IFT 6390B Fundamentals of Machine Learning
Ioannis Mitliagkas

Homework 1 - Theory

Submission: Aasheesh Singh

Solutions:

1. **Conditional probabilities and Bayes rule**

   (a) The conditional probability of a random variable **X** given another variable **Y** is defined as the probability of occurrence of X given that event Y has already happened.

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

   where P(X,Y) = Joint distribution of (X,Y) which can be expressed as:

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

   therefore we get,

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

   (b) Let X denote the event of getting a Heads in the first toss.
   Let Y denote the event of getting exactly 1 Head in the next two coin tosses. Therefore we get exactly 2 heads in the total three tosses.

   Given, $P(Head) = \frac{2}{3}, P(Tail) = \frac{1}{3}$

   We can thus calculate the conditional probability P(Y|X) as:

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

P(X,Y) = P(Head,Head,Tails) + P(Head, Tail, Head)

Since individual coin tosses are independent events we can write:

$$P(Head, Head, Tails) = P(Head)P(Head)P(Tail) = \frac{2}{3}*\frac{2}{3}*\frac{1}{3} = \frac{4}{27}$$

Similarly, P(Head, Tail, Head) $= \frac{2}{3} * \frac{1}{3} * \frac{2}{3} = \frac{4}{27}$

Combining we get:

$$P(Y|X) = \frac{P(Head, Head, Tail) + P(Head, Tail, Head)}{P(Head)} = \frac{4}{9}$$

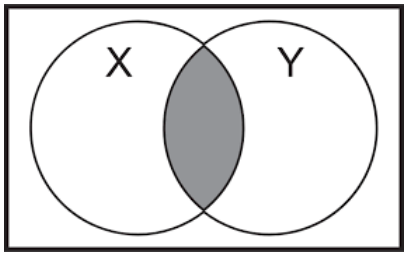
as the answer.

(c) The joint distribution P(X,Y) can be expressed through the conditional probability of individual events as following:

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

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

(d) Let X, Y denote two random variable where P(X∩Y) denotes the joint or shared probability of their occurrence.



Grey region denotes the joint distribution P(X, Y)

The joint distribution can thus be expressed through the of conditional probability of X given Y has already happened:

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

or,

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

The join distribution P(X,Y) can also be expressed given the vice-versa: probability of Y given X has already happened:

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

Substituting this in (1), we get:

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

which gives us the Baye's theorem for computing posterior distribution.

(e) Let X be a random variable denoting the university of a student in the set {UdeM, McGill}.
Let random variable Y denote the bilingual status of a student.

Given that, P(X=UdeM) = 0.6

i. P(X=McGill) = 1 - P(X=UdeM) = 0.4, is the answer

ii. We are given the following conditional probabilities as below:
P(Y=bilingual | X=UdeM) = 0.6
P(Y=bilingual | X=McGill) = 0.3

Now we want to calculate the probability that given a sampled student is bilingual, whats the probability of them being from McGill.

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

The denominator is a sum across both events, i.e.

P(Y=bilingual) = P(bilingual|McGill)*P(McGill) + P(bilingual|UdeM)*P(UdeM)

Plugging in the values in the equation we get,

$$P(X = McGill|Y = bilingual) = \frac{0.3 * 0.4}{0.3 * 0.4 + 0.6 * 0.6}$$

P(X=McGill | Y=bilingual) = 0.25, as the answer.

2. **Bag of words and topic model**

Let X denote the event of document topic being from set {Sports, Politics}.
Let Y denote the word category from the set {goal, kick, congress, vote, other}.

(a) From the likelihood table, we can get the probability:

$$P(Y = goal|X = Politics) = \frac{5}{1000}$$

as the answer.

(b) Frequency(Y=congress|X=Sports) = P(Congress|Sports)*(Num of trials)
$Frequency(Y = congress|X = Sports) = \frac{1}{1000} * (2000) = \mathbf{2}$, as the answer.

(c) P(Y=goal) = P(Y=goal|X=sports)P(X=sports) + P(Y=goal|X=Politics)P(X=Politics) plugging in the values we get,

$$P(Y = goal) = \frac{1}{100} * \frac{1}{3} + \frac{5}{1000} * \frac{2}{3}$$

,

$$P(Y = goal) = \frac{2}{300} = \frac{1}{150}$$

as the answer.

(d) From Baye's law we can write:

$$P(X = sports|Y = kick) = \frac{P(kick|sports)P(sports)}{P(kick|sports)P(sports) + P(kick|politics)P(politics)}$$

Pluging the values in, we get:

$$P(X = sports|Y = kick) = \frac{\frac{2}{100} * \frac{1}{3}}{(\frac{2}{100} * \frac{1}{3} + \frac{1}{1000} * \frac{2}{3})}$$

$$P(X = sports|Y = kick)] = \frac{10}{11} = 0.909$$

as the answer.

(e) There are two possible scenarios given that the first word drawn is "kick". Either the word could have been drawn from "sports" topic or "politics" topic. Since the second word "vote" is drawn from the same document, we can compute its probability as:

$$P(vote, doc|kick) = P(vote, sports|kick) + P(vote, politics|kick)$$

Now since drawing two words from the same document are independent events we get:

$$P(vote, doc|kick) = P(vote|sports)P(sport|kick)+P(vote|politics)P(politics|kick)$$

from above question, we have the result:

$$P(sports|kick) = \frac{10}{11}$$
$$P(politics|kick) = 1 - P(sports|kick) = \frac{1}{11}$$

Given, $P(vote|sports) = 3/1000$
$P(vote|politics) = 4/100$

Plugging these values in the equation we get:

$$P(vote, doc|kick) = (3/1000 * 10/11 + 4/100 * 1/11)$$

$P(vote, doc|kick) = 7/1100$ as the answer.

(f) The topic and word probabilities can be calculated by counting their occurrence frequency in the corpus. They can be defined as:

$$P(topic = politics) = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{topic[i]=politics\}}}{N}$$

The distribution of a word given a topic can be calculated as following:

$$P(word = goal|topic = politics) = \frac{\text{Count of word "goal" in all documents}}{\text{Total num of words across documents}}$$

3. **Maximum likelihood estimation**

(a) Since samples are drawn independently, their joint pdf will same as the product of their individual distributions:

$$f_\theta(x_1, x_2, ...x_n) = f_\theta(x_1) * f_\theta(x_2)... * f_\theta(x_n)$$

(b) Given the definition of MLE of θ_{MLE} and the above expansion of $f_\theta(x_1, x_2, ...x_n)$, we can write:

$$\theta_{MLE} = argmax \prod_{i=1}^{n} 2\theta x_i \exp^{-\theta x_i^2}$$

Since logarithm is a monotonically increasing function, argmax f(x) = $argmax \log f(x)$, therefore:

$$\theta_{MLE} = argmax \sum_{i=1}^{n} \log(2\theta x_i \exp^{-\theta x_i^2})$$

using the sum property of logarithm.

For max value of θ, the gradient of the above term wrt to it has to be zero, therefore:

$$\frac{\partial J}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} \log(2\theta x_i \exp^{-\theta x_i^2})}{\partial \theta}$$

$$\frac{\partial J}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{2\theta x_i \exp^{-\theta x_i^2}} (2x_i \exp^{-\theta x_i^2} - 2\theta x_i \exp^{-\theta x_i^2} x_i^2) = 0$$

dividing common terms in numerator and denominator, we get a simplified expression:

$$\frac{\partial J}{\partial \theta} = \sum_{i=1}^{n} \frac{1 - \theta x_i^2}{\theta} = 0$$

therefore we get,

$$\frac{\partial J}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i^2 = 0$$

thus,

$$\theta_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i^2}$$

is the answer.

4. **Maximum likelihood meets histograms**

   (a) The are underneath a pdf totals to 1, therefore for $0 \leq x \leq 1$, we can write

   $$\int_0^1 p_\theta(x) dx = 1$$

   The integral can be split across bin intervals, i.e. $0 \leq x < \frac{1}{N}$; $\frac{1}{N} \leq x < \frac{2}{N}$.. etc. and each integral can be computed separately in its limits. For example the first bin integral equals to:

   $$\int_0^{1/N} \theta_1 dx = \frac{\theta_1}{N}$$

   therefore the original integral can now be defined as:

   $$\frac{\theta_1}{N} + \frac{\theta_2}{N} + .. + \frac{\theta_N}{N} = 1$$

   thus,

   $$\theta_N = N - (\theta_1 + \theta_2 + .. + \theta_{N-1})$$

   is the answer.

(b) The log likelihood of i.i.d data D $\in\{X_1, X_2, X_3, ..X_n\}$ can be defined as:

$$\text{Log likelihood of D} = \log p(X_1, X_2, ..X_n|\theta_1, \theta_2, ..\theta_N)$$

Since samples are independent from each other, we can separate the joint distribution into their individual probabilities as below:

$$\text{Log likelihood} = \log[p(X_1|\theta_1, ..\theta_N)p(X_2|\theta_1, ..\theta_N)..p(X_n|\theta_1, ..\theta_N)]$$

$$= \sum_{i=1}^{n} \log p(X_i|\theta_1, \theta_2, ..\theta_N)$$

from given pdf definition, we know for any point i $\in$ {1, 2...n}

$$p(X_i) = \theta_j;$$

where j refers the bin number to which the point belongs taking value $\in$ {1, 2,..N}

Log likelihood = $\sum_{i=1}^{n} \log \theta_j$; j=bin number for data point X_i

Now from definition μ_j where j $\in$ {1,2..N} refers to the number of points in each bin, therefore the summation term over 'n' data points can be written as:

$$\text{Log likelihood} = \mu_1 \log \theta_1 + \mu_2 \log \theta_2 + .. + \mu_N \log \theta_N$$

where μ_1 points belong to 1st bin, μ_2 points belong to 2nd bin and so forth. To express it in terms of upto μ_{N-1} and θ_{N-1} we can use the below two conclusions:

$$\theta_N = N - (\theta_1 + \theta_2 + .. + \theta_{N-1}) \tag{1}$$
$$\mu_N = n - (\mu_1 + \mu_2 + .. + \mu_{N-1})$$

where n denotes the number of data points $\{X_1, X_2, ..X_n\}$.

Plugging these back in the log likelihood equation we can write the equation in terms of $\{\theta_1, \theta_2, ..\theta_{N-1}\}$ and $\{\mu_1, \mu_2, ..\mu_{N-1}\}$:

$$\log p(X_1, X_2, ..X_n|\theta_1, \theta_2..\theta_{N-1}) = \sum_{j=1}^{N-1} \mu_j \log \theta_j + [n - \sum_{j=1}^{N-1} \mu_j] \log(N - \sum_{j=1}^{N-1} \theta_j)$$

is the answer.

(c) Max likelihood estimate for any $\theta \in$ {1,2,..N} will thus be defined as:

$$\theta_j^{MLE} = argmax \sum_{j=1}^{N} \mu_j \log \theta_j$$

for MLE estimate, the gradient of log likelihood wrt any $\theta_j = 0$, j $\in \{1, 2, ..N\}$ therefore:

$$\frac{\partial L}{\partial \theta_j} = \frac{\partial \sum_{j=1}^{N} \mu_j \log \theta_j}{\partial \theta_j} = 0;$$

Calculating the partial derivative with respect to say, θ_1 we get:

$$\frac{\partial L}{\partial \theta_1} = \frac{\mu_1}{\theta_1} = 0$$

The only solution that will satisfy this will be $\mu_1 = 0$ which is not helpful in getting an estimate for θ_1. Thus we will use 1, to solve a constraint optimisation problem while maximising the log likelihood.

$$max(\sum_{j=1}^{N} \mu_j \log \theta_j)$$
$$\text{under constraint } (N - \sum_{j=1}^{N} \theta_j) = 0 \text{ (from 1)}$$

This optimization problem with linear constraint can thus be modeled using Lagrangian method as follows:

$$L(\theta_1, \theta_2..\theta_N, \lambda) = f(\theta_1, \theta_2..\theta_N) + \lambda g(\theta_1, \theta_2..\theta_N)$$

where; $f(\theta_1, \theta_2..\theta_N) = \sum_{j=1}^{N} \mu_j \log \theta_j$

$g(\theta_1, \theta_2..\theta_N) = N - \sum_{j=1}^{N} \theta_j$

Taking partial derivatives wrt θ_j for j $\in$ {1,2,..N} and λ, we get:

$$\frac{\partial L}{\partial \theta_j} = \frac{\mu_j}{\theta_j} - \lambda(1) = 0$$

$$\theta_j = \frac{\mu_j}{\lambda}$$

$$\frac{\partial L}{\partial \lambda} = 0 + (1)N - (1)\sum_{j=1}^{N} \theta_j = 0$$

substituting value of $\theta_j = \frac{\mu_j}{\lambda}$ in the above we get,

$$N - \frac{\sum_{j=1}^{N} \mu_j}{\lambda} = 0$$

since the sum of points in each bin $\sum_{j=1}^{N} \mu_j = n$;

$$\lambda = \frac{n}{N}$$

putting the value of λ back in θ_j expression we get,

$$\boxed{\theta_j^{MLE} = \frac{\mu_{jN}}{n}} \tag{2}$$

as the answer.

5. (a) Let I_x represent the indicator random variable such that:

$$I_x = 1 \ \forall \ \text{x} \in \text{S};$$

$$I_x = 0 \ \forall \ \text{x} \notin \text{S}$$

By definition, $\text{E}[I_x] = \text{E}[\mathbb{1}_{x \in S}]$

and expectation of any random variable is given by $E[z] = \sum zP(z)$, we can thus write

$E[\mathbb{1}_{x \in S}] = (1)P(I_X = 1) + (0)P(I_x = 0)$

where $P(I_x = 1) = P(x \in S)$
and $P(I_x = 0) = P(x \notin S)$

therefore we get the result: $E[\mathbb{1}_{x \in S}] = P(x \in S) + 0 = P(x \in S)$

(b) Let P denote the true probability of a point falling in bin i with volume V_i :

$$P = \int_{V_i} f(x)dx$$

where f(x) is the true pdf

Let 'k' denote the no. of points that belong to bin i out of the total 'n' points in the dataset, where k $\in$ {0,1,2,..n}. The estimated probability for the bin i will therefore be:

$$P_{estimate} = \frac{\text{No. of points in bin 'i'}}{\text{Total no of points}} = \frac{k}{n}$$

Now as $n \to \infty$, the estimated probability will tend to the true probability since the estimate will get better with more data, therefore:

$$\mathbf{E}[P_{estimate}] \approx P$$

The expectation of $P_{estimate}$ can thus be calculated by:
$\mathbf{E}[P_{estimate}] = \mathbf{E}[k/n]$

$$= \frac{1}{n}\mathbf{E}[kp(k)]$$

The random variable 'k', which is no of points in a bin follows a binomial distribution with:

$$p(k) = \binom{n}{k} P^k (1-P)^{n-k}$$

where P is the true probability in bin 'i'

$$= \frac{1}{n} \sum_{k=0}^{n} k \binom{n}{k} P^k (1-P)^{n-k}$$

$$= \frac{1}{n} * n * P \sum_{k=0}^{n} [\binom{n-1}{k-1} P^{k-1} (1-P)^{(n-1)-(k-1)}$$

$$= P \sum_{k=0}^{n} \binom{n-1}{k-1} P^{k-1} (1-P)^{(n-1)-(k-1)}$$

The expectation can be simplified by using the binomial theorem which for any x,y is defined as follows:

$$\sum_{k=1}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n$$

Using this result and substituting x = P and Y = 1-P, we get

$$\mathbf{E}[P_{estimate}] = P * (P + (1 - P))^n = P$$

or,

$$\mathbf{E}[P_{estimate}] = \int_{V_i} f(x)dx$$

as the answer.

(c) Total number of bins = 2^{784}
The no of digits can be counted by observing the pattern in powers of 2 as follows:

$\{2, 2^2, 2^3\}$ = 1 digit
$\{2^4, 2^5, 2^6\}$ = 2 digits
$\{2^7, 2^8, 2^9\}$ = 3 digits and so forth..

Therefore in general for any n, no of digits = n//3 + 1

For n=784, we get no of digits = 784//3 + 1 = 262 as the answer.

(d) Total % increase in accuracy = 80
step size = 5 %
total steps = 80/5 = 16
Points needed per 5% increase = 4* Total bins = $4*2^{784} = 2^{786}$

Total points to get 90% accuracy = (Total steps)*(Points per step)

$= 2^4 * 2^{786}$
$= 2^{790}$ is the answer

(e) Total no of bins = m^d
Let's say bin 'i' is empty, for that to happen, all 'n' points need to belong to every other bin except this one. Thus there are $m^d - 1$ options for each point.

Total no of ways 'n' points can be distributed in $m^d - 1$ bins = $(m^d - 1)^n$

Total no of ways 'n' points can be distributed in 'n' bins = m^{dn}

$$Pr(bin_i = empty) = \frac{(m^d - 1)^n}{m^{dn}}$$

$$Pr(bin_i = empty) = (1 - \frac{1}{m^d})^n$$

is the answer.

6. Given the below distributions:

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

$P(X = x|Y = O) = \mathcal{N}(\mu_0, \sum_0)$
$P(X = x|Y = 1) = \mathcal{N}(\mu_1, \sum_1)$

   (a) To calculate P(Y=0 | X=x) = ?

   Using baye's rule we can write P(Y|X) in terms of the likelihood P(X|Y) and prior P(X) as follows:

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

   where Prior P(X=x) can be calculated by the sum:

$$P(X = x) = P(X = x|Y = 0)P(Y = 0) + P(X = x|Y = 1)P(Y = 1)$$

$$P(Y = 0|X = x) = \frac{\mathcal{N}(\mu_0, \Sigma_0) * \frac{1}{2}}{\mathcal{N}(\mu_0, \Sigma_0) * \frac{1}{2} + \mathcal{N}(\mu_1, \Sigma_1) * \frac{1}{2}}$$

$$= \frac{\frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}}{(\frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)} + \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)})}$$

$$= \frac{1}{1 + \frac{\sqrt{det(\Sigma_0)}}{\sqrt{det(\Sigma_1)}} * e^{-\frac{1}{2}[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]}}$$

$$= \frac{1}{1 + \sqrt{\frac{det(\Sigma_0)}{det(\Sigma_1)}} * e^{-\frac{1}{2}[(x^T\Sigma_1^{-1}x - 2\mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1) - (x^T\Sigma_0^{-1}x - 2\mu_0^T\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0)]}}$$

$$= \frac{1}{1 + \sqrt{\frac{det(\Sigma_0)}{det(\Sigma_1)}} * e^{-\frac{1}{2}[x^T(\Sigma_1^{-1} - \Sigma_0^{-1})x + (2\mu_0^T\Sigma_0^{-1} - 2\mu_1^T\Sigma_1^{-1})x + (\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0)]}}$$

$$P(Y=0|X=x) = \frac{1}{1 + \sqrt{\frac{det(\Sigma_0)}{det(\Sigma_1)}} * e^{-\frac{1}{2}[x^TAx + Bx + C]}}$$

where, $A = \Sigma_1^{-1} - \Sigma_0^{-1}$

$B = (2\mu_0^T\Sigma_0^{-1} - 2\mu_1^T\Sigma_1^{-1})$

$C = (\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0)$

(b) $h_{Bayes}(x) = argmax(P(Y=0|X=x), P(Y=1|X=x)$

Given that, $\Sigma_1 = \Sigma_0 = \Sigma$, we get:
$A = \Sigma_1^{-1} - \Sigma_0^{-1} = 0$;

and $\sqrt{\frac{det(\Sigma_0)}{det(\Sigma_1)}} = 1$

Therefore we can write the below simplified expressions for P(Y=0|X=x) and P(Y=1|X=x):

$$P(Y = 0|X = x) = \frac{1}{1 + e^{-\frac{1}{2}(B_0 x + C_0)}};$$

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-\frac{1}{2}(B_1 x + C_1)}}$$

where $B_0 = (2\mu_0^T \Sigma^{-1} - 2\mu_1^T \Sigma^{-1}); B_1 = (2\mu_1^T \Sigma^{-1} - 2\mu_0^T \Sigma^{-1})$

and $C_0 = (\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0); C_1 = (\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1)$

We can therefore see that the Bayes classifier with equal covariance takes the form of logistic regression which is a linear classifier:

$$P(Y = y|X = x) = sigmoid((Bx + C)/2)$$

The decision boundary is therefore a hyperplane and the classifier is linear in x.