_i} f(x)dx = \mathbb{P}_{x\sim f}(x \in V_i)\end{aligned}$$

Thus, we can say that $\mathbb{E}_{x\sim f}[\mathbb{1}_{V_i}(x)] \to \mathbb{P}_{x\sim f}(x \in V_i)$ as $n \to \inf$ by Law of Large Numbers.

(c) Since each dimension is split into 2 bins and there are 784 dimensions, the total number of bins is simply $2 * 784 = 1568$. Thus, the total number of bins has 4 digits.

(d) Since the bins follow a uniform distribution, we can expect that if we were to drop n balls, assuming there are n bins, we would find a single ball in each bin. Thus, we simply need get a total number of points as such: #number of bins $\times\, k = nk$

(e) We first need to break up the domain over the two cases of the ratio between the size of the dataset and the total number of bins. As such, the probability that a particular bin is empty can be expressed as such:

$$P(x_i \notin B_j, \forall i \in \{1, ..., n\}) = \begin{cases} 1 & \text{if } md > n \text{ by Pigeonhole Principle} \\ (\frac{md-1}{md})^n & \text{otherwise} \end{cases}$$

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

Solution:

(a)

$$\begin{aligned} P(Y=0|X=x) &= \frac{P(X=x|Y=0)P(Y=0)}{P(X=x)} \\ &= \frac{f_{\mu_0,\Sigma_0}(X=x)\cdot\pi_0}{\sum \pi_k f_{\mu_k,\Sigma_k}(X=x)} \\ &= \frac{\pi_0 f_{\mu_0,\Sigma_0}(X=x)}{\pi_0 f_{\mu_0,\Sigma_0}(X=x)+\pi_1 f_{\mu_1,\Sigma_1}(X=x)} \end{aligned}$$

(b) The first part consists of finding the value of $\mathbb{P}(Y=y|X=x)$ in $h_{Bayes}(x) = \arg\max_{y\in\{0,1\}} \mathbb{P}(Y=y|X=x)$.

$$\mathbb{P}(Y=y|X=x) = \frac{\pi_y f_{\mu_y,\Sigma_0}(X=x)}{\pi_0 f_{\mu_0,\Sigma_0}(X=x)+\pi_1 f_{\mu_1,\Sigma_0}(X=x)}$$

$$= \frac{\pi_y \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(x-\mu_y)^T\Sigma_0^{-1}(x-\mu_y)}}{\pi_0 \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)} + \pi_1 \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma_0^{-1}(x-\mu_1)}}$$

By taking the log of the probability, we get the Bayes optimal classifier $\delta_y(X=x)$.

$$\begin{aligned} h_{Bayes}(x) &= \delta_y(X=x) \\ \delta_y(X=x) &= log(\mathbb{P}(Y=y|X=x)) \end{aligned}$$

However, since we are taking the argmax, the denominator of the probability and other constant values can be ignored as they are common to the

probability for all values of Y.

$$
\begin{aligned}
\delta_y(X=x) &= log(\mathbb{P}(Y=y|X=x)) \\
&= log(\mu_y \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(x-\mu_y)^T\Sigma_0^{-1}(x-\mu_y)}) + \cancel{log(denominator)} \\
&= log(\mu_y) - log((2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}) + log(e^{-\frac{1}{2}(x-\mu_y)^T\Sigma_0^{-1}(x-\mu_y)}) \\
&= log(\mu_y) - \frac{d}{2}log((2\pi) - \frac{1}{2}log(det(\Sigma_0))) - \frac{1}{2}(x-\mu_y)^T\Sigma_0^{-1}(x-\mu_y) \\
&= log(\mu_y) - \cancel{\frac{d}{2}log((2\pi)} - \cancel{\frac{1}{2}log(det(\Sigma_0)))} - \frac{1}{2}(x-\mu_y)^T\Sigma_0^{-1}(x-\mu_y) \\
&= log(\mu_y) - \frac{1}{2}(x-\mu_y)^T\Sigma_0^{-1}(x-\mu_y) \\
&= log(\mu_y) - \frac{1}{2}[x^T\Sigma_0^{-1}x - x^T\Sigma_0^{-1}\mu_y - \mu_y^T\Sigma_0^{-1}x + \mu_y^T\Sigma_0^{-1}\mu_y] \\
&= log(\mu_y) - \frac{1}{2}[\cancelto{0}{x^T\Sigma_0^{-1}x} - 2\mu_y^T\Sigma_0^{-1}x + \mu_y^T\Sigma_0^{-1}\mu_y] \\
&= log(\mu_y) + \mu_y^T\Sigma_0^{-1}x - \frac{1}{2}\mu_y^T\Sigma_0^{-1}\mu_y
\end{aligned}
$$

Thus, since x is to the power 1, the Bayes optimal classifier is linear in x.