Homework 1 - Theory

Reza Bayat

September 2021

1 Theoretical Part

1. (a) If X and Y are discrete random variables with joint PMF given by $P(X,Y)$, then the conditional probability mass function of X, given that Y, is denoted $P(X|Y)$ and given by:

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

(b) The Desired outputs are: $\{HTH, HHT\}$

$$P(H) = 1/4 \quad P(T) = 1/4$$

A = two of three outcomes be head.

$$P(A) = 2 \times (\frac{3}{4} \times \frac{1}{4} \times \frac{3}{4}) = \frac{9}{32}$$

(c)

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

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

(d) From above formulas, we can prove Bayes theorem:

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

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

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

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

(e) (i) The sum of all items of probability should be 1, $\frac{55}{100}$ of students are affiliated with UdeM, so the rest of the students which is $1 - \frac{55}{100} = \frac{45}{100}$ are affiliated with McGill.

(ii)
M = a student is affiliated with McGill.
B = a student is bilingual.
U = a student is affiliated with UdeM.

$$P(M|B) = \frac{P(M, B)}{P(B)}$$

Applying the Bayes role:

$$P(M|B) = \frac{P(B|M)P(M)}{P(B)}$$

$$P(B) = P(B|M)P(M) + P(B|U)P(U)$$

$$P(M|B) = \frac{P(B|M)P(M)}{P(B|M)P(M) + P(B|U)P(U)}$$

$$P(M|B) = \frac{\frac{50}{100} \times \frac{45}{100}}{\frac{50}{100} \times \frac{45}{100} + \frac{80}{100} \times \frac{55}{100}} = \frac{50 \times 45}{50 \times 45 + 80 \times 55} \simeq 0.338$$

2. (a) The probability that a word is chosen at random in a document is "goal" is 0 if the topic of the document is *politics*.

   (b) The probability that a word chosen at random in a document is "goal" is 2/100 if the topic of the document is *sports*, so expectation is $E = 2/100 \times 200 = 4$, which means we expect to see the "goal" four times in that document.

   (c) We should consider cases that the drawn document is *sports* or *politics*, then compute the total probability that a random word of the document is "goal", which is.

$$P(goal) = P(goal|sports)P(sports) + P(goal|politics)P(politics)$$

   We know the following probabilities:

$$P(goal|sports) = 2/100, \quad P(goal|politics) = 0$$

$$P(sports) = 2/3, \quad P(politics) = 1/3$$

$$P(goal) = \frac{2}{100} \times \frac{2}{3} + 0 \times \frac{1}{3} = \frac{4}{300}$$

   (d) The probability that the topic of the document is *sports*, given that the drawn random word is "kick":

$$P(sports|kick) = ?$$

   We can apply the Bayes theorem.

$$P(sports|kick) = \frac{P(kick|sports)P(sports)}{P(kick)}$$

$$P(kick|sports) = 1/200 \quad P(sports) = 2/3 \quad P(kick) = ?$$

   The $P(kick)$ is undefined, so we should first compute this probability.

$$P(kick) = P(kick|sports)P(sports) + P(kick|politics)P(politics)$$

We know the following probabilities:

$$P(kick|sports) = 1/200, \quad P(kick|politics) = 5/1000$$

$$P(sports) = 2/3, \quad P(politics) = 1/3$$

$$P(kick) = \frac{1}{200} \times \frac{2}{3} + \frac{5}{1000} \times \frac{1}{3} = \frac{1}{300} + \frac{5}{3000} = \frac{15}{3000}$$

Now we can compute the $P(sports|kick)$:

$$P(sports|kick) = \frac{\frac{1}{200} \times \frac{2}{3}}{\frac{15}{3000}} = \frac{2}{3}$$

(e)

$$P(goal|kick) = \frac{P(goal, kick)}{P(kick)}$$

$$P(goal|kick) = \frac{P(goal, kick|sports)P(sports) + P(goal, kick|politics)P(politics)}{P(kick)}$$

Words in a document are independent from one another given the topic of the document, so:

$$P(goal, kick|sports) = P(goal|sports)P(kick|sports) = \frac{2}{100} \times \frac{1}{200} = \frac{1}{10000}$$

$$P(goal, kick|politics) = P(goal|politics)P(kick|politics) = 0 \times \frac{5}{1000} = 0$$

From previous section:

$$P(kick) = \frac{15}{3000}$$

$$P(goal|kick) = \frac{\frac{1}{10000} \times \frac{2}{3}}{\frac{15}{3000}} = \frac{1}{5}$$

(f) (i) The conditional probabilities $P(word = \text{"}goal\text{"}|topic = politics)$ can be defined by the definition of conditional probability, than means:

$$P(goal|politics) = \frac{P(goal, politics)}{P(politics)}$$

For $P(goal, politics)$, we divide the number of all documents that their topics are *politics* and they contain "goal" by the number of all documents.

And for $P(politics)$, we divide the number of all documents that their topics are *politics* by the number of all documents.

$$P(goal, politics) = \frac{\#\ documents\ with\ "politics"\ topic\ that\ contain\ "goal"}{\#\ all\ documents}$$

$$P(topic = politics) = \frac{\#\ documents\ with\ "politics"\ topic}{\#\ all\ documents}$$

$$P(goal|politics) = \frac{\#\ documents\ with\ "politics"\ topic\ that\ contain\ "goal"}{\#\ documents\ with\ "politics"\ topic}$$

(ii) The topic probabilities $P(topic = politics)$ can also be defined by dividing the number of documents that their topics are *politics* by the number of all documents.

$$P(topic = politics) = \frac{\#\ documents\ with\ "politics"\ topic}{\#\ all\ documents}$$

3. (a) Since $D = \{x_1, ..., x_n\}$ are drawn independently according to $f_\theta(x)$, we can write the joint distribution $f_\theta(x_1, x_2, ..., x_n)$ as a product of probabilities of all $\{x_1, ..., x_n\}$, which means:

$$f_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n)$$

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

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

(b)

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

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} \prod_{i=1}^{n} f_\theta(x_i)$$

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} (\frac{1}{\theta})^n$$

Since $0 \leq x \leq \theta$, so the maximum likelihood estimate of θ is $\max\left(x_1, \ldots, x_n\right)$; because $x \leq \theta$, which means θ can not be less that maximum values of $(x_i, \ldots, x_n)$.

4. (a) The total area underneath a probability density function is 1:
The length of each bin is $\frac{1}{N}$

$$\frac{1}{N}\sum_{i=1}^{N}\theta_i = 1$$
$$\theta_N = N - (\theta_1 + \cdots + \theta_{N-1})$$
$$\theta_N = N - \sum_{i=1}^{N-1}\theta_i$$

(b) Likelihood:

$$p(D_n;\theta_1,\ldots,\theta_N) = p(x_1,\ldots,x_n;\theta_1,\ldots,\theta_N) = \prod_{i=1}^{n} f(x_i;\theta_1,\ldots,\theta_N)$$

Log-Likelihood:

$$\ell(\theta_1,\ldots,\theta_N) = log(p(D_n;\theta_1,\ldots,\theta_N)) = \sum_{i=1}^{n} log\, p(x_i;\theta_1,\ldots,\theta_N)$$

$$\ell(\theta_1,\ldots,\theta_N) = \sum_{i=1}^{n} log\, \theta_j(x_i \in B_j)$$

We know:

$$\mu_j = \frac{1}{N}\theta_j \quad \theta_j = N\mu_j$$

$$\ell(\theta_1,\ldots,\theta_N) = \sum_{i=1}^{n} log\, N\mu_j(x_i \in B_j)$$

(c)

$$\nabla\ell = \nabla(\sum_{i=1}^{n} log\, N\mu_j(x_i \in B_j))$$
$$\nabla\ell(\theta_1,\ldots,\theta_N) = \sum_{i=1}^{n}\nabla(log\, N\mu_j(x_i \in B_j))$$
$$\nabla\ell(\theta_1,\ldots,\theta_N) = \sum_{i=1}^{n}\nabla(log\, \theta j(x_i \in B_j))$$

for $\theta_j, j \in 1,2,\ldots,N$:

$$\nabla\ell(\theta_1,\ldots,\theta_N)_{\theta_j} = \frac{1}{N\theta_j}$$

5. (a)

$$\mathbb{E}[1_{\{x\in S\}}] = \sum x P_{\{x\in S\}}$$

$$\mathbb{E}[1_{\{x\in S\}}] = 1\cdot P_{\{x\in S\}}(1) + 0\cdot P_{\{x\in S\}}(0)$$

$$\mathbb{E}[1_{\{x\in S\}}] = 1\cdot P(E) + 0\cdot P(E^c)$$

$$\mathbb{E}[1_{\{x\in S\}}] = P(x\in S)$$

(b) I had two suggestion for this question:
First:

As we know:

$$\mathbb{E}1_{\{x\in S\}} = \mathbb{P}(x\in S)$$

And based on the Law of Large Numbers we have:

$$\mathbb{P}(x\in S) = \frac{Volume\ of\ a\ bin}{Total\ volume} = \frac{(\frac{1}{m})^d}{1^d} = (\frac{1}{m})^d$$

Second:

$$P(x\in B_i) = \sum_{j=1}^{m} 1_{\{x\in B_i\}} P(x\in B_j) = \mathbb{E}[1_{\{x\in B_i\}}]$$

$$\mathbb{E}[x\in B_i] = \sum_{k=1}^{n} \frac{P(x_k\in B_i)}{n}$$

We can take $\lim_{n\to\inf}$, so it would be integral on volume of a bin.

(c) Since we divide each dimension to two bins, so:

$$Number\ of\ all\ bins = 2^{784}$$

$$Number\ of\ digits = 237$$

(d) Since we have 2^{784} bins:
For reaching the accuracy from 10% to 90% we need the following number of examples:

$$Number\ of\ data\ point = \frac{(80/5) \times 4 \times 2^{784}}{784} = \frac{16 \times 4 \times 2^{784}}{784}$$

And for reaching the accuracy from 0% to 90%:

$$Number\ of\ all\ data\ point = \frac{(90/5) \times 4 \times 2^{784})}{784} = \frac{18 \times 4 \times 2^{784}}{784}$$

Note: the 784 division comes from the fact that each data points cover 784 dimentions.

(e) We have m^d bins, and one sample falls into d bins, so with n samples, nd bins will be covered:
So probability to fall into a bin is $\frac{nd}{m^d}$

$$P(bin\ i\ is\ empty) = (1 - \frac{nd}{m^d})$$

$$P(bin\ i\ is\ empty) = \frac{m^d - nd}{m^d}$$

6. (a)

$$P(Y=0|X=x) = \frac{P(Y=0, X=x)}{P(X=x)}$$

Applying the Bayes theorem:

$$P(Y=0|X=x) = \frac{P(X=x|Y=0)P(Y=0)}{P(X=x)}$$

$$P(Y=0|X=x) = \frac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=0)P(Y=0) + P(X=x|Y=1)P(Y=1)}$$

Since we have flipped the balanced coin for generating the data:

$$P(Y=0) = P(Y=1) = \frac{1}{2}$$

$$P(Y=0|X=x) = \frac{P(X=x|Y=0)}{P(X=x|Y=0) + P(X=x|Y=1)}$$

$$P(Y=0|X=x) = \frac{f_{\mu_0,\Sigma_0}(\mathbf{x})}{f_{\mu_0,\Sigma_0}(\mathbf{x}) + f_{\mu_1,\Sigma_1}(\mathbf{x})}$$

$$P(Y=0|X=x) = \frac{\mathcal{N}_d\left(\mu_0, \Sigma_0\right)}{\mathcal{N}_d\left(\mu_0, \Sigma_0\right) + \mathcal{N}_d\left(\mu_1, \Sigma_1\right)}$$

(b) We can define the decision boundary:

$$d(x) = \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$$

When $d(x) > 1$ class is 1 and vice versa. We take a log on both sides of the equation.

$$log\ d(x) = log\ P(Y=1|X=x) - log\ P(Y=0|X=x)$$

Applying the Bayes rule:

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

$$\begin{aligned} log\ d(x) = &(log\ P(X=x|Y=0) + log\ P(Y=0) - log\ P(X=x)) \\ &- (log\ P(X=x|Y=1) + log\ P(Y=1) - log\ P(X=x)) \end{aligned}$$

$$
\begin{aligned}
log\, d(x) = &(log\, P(X = x|Y = 0) + log\, P(Y = 0)) \\
&- (log\, P(X = x|Y = 1) + log\, P(Y = 1))
\end{aligned}
$$

Denoting $\pi_j = P(Y = j)$.

$$d(x) = \pi_0 + \frac{1}{2}\left(\mathbf{x} - \mu_0\right)^T \Sigma_0^{-1}\left(\mathbf{x} - \mu_0\right) - \frac{1}{2}\left(\mathbf{x} - \mu_1\right)^T \Sigma_1^{-1}\left(\mathbf{x} - \mu_1\right) - \pi_1$$

We set $d(x) = 0$ for finding the decision boundary.

$$
\begin{gathered}
C + x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 = x^T\Sigma^{-1}x - 2\mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 \\
\left[2\left(\mu_0 - \mu_1\right)^T\Sigma^{-1}\right]x - \left(\mu_0 - \mu_1\right)^T\Sigma^{-1}\left(\mu_0 - \mu_1\right) = C
\end{gathered}
$$

In case that $\Sigma_1 = \Sigma_0$, the quadratic terms will be canceled, and it would be a linear equation.

$$a^T x - b = 0$$