Homework Assignment # 1

Submitted by: Prerit Anwekar

Question 1. [10 MARKS]

Let $\Omega_X = \{a, b, c\}$ and $p_X(a) = 0.1, p_X(b) = 0.2$, and $p_X(c) = 0.7$. Let

$$f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

(a) [3 MARKS] What is $E[f(x)]$?

Answer(a). Given, $\Omega_X = \{a, b, c\}$ and $p_X(a) = 0.1, p_X(b) = 0.2$, and $p_X(c) = 0.7$. and

$$f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

$$\begin{aligned} E[f(x)] &= \sum_{x \in \Omega} f(x)\, p_X(x) \\ &= f(a)\, p_X(a) + f(b)\, p_X(b) + f(c)\, p_X(c) \\ &= 10 \times 0.1 + 5 \times 0.2 + \tfrac{10}{7} \times 0.7 \\ &= 1 + 1 + 1 \\ &= 3 \end{aligned}$$

(b) [3 MARKS] What is $E[1/p_X(x)]$?

Answer(b). According to given condition we need to change our pmf $f(x)$ such that,

$$f(x) = \frac{1}{p_X(x)} = \begin{cases} \frac{1}{0.1} & \text{if } x = a \\ \frac{1}{0.2} & \text{if } x = b \\ \frac{1}{0.7} & \text{if } x = c \end{cases}$$

$$\therefore \quad f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

Hence,

$$\begin{aligned} E[f(x)] &= \sum_{x \in \Omega} f(x)\, p_X(x) \\ &= f(a)\, p_X(a) + f(b)\, p_X(b) + f(c)\, p_X(c) \\ &= 10 \times 0.1 + 5 \times 0.2 + \tfrac{10}{7} \times 0.7 \\ &= 1 + 1 + 1 \\ &= 3 \end{aligned}$$

(c) [4 MARKS] For an arbitrary pmf $p_X(x)$, what is $E[1/p_X(x)]$?

Answer(c). If we have an arbitrary pmf we don't know the value of pmf at the given discrete points. and we know that for a pmf $\sum_{x\in X} p_X(x) = 1$.

$\therefore$ *Let* $g(x) = \frac{1}{p_X(x)}$

$$\begin{aligned} E[g(x)] &= \sum_{x\in\Omega} g(x)\, p_X(x) \\ &= \frac{1}{p_x(a)}\, p_X(a) + \frac{1}{p_x(b)}\, p_X(b) + \frac{1}{p_X(c)}\, p_X(c) \\ &= 1 + 1 + 1 \\ &= 3 \end{aligned}$$

Question 2. [15 MARKS]

Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be independent multivariate Gaussian random variables, with $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, with $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d\times d}$ for dimension $d \in \mathbb{N}$. Define $\mathbf{X} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_m\mathbf{X}_m$ as a convex combination, $a_i \geq 0$ and $\sum_{i=1}^m a_i = 1$.

(a) [5 MARKS] Write the expected value $E[\mathbf{X}]$ in terms of the givens $a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$. Show all you steps. What is the dimension of $E[\mathbf{X}]$?

Answer(a). Given, $\mathbf{X} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_m\mathbf{X}_m$ where each of $\mathbf{X}_1, \ldots, \mathbf{X}_m$ are random vectors of Gaussian random variables and representing independent multivariate Gaussian random variables.

$\therefore$ Let, $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})$ be the random vector of i^{th} multivariate Gaussian.

$$\begin{aligned} X &= a_1 \begin{bmatrix} X_{11} \\ X_{12} \\ . \\ . \\ . \\ X_{1d} \end{bmatrix} + a_2 \begin{bmatrix} X_{21} \\ X_{22} \\ . \\ . \\ . \\ X_{2d} \end{bmatrix} + \ldots + a_m \begin{bmatrix} X_{m1} \\ X_{m2} \\ . \\ . \\ . \\ X_{md} \end{bmatrix} \\ &= \begin{bmatrix} a_1X_{11} \\ a_1X_{12} \\ . \\ . \\ . \\ a_1X_{1d} \end{bmatrix} + \begin{bmatrix} a_2X_{21} \\ a_2X_{22} \\ . \\ . \\ . \\ a_2X_{2d} \end{bmatrix} + \ldots + \begin{bmatrix} a_mX_{m1} \\ a_mX_{m2} \\ . \\ . \\ . \\ a_mX_{md} \end{bmatrix} \\ &= \begin{bmatrix} a_1X_{11} + a_2X_{21} + \ldots + a_mX_{m1} \\ a_1X_{12} + a_2X_{22} + \ldots + a_mX_{m2} \\ . \\ . \\ . \\ a_1X_{1d} + a_2X_{2d} + \ldots + a_mX_{md} \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} \sum_{i=1}^{m} a_i X_{i1} \\ \sum_{i=1}^{m} a_i X_{i2} \\ . \\ . \\ . \\ \sum_{i=1}^{m} a_i X_{id} \end{bmatrix}$$

$$EX = E \begin{bmatrix} \sum_{i=1}^{m} a_i X_{i1} \\ \sum_{i=1}^{m} a_i X_{i2} \\ . \\ . \\ . \\ \sum_{i=1}^{m} a_i X_{id} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{m} a_i E[X_{i1}] \\ \sum_{i=1}^{m} a_i E[X_{i2}] \\ . \\ . \\ . \\ \sum_{i=1}^{m} a_i E[X_{id}] \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{m} a_i \mu_{i1} \\ \sum_{i=1}^{m} a_i \mu_{i2} \\ . \\ . \\ . \\ \sum_{i=1}^{m} a_i \mu_{id} \end{bmatrix}$$

Hence, the convex combination of multivariate Gaussian will always give us a multivariate Gaussian with each of it's dimensions mixed in the same proportion, given a_i.

We can simply write the $E[\mathbf{X}] = E[a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_m\mathbf{X}_m] = \sum_{i=1}^{m} a_i\boldsymbol{\mu}_i$

$\therefore$ $E[\mathbf{X}]$ has a dimension of $d \times 1$.

(b) [10 MARKS] Write the covariance Cov[$\mathbf{X}$] in terms of the givens $a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$. Show all you steps. What is the dimension of Cov[$\mathbf{X}$]? Briefly explain how the result for Cov[$\mathbf{X}$] would be different if the variables X_1 and X_2 are not independent and have covariance Cov[$\mathbf{X}_1, \mathbf{X}_2$] = $\boldsymbol{\Lambda}$ for $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$.

Answer(b).

We know that,

$$\begin{aligned} Cov[\mathbf{X}] &= Cov[\mathbf{X}, \mathbf{X}] \\ &= E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T] \end{aligned}$$

Let us first calculate $\mathbf{X} - E(\mathbf{X})$,

$$\mathbf{X} - E(\mathbf{X}) = \begin{bmatrix} X_1 & X_2 & . & . & . & X_m \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_m \end{bmatrix} - \begin{bmatrix} \mu_1 & \mu_2 & . & . & . & \mu_m \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_m \end{bmatrix}$$

$$= \begin{bmatrix} X_1 - \mu_1 & X_2 - \mu_2 & . & . & . & X_m - \mu_m \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_m \end{bmatrix} (\because AB - CB = (A - C)B)$$

$$(\mathbf{X} - E(\mathbf{X}))^T = \begin{bmatrix} a_1 & a_2 & . & . & . & a_m \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ . \\ . \\ . \\ X_m - \mu_m \end{bmatrix} (\because (AB)^T = B^T A^T)$$

Now, $(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T =$

$$\begin{bmatrix} X_1 - \mu_1 & X_2 - \mu_2 & . & . & . & X_m - \mu_m \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_m \end{bmatrix} \begin{bmatrix} a_1 & a_2 & . & . & . & a_m \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ . \\ . \\ . \\ X_m - \mu_m \end{bmatrix}$$

$$= \begin{bmatrix} X_1 - \mu_1 & X_2 - \mu_2 & . & . & . & X_m - \mu_m \end{bmatrix} \begin{bmatrix} {a_1}^2 & a_1 a_2 & . & . & . & a_1 a_m \\ a_2 a_1 & {a_2}^2 & . & . & . & a_2 a_m \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_m a_1 & a_m a_2 & . & . & . & {a_m}^2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ . \\ . \\ . \\ X_m - \mu_m \end{bmatrix}$$

$$= \begin{bmatrix} X_1 - \mu_1 & X_2 - \mu_2 & . & . & . & X_m - \mu_m \end{bmatrix} \begin{bmatrix} {a_1}^2(X_1 - \mu_1) + a_1 a_2(X_2 - \mu_2) + ... + a_1 a_m(X_m - \mu_m) \\ a_2 a_1(X_1 - \mu_1) + {a_2}^2(X_2 - \mu_2) + ... + a_2 a_m(X_m - \mu_m) \\ . \\ . \\ . \\ a_m a_1(X_1 - \mu_1) + a_m a_2(X_2 - \mu_2) + ... + {a_m}^2(X_m - \mu_m) \end{bmatrix}$$

=

$$\begin{aligned} &{a_1}^2(X_1 - \mu_1)^2 + a_1 a_2(X_2 - \mu_2)(X_1 - \mu_1) + ... + a_1 a_m(X_m - \mu_m)(X_1 - \mu_1) \\ &+ a_2 a_1(X_1 - \mu_1)(X_2 - \mu_2) + {a_2}^2(X_2 - \mu_2)^2 + ... + a_2 a_m(X_m - \mu_m)(X_2 - \mu_2) \\ &+ ... + a_m a_1(X_1 - \mu_1)(X_m - \mu_m) + a_m a_2(X_2 - \mu_2)(X_m - \mu_m) + ... + {a_m}^2(X_m - \mu_m)^2 \end{aligned}$$

$$Cov[\mathbf{X}] = E[(X - EX)(X - EX)^T]$$

$= {a_1}^2 E[(X_1 - \mu_1)^2] + a_1 a_2 E[(X_2 - \mu_2)(X_1 - \mu_1)] + ... + a_1 a_m E[(X_m - \mu_m)(X_1 - \mu_1)]$
$+ a_2 a_1 E[(X_1 - \mu_1)(X_2 - \mu_2)] + {a_2}^2 E[(X_2 - \mu_2)^2] + ... + a_2 a_m E[(X_m - \mu_m)(X_2 - \mu_2)]$
$+ ... + a_m a_1 E[(X_1 - \mu_1)(X_m - \mu_m)] + a_m a_2 E[(X_2 - \mu_2)(X_m - \mu_m)] + ... + {a_m}^2 E[(X_m - \mu_m)^2]$

$= {a_1}^2 \Sigma_1 + a_1 a_2 \Sigma_{12} + ... + a_1 a_m \Sigma_{m1} + a_2 a_1 \Sigma_{12} + {a_2}^2 \Sigma_2 + ... + a_2 a_m \Sigma_{m2} + ... + a_m a_1 \Sigma_{m1} + a_m a_2 \Sigma_{m2} + ... + {a_m}^2 \Sigma_m$

Since $X_1, X_2, ..., X_m$ are independent our expression will end up being

$Cov[\mathbf{X}] = {a_1}^2 \Sigma_1 + {a_2}^2 \Sigma_2 + ... + {a_m}^2 \Sigma_m$

$= \sum_{i=1}^{m} a_i^2 \Sigma_i$

The dimensions of the covariance matrix would be d x d.

Now, when $Cov[\mathbf{X_1}, \mathbf{X_2}] = \Lambda$

$Cov[\mathbf{X}] = \sum_{i=1}^{m} a_i^2 \Sigma_i + a_1 a_2 \Lambda + a_2 a_1 \Lambda$

$Cov[\mathbf{X}] = \sum_{i=1}^{m} a_i^2 \Sigma_i + 2 a_1 a_2 \Lambda$

Question 3. [10 MARKS]

This question involves some simple simulations, to better visualize random variables and get some intuition for sampling, which is a central theme in machine learning. Use the attached code called `simulate.py`. This code is a simple script for sampling and plotting with python; play with some of the parameters to see what it is doing. Calling `simulate.py` runs with default parameters; `simulate.py 1 100` simulates 100 samples from a 1d Gaussian.

(a) [5 MARKS] Run the code for 10, 100 and 1000 samples with dim=1 and $\sigma = 1.0$. Next run the code for 10, 100 and 1000 samples with dim=1 and $\sigma = 10.0$. What do you notice about the sample mean?

Answer(a).

dim,σ	10 samples	100 samples	1000 samples
1, 1.0	0.0198663747549	0.143975586371	-0.0431212935257
1, 10.0	4.02051073576	-0.139705315861	0.293219870959

As we increase the size of our samples, the sample mean starts to converge to the population mean. In our case the mean approaches zero. From the formula of sample variance $\sigma/\sqrt{(N)}$, we can also conclude that the variance starts decreasing as N increases.

(b) [5 MARKS] The current covariance for dim=3 is

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

What does that mean about the multivariate Gaussian (i.e., about X, Y and Z)?

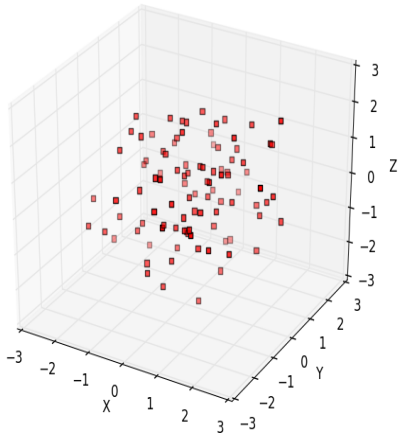
Change the covariance to

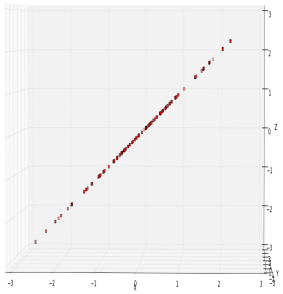
$$\Sigma = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

What happens?

Answer(b). Firstly, we observe that X,Y,Z are independent and each of them has a variance of 1. Also, from the scatter plot we see that the points are scattered all around and thus gives us an idea that they are not correlated, hence independent. Secondly, when we change the covariance

matrix, we are making X and Z correlated by introducing Cov[X,Z] = 1. From the scatter plot we see that X and Z are positively correlated with $\rho = 1.0$, since points are in straight line.





Question 4. [30 MARKS]

Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with an unknown mean λ. Based on previous experience in similar industrial plants, suppose that our initial feelings about the possible value of λ can be expressed by an exponential distribution with parameter $\theta = \frac{1}{2}$. That is, the prior density is

$$f(\lambda) = \theta e^{-\theta\lambda}$$

where $\lambda \in (0, \infty)$.

[Attached Python Code]

(a) [5 MARKS] Before observing any data (any reported accidents), what is the most likely value for λ?

Answer(a). Since, we don't have any other information except the prior we need to find the value of λ which will give the maximum prior probability.

so we have,

$$\begin{aligned} f(\lambda) &= \theta e^{-\theta\lambda} \\ &= \frac{\theta}{e^{\theta\lambda}} \end{aligned}$$

To maximize this value we need to minimize the denominator. So the most likely value of λ will be zero. and $f(\lambda) = \theta = \frac{1}{2}$

(b) [5 MARKS] Now imagine there are 79 accidents over 9 days. Determine the maximum likelihood estimate of λ.

Answer(b).

To find the maximum likelihood estimate of λ we have,

$$\lambda_{ML} = \arg\max_{\lambda \in (0,\infty)} \{p(D|\lambda)\}$$
$$p(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} e^{-n\lambda}}}{\prod_{i=1}^{n} x_i!}$$

We know that, the posterior probability is propotional to

$$\begin{aligned} p(\lambda|D) &\propto p(D|\lambda) \\ ln\, p(\lambda|D) &\propto ln\, p(D|\lambda) \\ &= \sum_{i=1}^{n} x_i\, ln\, \lambda - n\lambda - \sum_{i=1}^{n} ln\, x_i! \end{aligned}$$

Taking the partial derivative w.r.t λ

$$\frac{\partial ll}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

Now equating, $\frac{\partial ll}{\partial \lambda} = 0$ we get,

$$\begin{aligned} \frac{\sum_{i=1}^{n} x_i}{\lambda_{ML}} - n &= 0 \\ \frac{\sum_{i=1}^{n} x_i}{\lambda_{ML}} &= n \\ \lambda_{ML} &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

$\sum_{i=1}^{9} x_i = 79$ and $n = 9$
we get,
$\lambda_{ML} = \frac{79}{9} = 8.778$

(c) [5 MARKS] Again imagine there are 79 accidents over 9 days. Determine the maximum a posteriori (MAP) estimate of λ.

Answer(c).

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(D|\lambda)p(\lambda)\}$$
$$p(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} e^{-n\lambda}}}{\prod_{i=1}^{n} x_i!}$$
$$p(\lambda) = \theta e^{-\lambda\theta}$$

We know that, the posterior probability is given by

$$p(\lambda|D) \propto p(D|\lambda)p(\lambda)$$

$$
\begin{aligned}
ln\, p(\lambda|D) &\propto ln\, p(D|\lambda) + ln\, p(\lambda) \\
&= \sum_{i=1}^{n} x_i\, ln\, \lambda - n\lambda - \sum_{i=1}^{n} ln\, x_i! + ln\, \theta - \lambda\theta
\end{aligned}
$$

Taking the partial derivative w.r.t λ

$$
\frac{\partial ll}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n - \theta
$$

Now equating, $\frac{\partial ll}{\partial \lambda} = 0$ we get,

$$
\begin{aligned}
\frac{\sum_{i=1}^{n} x_i}{\lambda_{MAP}} - n - \theta &= 0 \\
\frac{\sum_{i=1}^{n} x_i}{\lambda_{MAP}} &= n + \theta \\
\lambda_{MAP} &= \frac{\sum_{i=1}^{n} x_i}{n + \theta}
\end{aligned}
$$

$\sum_{i=1}^{9} x_i = 79$, $n = 9$ and $\theta = \frac{1}{2}$
we get,
$\lambda_{MAP} = \frac{79}{9+\frac{1}{2}} = \frac{79}{9.5} = 8.316$

(d) [5 MARKS] Imagine you now want to predict the number of accidents for tomorrow. How can you use the maximum likelihood estimate computed above? What about the MAP estimate? What would they predict?

Answer(d). The maximum likelihood estimate of λ can be used to find the input at which the likelihood function attains the maximum value. It is a representation of how well the In our predictor, it will strengthen the posterior probability from our distribution. On the other hand, the MAP estimate will help us in finding the most probable model that fits our data. We can calculate the goodness of fit using Pearson's chi-squared test or likelihood ratio test. This will give us confidence of how proper our model fits the observed data.

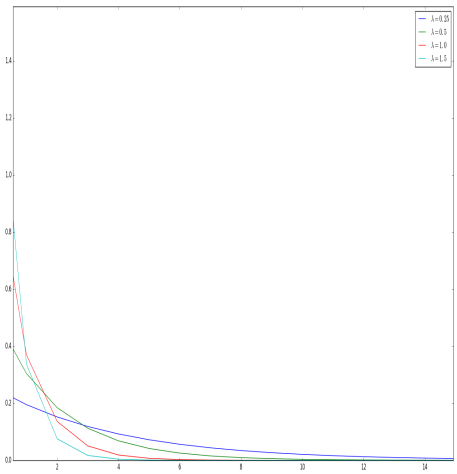
The λ_{ML} will predict that there will 8.778 accidents on a average tomorrow and λ_{MAP} will predict that there will be 8.316 accidents on an average tomorrow with highest probability. The posterior probability calculated from MAP will strongly depend on the value of θ. As you will see in the subsection (f) of this question.

(e) [5 MARKS] For the MAP estimate, what is the purpose of the prior once we observe this data?

Answer(e). Prior is the information that we know before we take any observations into the account. Once we have the data, prior is used to avoid over fitting our hypothesis/model about the data's distribution. The stronger our prior the strongly it will pull the posterior probability towards itself.

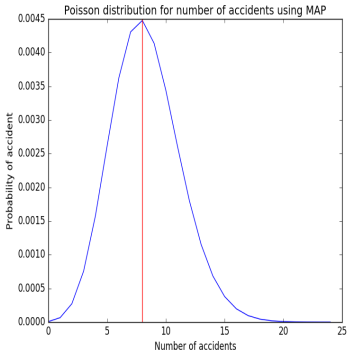
(f) [5 MARKS] Look at the plots of some exponential distributions to better understand the prior chosen on λ. Imagine that now new safety measures have been put in place and you believe that the number of accidents per day should sharply decrease. How might you change θ to better reflect this new belief about the number of accidents?

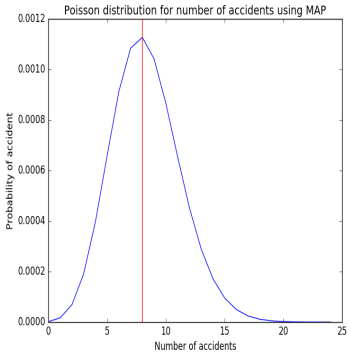
Answer(f). Let's plot some exponential distribution for different values of θ.

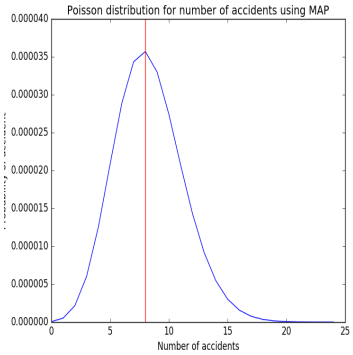


The x-axis represents the number of accidents and the y-axis represents the probability of happening of the accidents.

In my opinion, after including safety measures in order to depict the decrease in number of accidents everyday we should increase the value of θ because as you start to increase the value of theta, the posterior probability will start to decrease and hence the probability of accidents happening will decrease sharply. Here are some of the plots for different values of θ. diagrams in order theta = [0.25,0.5,1.0]







Check the peak of the distribution corresponding to the λ_{ML} is decreasing with increase in value of θ

Question 5. [20 MARKS]

Imagine that you would like to predict if your favorite table will be free at your favorite restaurant. The only additional piece of information you can collect, however, is if it is sunny or not sunny. You collect paired samples from visit of the form (is sunny, is table free), where it is either sunny (1) or not sunny (0) and the table is either free (1) or not free(0).

(a) [10 MARKS] How can this be formulated as a maximum likelihood problem?

Answer(a). In the given problem , we are trying to make a predictor based on given the evidence of weather being sunny what's the probability that the table is free. To formulate this as a maximum likelihood problem we define following random variables.

Let X_1 = If it's sunny or not and let X_2 = if the table is free or not.Clearly, $X_1, X_2 \in \{0, 1\}$.

Also, let $\hat{\theta}$ be the point estimator of θ. Then, we can write

$$\hat{\theta} = g(X_1, X_2)$$

Since, X_1, X_2 are i.i.d. , we define it's pdf as $f(x;\theta)$. Now, we can write the likelihood function as, Let, $\mathbf{D} = (x_{1i}, x_{2i})$ be the dataset for i = 1 ... n.

$$L(\theta) = \prod_{i=1}^{n} f(X_1 = x_{1i}, X_2 = x_{2i}; \theta)$$

Now from Bayes theorem,

$$P(X_2|X_1) = \frac{P(X_1|X_2)P(X_2)}{P(X_1)}$$

Once we have the data, we can easily estimate $P(X_2)$ and $P(X_1)$ and the problem reduces to finding the likelihood function because these value are constant and we drop these value from our calculations. And in order to find the maximum value for which we have maximum value of likelihood function we find argmax for the parameter θ. Therefore, this problem is just a maximum likelihood problem. Given by:

$$\begin{aligned}\hat{\theta}_{ML} &= \arg\max_{X_2}\{P(X_1|X_2;\theta)\} \\ &= \arg\max_{X_2} L(\theta)\end{aligned}$$

Here, the parameter values are $P(X_2)$ and $P(X_1|X_2)$.

After estimation of parameter we can use this value to maximize and likelihood function and use the data to find the prior and the marginal probability. Thus giving us all the probabilities to find the posterior probability.

Now, since $X_1, X_2 \in 0, 1$, they both will have a bernoulli distribution. and the joint probability of X_1, X_2 will be a binomial distribution.

$$\begin{aligned}L(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{(1-\sum_{i=1}^{n} x_i)} \\ &= p^{y}(1-p)^{(n-(y))}\end{aligned}$$

Now, taking log of the likelihood we get

$$logL(p) = (y)log(p) + (n-(y))log(1-p)$$

differentiating it w.r.t. p and equating to zero. we get,

$$\frac{\partial}{\partial p}(y)log(p) + (n-(y))log(1-p) =$$

$$= \frac{(y)}{p} - \frac{(n-(y))}{(1-p)} = 0$$

$$\frac{y}{p} - \frac{n-y}{(1-p)} = 0$$
$$\frac{y}{p} = \frac{n-y}{(1-p)}$$
$$y - yp = np - yp$$
$$p = \frac{y}{n}$$

(b) [5 MARKS] Assume you have collected data for the last 10 days and computed the maximum likelihood solution to the problem formulated in (a). If it is sunny today, how would you predict if your table will be free?

Answer(b). Now that we have collected the data we can find the posterior probabilities for $X_2 = 0$ and $X_2 = 1$,

$$P(X_2 = 0|X_1 = 1; \hat{p}_{ML}) = P(X_1 = 1|X_2 = 0)P(X_2 = 0) \quad (1)$$
$$P(X_2 = 1|X_1 = 1; \hat{p}_{ML}) = P(X_1 = 1|X_2 = 1)P(X_2 = 1) \quad (2)$$

Out of the two posterior probability, the one which is greater will be the class of the given data point.

(c) [5 MARKS] Imagine now that you could further gather information about if it is morning, afternoon, or evening. How does this change the maximum likelihood problem?

Answer(c). We define new random variables X_3, X_4, X_5 representing if it's morning, afternoon and evening respectively. Now are we include more Bernoulli random variables we will end up with a Multinomial joint distribution.

Now our problem of maximum likelihood will include finding the parameters (say, p_1, p_3, p_4, p_5) for all the underlying conditional probability distributions. We need to find MLE solution for multinomial joint distribution. The joint probability can be expressed as a gamma function where, C is a constant.

$$L(p_1, p_3, p_4, p_5) = C{p_1}^{x_1}{p_3}^{x_3}{p_4}^{x_4}{p_5}^{x_5}$$
$$logL(p_1, p_3, p_4, p_5) = logC + x_1 logp_1 + x_3 logp_3 + x_4 logp_4 + x_5 logp_5$$

We solve this log likelihood to find the maximum likelihood of the function parameters.

Question 6. [15 MARKS]

Suppose you have three coins. Coin A has a probability of heads of 0.75, Coin B has a probability of heads of 0.5, and Coin C has a probability of heads of 0.25.

(a) [5 MARKS] Suppose you flip all three coins at once, and let X be the number of heads you see (which will be between 0 and 3). What is the expected value of X, $E[X]$?

Answer(a). Given,
Coin A : P(H) = 0.75 , P(T) = 0.25

Coin B : P(H) = 0.5 , P(T) = 0.5
Coin C : P(H) = 0.25 , P(T) = 0.75
$\Omega = \{TTT, TTH, THT, THH, HTT, HTH, HHT, HHH\}$
Let, X = Number of Heads
P(X = 0) = $\frac{1}{8}$ P(X = 1) = $\frac{3}{8}$ P(X = 2) = $\frac{3}{8}$ P(X = 3) = $\frac{1}{8}$

$$\begin{aligned} E[X] &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\ &= 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} \\ &= \frac{12}{8} \\ &= 1.5 \end{aligned}$$

(b) [10 MARKS] Suppose instead you put all three coins in your pocket, select one at random, and then flip that coin 5 times. You notice that 3 of the 5 flips result in heads while the other 2 are tails. What is the probability that you chose Coin C?

Answer(b). Let X be the random variable denoting number of heads and let Y be the random variable denoting number of tails. We calculate, using Bayes Theorem.

$$\begin{aligned} P(C|X=3, Y=2) &= \frac{P(X=3, Y=2|C)P(C)}{P(X=3, Y=2)} \\ &= \frac{1}{\binom{3}{1}} \times \frac{\binom{5}{3}(0.25)^3(0.75)^2}{\binom{5}{3}\frac{1}{\binom{3}{1}}[(0.75)^3(0.25)^2 + (0.5)^3(0.5)^2 + (0.25)^3(0.75)^2]} \\ &= \frac{0.00879}{0.06641} \\ &= 0.13236 \end{aligned}$$