# Homework Assignment #1

## Answer to the Question #1

a) $E[f(x)] = \sum_{i=1}^{n} f(x_i)p(x_i)$
$= f(a)p(a) + f(b)p(b) + f(c)p(c)$
$= 10 * 0.1 + 5*0.2 + (10/7) * 0.7$
$= 1+1+1$
$=3$

b) $E\left[\frac{1}{p(x)}\right] = \frac{1}{p(a)} * p(a) + \frac{1}{p(b)} * p(b) + \frac{1}{p(c)} * p(c)$
$= \frac{1}{0.1} * 0.1 + \frac{1}{0.2} * 0.2 + \frac{1}{0.7} * 0.7$
$= 1+1+1$
$=3$

c) $E\left[\frac{1}{p(x)}\right] = \sum_{i=1}^{n} \frac{1}{p(x_i)} * p(x_i)$
$= \sum_{i=1}^{n} 1$
$=n$

## Answer to the Question #2

a) $E[X] = E[a_1X_1 + a_2X_2 + \cdots a_mX_m]$
$= E[a_1X_1] + E[a_2X_2] + \cdots + E[a_mX_m]$
$= a_1E[X_1] + a_2E[X_2] + \cdots a_mE[X_m]$
$= a_1\begin{bmatrix} X_{11} \\ X_{12} \\ X_{1d} \end{bmatrix} + a_2E\begin{bmatrix} X_{21} \\ X_{22} \\ X_{2d} \end{bmatrix} + \cdots + a_mE\begin{bmatrix} X_{m1} \\ X_{m2} \\ X_{md} \end{bmatrix}$
$= a_1\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{1d} \end{bmatrix} + a_2\begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{2d} \end{bmatrix} + \cdots + a_m\begin{bmatrix} \mu_{m1} \\ \mu_{m2} \\ \mu_{md} \end{bmatrix}$
$= \begin{bmatrix} a_1\mu_{11} \\ a_1\mu_{12} \\ a_1\mu_{1d} \end{bmatrix} + \begin{bmatrix} a_2\mu_{21} \\ a_2\mu_{22} \\ a_2\mu_{2d} \end{bmatrix} + \cdots + \begin{bmatrix} a_m\mu_{m1} \\ a_m\mu_{m2} \\ a_m\mu_{md} \end{bmatrix}$
$= \begin{bmatrix} a_1\mu_{11}+a_2\mu_{21} & +a_m\mu_{m1} \\ a_1\mu_{12}+a_2\mu_{22} & \ldots +a_m\mu_{m2} \\ a_1\mu_{1d}+a_2\mu_{2d} & +a_m\mu_{md} \end{bmatrix}$
$= \begin{bmatrix} \sum_{i=1}^{m} a_i\mu_{i1} \\ \sum_{i=1}^{m} a_i\mu_{i2} \\ \sum_{i=1}^{m} a_i\mu_{id} \end{bmatrix}$
$= \sum_{i=1}^{m} a_i \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{id} \end{bmatrix}$
$= \sum_{i=1}^{m} a_i\, \mu_i$

b) $Cov\,[X_1 + X_2 + \cdots + X_m]$
$= \sum_{i=1}^{m} \sum_{j=1}^{m} Cov\,[X_i\, X_j]$
$= \sum_{i=1}^{m} V[X_i] + 2 \sum Cov\,[X_i\, X_j]$
if $X_i$ and $X_j$ are independent $Cov\,[X_i\, X_j] = 0$
$Cov\,[X] = \sum_{i=1}^{m} V[\,a_i X_i] + 2 \sum Cov\,[\,a_i X_i,\, a_j X_j]$

As $X_i$ and $X_j$ are independent $Cov\,[\,a_i X_i\,,\, a_j X_j] = 0$
$Cov\,[X] = \sum_{i=1}^{m} V[\,a_i X_i]$
$= \sum_{i=1}^{m} a_i^2 V[X_i]$
$= \sum_{i=1}^{m} a_i^2\, \Sigma_i$

Answer to the Question #3

a) Running the code for 10,100,1000 samples with dim=1 and σ = 1.0 and observing the outputs we can see that, the sample mean is going nearer to zero with the increase of sample size.
Next running the code for 10,100,1000 samples with dim=1 and σ = 10.0 and observing the outputs we can see that, the sample mean is going nearer to zero with the increase of sample size. And the samples go closer to other samples with the increase of sample size.
We can also observe that, the samples are more scattered when σ gets large. Which means the samples goes closer to the mean when sample size increases and σ decreases.

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

$\Sigma_{ij} = \begin{cases} 1\ if\ i = j \\ 0\ otherwise \end{cases}$
$\Sigma_{11} = 1$
This means positive covariance.

Cov[X,Y] = $\Sigma_{12} = 0$
$Cov\,[X, Y] = E[XY] - E[X]E[Y]$
$\Rightarrow E[XY] = E[X]E[Y]$
This means zero covariance i.e. independent.
The meaning of the covariance matrix is, the dimensions (Vector random variables X, Y, Z) are independent of each other.

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

$\Sigma_{13} = \Sigma_{31} = 1$
This means positive covariance.

$\Sigma_{12} = 0$
This means zero covariance i.e. independent.

The meaning of the covariance matrix is, the Vector random variables X and Z are positively correlated to each other and Y is independent of X and Z.

Answer to the Question #4

a) Before observing any data the likely value of λ to be distributed exponentially. We need to estimate the λ that maximizes the likelihood of $P(\lambda)$ to be exponentially distributed.

$P(\lambda|\theta) = \theta e^{-\theta\lambda}$
$lnP(\lambda|\theta) = \ln(\theta e^{-\theta\lambda})$
$lnP(\lambda|\theta) = \ln\theta - \theta\lambda$

Taking the first derivative
$\frac{d}{d\theta}(\ln P(\lambda|\theta) = \frac{d}{d\theta}\ln\theta + \frac{d}{d\theta}(-\theta\lambda)$

Equating the derivative to zero we get
$\frac{1}{\theta} = \lambda$
$\lambda = \frac{1}{\frac{1}{2}}$
$\lambda = 2$
Implies before observing any data the most likely value of λ is 2.

b) We have,
$\sum_{i=1}^{n} x_i = 79$ and $n = 9$

$\lambda_{MLE} = \underset{\lambda \in (0,\alpha)}{argmax} P(D|\lambda)$

We can write the likelihood function as:
$P(D|\lambda) = P(\{x_i\}_{i=1}^{n}|\lambda)$
$P(D|\lambda) = \prod_{i=1}^{n} P(x_i|\lambda)$
$P(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} * e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$
$ll(D,\lambda) = \ln P(D|\lambda)$
$= \ln\left(\frac{\lambda^{\sum_{i=1}^{n} x_i} * e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}\right)$
$= \ln\lambda^{\sum_{i=1}^{n} x_i} + \ln e^{-n\lambda} - \ln\prod_{i=1}^{n} x_i!$
$= \sum_{i=1}^{n} x_i \ln\lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i)!$

Taking the first derivative of log likelihood and equating to zero we get:
$\frac{dll(D|\lambda)}{d\theta} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$
$\frac{1}{n}\sum_{i=1}^{n} x_i = \lambda$
Substituting n=9 and $\sum_{i=1}^{n} x_i = 79$ we get

$\lambda_{MLE} = \frac{79}{9} = 8.78$

This is the most likely value of λ given data.

c) First, we write the probability density function of exponential distribution as:

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

The MAP estimate parameters can be found as

$\lambda_{MAP} = \underset{\lambda \in (0,\alpha)}{argmax} P(D|\lambda)P(\lambda)$

We can write the likelihood function as

$P(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} * e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$

And the prior distribution as

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

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

$\ln P(\lambda|D) \propto \ln P(D|\lambda) + \ln P(\lambda)$

Taking the first derivative and equating it to zero we get

$\frac{dll(\lambda|D)}{d\lambda} = \ln\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln x_i + \lambda \ln\theta - \theta\lambda$

$0 = \sum_{i=1}^{n} x_i \frac{1}{\lambda} - n + 0 + 0 - \theta$

$n + \theta = \sum_{i=1}^{n} x_i \frac{1}{\lambda}$

$\lambda_{MAP} = \frac{1}{n+\theta} \sum_{i=1}^{n} x_i$

Substituting the values of $\sum_{i=1}^{n} x_i$ and n we get MAP estimate of λ

$\lambda_{MAP} = \frac{79}{9.5} = 8.315$

d) Let X= $\{x_1, x_2, \ldots x_n\}$ is the Random variable that represents the number of accident.

$E[X] = \sum_{i=1}^{n} x_i \; p(x_i)$

$E[X] = \sum_{i=1}^{n} x_i \left(\frac{\lambda_{MLE}^{x} * e^{-n\lambda_{MLE}}}{x!}\right)$

We can use $\lambda_{MLE}$ to compute the probability of x accident tomorrow and thus the expected value of the number of accidents tomorrow.

Probability of accident using MAP estimation

$P(X) = \frac{\lambda^{\sum_{i=1}^{n} x_i} * e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$

Where X is a random variable represents the event of accident and $x_i$ is the number of accident in $i^{th}$ day. Both estimates will predict the number of accident in one day.

e) To prefer one model to another the prior is needed. Even once the data being observed using extra information of prior can be used for fine-tuning of MLE estimate.
As P(λ) is a prior that is found from another industries, so it is a good prior. It gives a weight to λ so that λ goes nearer to actual $\lambda^*$ faster.

f) From Q#1 we can see that,
$\lambda = \frac{1}{\theta}$
$\lambda = \frac{1}{\frac{1}{2}} = 2 \; when \; \theta = \frac{1}{2}$
$\lambda = \frac{1}{1} = 1 \; when \; \theta = 1$
$\lambda = \frac{1}{\frac{1}{4}} = 4 \; when \; \theta = \frac{1}{4}$
Which means λ and θ are inversely proportional (i.e. if λ goes down θ should go up). So to represent the sharp decrease of the number of accident per day θ should be increased in order to decrease λ

Answer to the Question #5

a) To formulate the problem we will use two random variable S and F to represent the Day(sunny or not sunny) and Table (free or not free):
As S takes two values we can consider P(S) a Bernoulli distribution.
*P(S,a) = a*s + (1-a)(1-s) for S = {0,1}*

Similarly the probability distribution of random variable F can also be expressed as Bernoulli distribution. Here F should be expressed as conditional distribution as the probability of the table to be free depends on S.

| F = {0,1} | P(F\|S=1) | P(F\|S=0) |
|---|---|---|
| Free (F=1) | *b* | c |
| Not Free(F=0) | *1-b* | *1-c* |

We can formulate the likelihood function as:
$\lambda_{MLE} = \; _{argmax} P(D|\lambda)$
here λ={*a,b,c*}

$P(D|\lambda) = \prod_{i=1}^{n} P\,(F,S|\lambda) = \prod_{i=1}^{n} P\,(F|S,\lambda)\,P(S)$
Here our goal is to estimate the unknown parameters $a_{MLE}$, $b_{MLE}$, $c_{MLE}$

b) We have
$a_{MLE}$, $b_{MLE}$, $c_{MLE}$ and
P(F,S) for 10 days

We will calculate $P(F_1,S_1)$ to predict if its sunny today the will the table be free.
$P(F_1,S_1) = P(F_1|S_1)\,P(S_1)$
= $b^* a$

If the $P(F_1,S_1) > 0.5$ the table will be predicted as free otherwise the table is not free.

c) Now we have another information about the time of the day. A random variable T represents the time of a day.

$$T = \begin{cases} 0 & when\ T = morning \\ 1 & when\ T = afternoon \\ 2 & when\ T = evening \end{cases}$$

We can assume P(T) to be a uniform distribution with pmf = 1/3

Now the *cocnditional* probability table will be changed as

| Sunny (S) | Time (T) | P(F=1) | P(F=0) |
|---|---|---|---|
| 1 | 0 | *a* | *1-a* |
| 1 | 1 | *b* | *1-b* |
| 1 | 2 | *c* | *1-c* |
| 0 | 0 | *d* | *1-d* |
| 0 | 1 | *e* | *1-e* |
| 0 | 2 | *f* | *1-f* |

Now the likelihood function will be changed as

$\lambda_{MLE} = \ _{argmax} P(D|\lambda)$

here $\lambda = \{a,b,c,d,e,f\}$

$P(D|\lambda) = \prod_{i=1}^{n} P\ (F,S,T|\lambda)$

Our goal is to estimate the unknown parameter $\lambda_{MLE}$.

d) We have generated 1000 samples from a d-dimensional multivariate Gaussian distribution with mean 0 and identity covariance matrix and computed the average distance of each point to the origin. We have repeated the experiment for the given dimensions and observed the result.

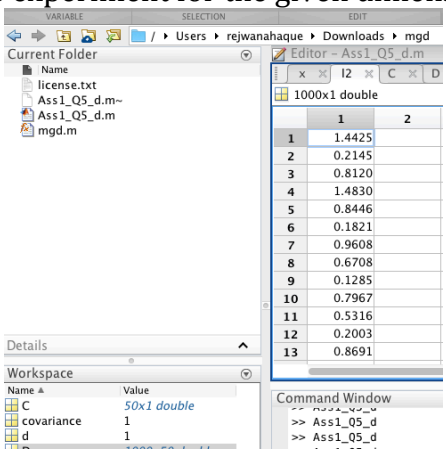


Fig1: Dimension 1

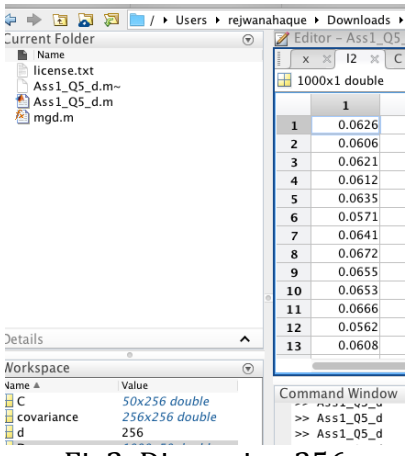


Fig2: Dimension 256

Seeing the distance we can observe that, when the dimension increases the average distance of each sample to the origin decreases (goes near zero) and becomes almost equal.

For k-means clustering, when the dimension increases the distance from each sample to the centroid of each cluster becomes similar. Which means the gap of one cluster to another becomes smaller.

e) We have,
Dimension = d
First hypercube side = 1
Second hypercube side = 1- ε ,where 0≤ε≤1

Now
Volume of First hypercube $Volume_1 = 1^d$
Volume of Second hypercube $Volume_2 = (1- \varepsilon)^d$

Ratio of volume of second hypercube to the ratio of the volume of first hypercube = $Volume_2 / Volume_1$
$= (1- \varepsilon)^d / 1^d$
$= (1- \varepsilon)^d$

As the ratio is exponential function with base $(1- \varepsilon) \leq 1$, the ratio will decrease exponentially with the increase of d.
From the formula of average distance $(\sqrt[d]{\sum_{i=1}^{n}(x_i - c)^d})/d$, we can see $(x_i - c)^d$ as the volume of a hypercube and the ratio can be considered as the ratio of two hypercube.
Here the denominator increases with the increase of d and the numerator decreases. So we can conclude, "If the distance is small the distance decreases with the increases of d." This completely agrees with the result of hypercube volume ratio.