1. Probability warm-up: conditional probabilities and Bayes Rule

a)

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

b)

A= event of exactly two head occurs = {[h,h,t],[h,t,h],[t,h,h]}

B = event of first outcome is head = {[h,t,t],[h,t,h],[h,h,h],[h,h,t]]}

P(A,B)=set{[h,h,t],[h,t,h]}

P(A|B) = $\frac{P(A,B)}{P(B)} = \frac{2*\frac{2}{3}^2*\frac{1}{3}}{\frac{2}{3}} = \frac{4}{9}$

c)

from part(a) we have $P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$ and $P(Y|X) = \frac{P(Y \cap X)}{P(X)}$

since P(X,Y) = P($X \cap Y$) = P($Y \cap X$) we have:

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

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

d)

from the previous part we have

P(X,Y) = P(X,Y) =>

P(Y|X)*P(X) = P(X|Y)*P(Y)

This implies:

P(X|Y) = $\frac{P(Y|X)*P(X)}{P(Y)}$ (s.t P(Y)!=0)

e)

i)

P(McGill) = 1- P(Udem) = 0.4

since 0.4 percent of students are from McGill so this probability is equal to 0.4.

ii)

P(McGill) = 0.4

P(UDEM) = 0.6

P(Bilingual|UDEM) = 0.6

P(Bilingual | McGill) = 0.3

P(McGill | Bilingual) = $\frac{P(Bilngual\ |McGill)*P(McGill)}{P(Bilngual)} = \frac{P(Bilngual\ |McGill)*P(McGill)}{P(Bilngual\ |McGill)*P(McGill)+ P(Bilngual\ |UDEM)*P(UDEM)} = \frac{0.3*0.4}{0.3*0.4+0.6+0.6}$= 0.25

2. Bag of words and single topic model

a)

from the table we have:

p("Goal" | "politic") = 0.005

b)

from the table we have:

P("Congress" | "sport") = 0.001

In a document of 2000 independents words (each word is independent from others) we have 200 repetitions of the above events, so 2000*0.001 = 2 times "Congress" will appear in document with sport topic.

c)

P(“Goal”) = P(“Goal” | “Sport”) * P(“Sport”) + P(“Goal” | “Politic”) * P(“Politics”) = (1/100)*(1/3) + (5/1000)*(2/3) = 0.0066

d)

P(“Sport” | “Kick”) = $\frac{P(\text{"Kick"} \mid \text{"Sport"}) *P(\text{"Sport"})}{P(\text{"Kick"} \mid \text{"Sport"}) *P(\text{"sport"}) + P(\text{"Kick"} \mid \text{"Politic"}) * P(\text{"Politic})}$ =

$\frac{\left(\frac{2}{100}\right)*\left(\frac{1}{3}\right)}{\left(\frac{2}{100}\right)*\left(\frac{1}{3}\right) + (1/1000)*(2/3)}$ = 0.90

e)

P(word2 = “Vote” | word1 = “Kick”) =

$\frac{P(word2\text{= “}Vote\text{” },word1 = \text{“}Kick\text{”})}{P(word1\text{="}kick\text{"}}$ =

$\frac{P(word2\text{= “}Vote\text{” },word1 = \text{“}Kick\text{” } \mid \text{"}Sport\text{"})P(\text{"Sport"}) + P(word2\text{= “}Vote\text{” },word1 = \text{“}Kick\text{” } \mid \text{"}Politic\text{"})P(\text{"Politic"})}{P(word1\text{=kick})}$

Since the words are independent of each other:

$$\frac{P(word2\ =\ \text{“}Vote\text{” } \mid \text{"Sport"})P(\ word1\ =\ \text{“}Kick\text{” } \mid \text{"}Sport\text{"})P(\text{"Sport"})\ +\ P(word2\ =\ \text{“}Vot}{P(word1\ =\ \text{"kick"} \mid \text{"Politic"})P(\text{"Politic"}) + P(word1\ =\ \text{"k}}$$

$=\frac{\frac{3}{1000}*\frac{2}{100}*\frac{1}{3}+\frac{4}{100}*\frac{1}{1000}*\frac{2}{3}}{\frac{1}{1000}*\frac{2}{3}+\frac{2}{100}*\frac{1}{3}}$= 0.006

f)

Without any previous knowledge we have to do density estimation.

So for topic probabilities we compute the frequency of each topic among all the documents and we have:

P(“Politic”) = $\frac{N_{Topics==\text{"}Politic\text{"}}}{N}$

P("Sport") = $\frac{N_{Topics=="Sport"}}{N}$

And we do the same for conditional probability like if we want to calculate the probability of observing word="kick" in a document of "Sport" topic we should do:

P("Kick" | "Sport") = $\frac{N_{Topics=="Sport"}\ And\ N_{Word=="Kick"}}{N_{Topics=="Sport"}}$

Which means that we count the number of times word="kick" appears in documents with topic ="sport" and we divide it by total number of words that documents with topic="sport" have. we do the exact same process for each W∈{"goal", "kick", "congress", other} and documents with T ∈ {"Sport", "Politic"} to calculate P(Word=W | Topic= T).

3. Maximum likelihood estimation

a)

Because x_is are IID so the joint density distribution is :

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

b)

θ_{MLE} = argmax($F_\Theta(x_1,\ldots,x_n)$) = argmax(Log l(D,Θ))

Log l(D,Θ) = $\sum_{i=1}^{n} log(F_\theta(x_i))$ = $\sum_{i=1}^{n}(log(2) + log(\theta) + log(x_i) - \theta x_i^2)$

To get the maximum point of this function we need to set its gradient with respect to its parameter to zero so we have:

$\frac{\delta \text{Log l(D,}\Theta)}{\delta\Theta}$ =0 -> $\sum_{i=1}^{n}((1/\Theta)- x_i^2)$ =0 -> $\theta_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i^2}$

4. Maximum likelihood meets histograms

a)

the width of each bin is 1/N since there are N bins between 0 and 1.

Because of the fact that the total area underneath a probability density function is 1 and indeed the area of histogram is sum of the area of each histogram bin which should add up to 1 so we have:

$\sum_{j=1}^{N} \theta_j * \frac{1}{N} = \sum_{j=1}^{N-1} \theta_j * \frac{1}{N} + \theta_N * \frac{1}{N}$ =1

θ_N = N - $\sum_{j=1}^{N-1} \theta_j$

b)

log likelihood= $\sum_{i=1}^{n} \log(p(x_i , \theta_1, \theta_2, \dots , \theta_N)))$ =

$\sum_{i=1}^{n} \sum_{j=1}^{N} \log(\theta_j) I(x_i \epsilon\, B_j)$ = $\sum_{j=1}^{N} \sum_{i=1}^{n} \log(\theta_{j)}) I(x_i \epsilon\, B_j)$
=$\sum_{j=1}^{N} \log\left(\theta_j \right) \sum_{i=1}^{n} I\left(x_i \epsilon\, B_j\right)$ => by definition stated in the question the number of data points in B_j is μ_j .

$= \sum_{j=1}^{N} \log(\theta_j)\, \mu_j$

=$\sum_{j=1}^{N-1} \log(\theta_j)\, \mu_j$ + $\log(\theta_N)(\mu_N)$

By the previous step we have θ_N = N - $\sum_{j=1}^{N-1} \theta_j$ and since there are n total points μ_N= $n - \mu_1 - \mu_2 - \cdots - \mu_{N-1}$ so we have:

log likelihood = $\sum_{j=1}^{N-1} \log(\theta_j)\, \mu_j$ + log(N-θ_1-$\theta_2 - \cdots - \theta_{N-1}$)$(n - \mu_1 - \mu_2 - \cdots - \mu_{N-1})$

c)

from the previous step we just take the gradient of calculated log likelihood with respect to θ_j:

$\frac{\delta}{\delta\theta_j}(\sum_{j=1}^{N-1} \log(\theta_j)\, \mu_j$ + log(N-θ_1-... $-\,\theta_{N-1})(n-\mu_1-\cdots-\mu_{N-1}))$

= $\frac{\mu_j}{\theta_j}$ - $\frac{n-\mu_1-\cdots\mu_{j-1}-\mu_{j+1}-\cdots.-\mu_{N-1}}{N-\theta_1-\cdots\theta_{j-1}-\theta_{j+1}-\cdots-\theta_{N-1}}$ = 0

$\theta_j = \mu_j \frac{N-\theta_1-\cdots\theta_{j-1}-\theta_{j+1}-\cdots-\theta_{N-1}}{n-\mu_1-\cdots\mu_{j-1}-\mu_{j+1}-\cdots.-\mu_{N-1}}$

5. Histogram methods

a)

E[$1_{\{x\in s\}}$] =$\int 1 * P(x \in s)dx \quad if\ x \in s\ + \int 0 * P(x\ !\in s)dx \quad if\ x\ ! \in s$ =

p($x \in s$)

b)

from the previous step we know that P($x \in V_i$)= $\int_{V_i} P(x)dx$

and form the law of large number we have: $\lim_{n\to\infty} p(x) = f(x)$

combining the two equation we have:

$$\lim_{n\to\infty} \mathrm{P}(x \in V_i) = \int_{V_i} f(x)dx$$

c)

with 2^{784} bins in total we have $\log_{10}(2^{784}) \cong 237$ digits.

d)

With 2^{784} bins in total, in order to increase the accuracy 5% by adding k=4 samples to each bin we need to add $k*2^{784}$ data points in total. So starting from 10% accuracy in order to reach 90% accuracy we need to $(\frac{90-10}{5})$ times increase the accuracy by 5% (or 16 times add 4 points to each bin) which mean we need to add $16*4*2^{784}$ new data points in total.

e)

the probability of a bin containing specific data point is: $\frac{1}{number\ of\ bins} = \frac{1}{m^d}$

and the probability of a bin doesn't contain that specific data point is:$1-\frac{1}{m^d}$

now since the data points are independent of each other we can say that the probability of a bin doesn't contain any data point =

$\prod_{i=1}^{n}$ probability of a bin doesn't contain data point(i) = $\prod_{i=1}^{n} 1 - \frac{1}{m^d}$ =

$(1 - \frac{1}{m^d})^n$

6. Gaussian Mixture

a)

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

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

$$\frac{\frac{1}{2}*\frac{e^{\frac{-1(x-\mu 0)^T sigma0^{-1}(x-\mu 0)}{2}}}{2\pi^{\frac{d}{2}}|sigma0|^{1/2}}}{\frac{1}{2}*\frac{e^{\frac{-1(x-\mu 0)^T sigma0^{-1}(x-\mu 0)}{2}}}{2\pi^{\frac{d}{2}}|sigma0|^{\frac{1}{2}}}+\frac{1}{2}*\frac{e^{\frac{-1(x-\mu 1)^T sigma1^{-1}(x-\mu 1)}{2}}}{2\pi^{\frac{d}{2}}|sigma1|^{1/2}}} =$$

$$\frac{\frac{e^{\frac{-1(x-\mu 0)^T sigma0^{-1}(x-\mu 0)}{2}}}{|sigma0|^{1/2}}}{\frac{e^{\frac{-1(x-\mu 0)^T sigma0^{-1}(x-\mu 0)}{2}}}{|sigma0|^{\frac{1}{2}}}+\frac{e^{\frac{-1(x-\mu 1)^T sigma1^{-1}(x-\mu 1)}{2}}}{|sigma1|^{1/2}}}$$

b)

when two covariance matrices are equal we have:

p(Y=0 |X=x) = $\frac{e^{\frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2}}}{e^{\frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2}}+e^{\frac{-1(x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2}}} =$

$$\frac{1}{1+e^{\frac{-1(x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2}-\frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2}}}$$

$$p(Y=1 \mid X=x) = \frac{e^{\frac{-1(x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2} - \frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2}}}{1+e^{\frac{-1(x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2} - \frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2}}}$$

assume we classify data point x to y=1 if p(y=1|x) > p(y=0|x) or likewise $\frac{\mathrm{p(y=1|x)}}{\mathrm{p(y=0|x}}$ >1 or log($\frac{\mathrm{p(y=1|x)}}{\mathrm{p(y=0|x}}$)>0:

$$\log\left(\frac{\mathrm{p(y=1|x)}}{\mathrm{p(y=0|x}}\right) = \log\left(e^{\frac{-1(x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2} - \frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2}}\right) =$$

$$\frac{-1(x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2} - \frac{-1(x-\mu 0)^T sigma^{-1}(x-\mu 0)}{2} =$$

$$\frac{(x-\mu 0)^T sigma^{-1}(x-\mu 0) - (x-\mu 1)^T sigma^{-1}(x-\mu 1)}{2} =$$

$$\frac{x^T sigma^{-1} x - x^T sigma^{-1}\mu 0 - \mu 0^T sigma^{-1} x + \mu 0^T sigma^{-1}\mu 0 - x^T sigma^{-1} x + x^T sign}{2}$$

$$= \frac{2x^T sigma^{-1}(\mu 1-\mu 0)+\mu 0^T sigma^{-1}\mu 0 - \mu 1^T sigma^{-1}\mu 1}{2}$$

Which shows that our classifier is linear in x.