CMPUT 551 Homework Assignment #1

Name: Amir Samani
Student ID: 1476213
CCID: samani
Date: 26 September 2017
Contributers: Pouneh Gorji, Farzane Aminmansour

Question 1

(a)

$$E[f(X)] = \sum_{x\in\Omega} f(x)p(x) = f(a)p(a) + f(b)p(b) + f(c)p(c) = (10)(0.1) + (5)(0.2) + (\frac{10}{7})(0.7) = 3$$

(b)

$$E[1/p(X)] = \sum_{x\in\Omega} \frac{1}{p(x)}p(x) = \frac{1}{p(a)}p(a) + \frac{1}{p(b)}p(b) + \frac{1}{p(c)}p(c) = 1 + 1 + 1 = 3$$

(c)

$$E[1/p(X)] = \sum_{x\in\Omega} \frac{1}{p(x)}p(x) = \frac{1}{p(a)}p(a) + \frac{1}{p(b)}p(b) + \frac{1}{p(c)}p(c) = 1 + 1 + 1 = 3$$

If the outcome space size changes, then this value would be n instead of 3, in which n is the outcome space size.

Question 2

(a)

Since all X_i have the same dimension d we can have:

$$E[X] = E[\sum_{i=1}^{m} a_i X_i] = \sum_{i=1}^{m} E[a_i X_i] = \sum_{i=1}^{m} a_i E[X_i] = \sum_{i=1}^{m} a_i \mu_i$$

The dimension of $E[X]$ will also be d as $\sum_{i=1}^{m} a_i\mu_i \in \mathbb{R}^d$.

(b)

$$\begin{aligned} Cov[X] &= Cov[\sum_{i=1}^{m} a_i X_i] = \sum_{i=1}^{m}\sum_{j=1}^{m} Cov[a_i X_i, a_j X_j] = \sum_{i=1}^{m} V[a_i X_i] + 2\sum_{1\le i<j\le m} Cov[a_i X_i, a_j X_j] \\ &= \sum_{i=1}^{m} a_i^2 V[X_i] + 2\sum_{1\le i<j\le m} a_i a_j Cov[X_i, X_j] \end{aligned}$$

Since $X_1, ..., X_m$ are independent, $Cov[X_i, X_j] = 0$ when $i \neq j$. So we have:

$$Cov[X] \quad = \quad \sum_{i=1}^{m} a_i^2 V[X_i] = \sum_{i=1}^{m} a_i^2 \Sigma_i$$

The dimension of $Cov[X]$ will also be $d \times d$ as $\sum_{i=1}^{m} a_i^2 \Sigma_i^2 \in \mathbb{R}^{d \times d}$. If X_1 and X_2 are not independent with $Cov[X_1, X_2] = \Lambda$, then we need to add the non-zero covariance of X_1 and X_2 to the summation for $Cov[X]$. So we have:

$$Cov[X] \quad = \quad \sum_{i=1}^{m} a_i^2 V[X_i] + 2a_1 a_2 Cov[X_1, X_2] = \sum_{i=1}^{m} a_i^2 \Sigma_i + 2a_1 a_2 \Lambda$$

Question 3

(a)

With higher variance (σ), the mean of the samples that we get from a Gaussian distribution is less likely to be close to the real model mean. To clarify, with a relatively low variance like $\sigma = 1$, we can have a proper estimation of underlying distribution mean, even with a small number of samples (such as 10 or 100). On the other hand, a relatively high variance like $\sigma = 10$, makes the mean estimation from a small number of samples significantly different from the mean from underlying distribution. The experiment shows that when the underlying distribution has high variance, we need to have more samples to have a accurate estimation of the sample mean. In other words, we can reduce the impact of variance by having more samples.

(b)

The covariance matrix (Σ) shows the covariance of each individual random variable (*i.e.*, X,Y, and Z) in a multivariate random variable.

$$\Sigma = \begin{bmatrix} V[X] & Cov[X,Y] & Cov[X,Z] \\ Cov[Y,X] & V[Y] & Cov[Y,Z] \\ Cov[Z,X] & Cov[Z,Y] & V[Z] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

As indicated in the covariance matrix Σ, random variables X, Y, and Z are independent, since $Cov[X, Y] = 0$, $Cov[X, Z] = 0$, and $Cov[Y, Z] = 0$.

(c)

Changes in the covariance matrix Σ are as follows:

$$\Sigma = \begin{bmatrix} V[X] & Cov[X,Y] & Cov[X,Z] \\ Cov[Y,X] & V[Y] & Cov[Y,Z] \\ Cov[Z,X] & Cov[Z,Y] & V[Z] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

As showed in the covariance matrix, $Cov[X, Z] = 1$. This result in the following:

$$Corr[X, Z] = \frac{Cov[X, Z]}{\sqrt{V[X]}\sqrt{V[Z]}} = \frac{1}{1 \times 1} = 1$$

The strong positive correlation between two random variables X and Z is illustrated in Figure 1. To shed light on the matter, by changing $Cov[X, Z] = -1$, we have $Corr[X, Z] = -1$. Now we have a strong negative correlation as illustrated in Figure 2.

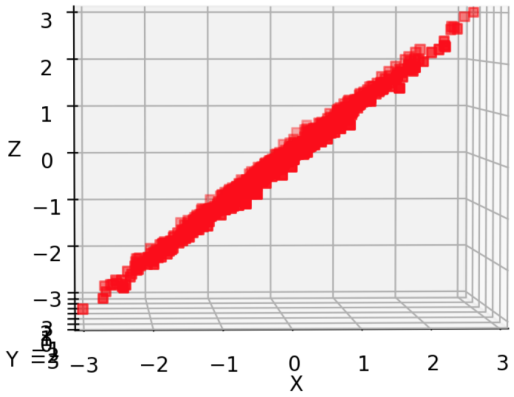

Figure 1: Samples from a multivariate random variable (X,Y,Z) with strong positive correlation between X and Z

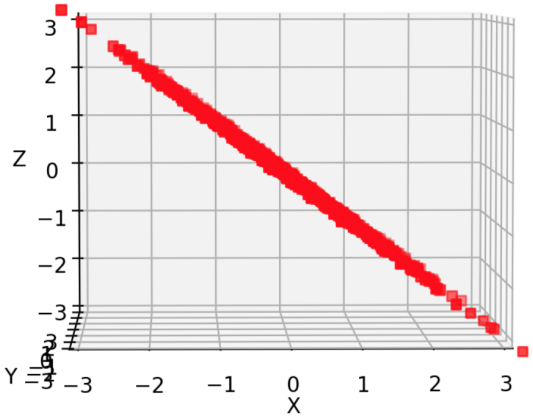

Figure 2: Samples from a multivariate random variable (X,Y,Z) with strong negative correlation between X and Z

Question 4

(a)

Since our initial feelings about the possible value of λ is said to be expressed by the exponential distribution and the mode of exponential distribution is 0. Most likely value for λ is 0.

(b)

The input data is $\mathcal{D}$, the sum of data is 79 and we have 9 instances of data entries. We can find the maximum likelihood estimation for λ in the following:

$$\lambda_{ML} = \arg\max_{\lambda}\{p(\mathcal{D}|\lambda)\}$$

The $\mathcal{D} = \{x_1, ..., x_9\}$ and these x_i instances are (believed) i.i.d samples from a Poisson distribution with parameter λ. So we can rewrite $p(\mathcal{D}|\lambda)$ in the following:

$$p(\mathcal{D}|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda)$$

Since we are using Poisson distribution for MLE, the probability of each x_i can be calculated by $p(x_i|\lambda) = \frac{\lambda^{x_i}e^{-\lambda}}{x_i!}$. Now we can rewrite $p(\mathcal{D}|\lambda)$ in the following:

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

In order to maximize (optimize) this *likelihood* function, we need to take derivative for the function with respect to λ, however, taking log from it and forming *log-likelihood* makes our calculation simpler.

$$ll(\lambda) = log(\frac{\lambda^{\sum_{i=1}^{n} x_i}e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}) = log(\lambda)\sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} log(x_i!)$$

Now by taking derivative from the *log-likelihood* function with respect to λ and solving it for value 0, we can find the optimal λ.

$$\frac{\partial ll}{\partial \lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

For our problem setting, we know that the sum of number of accidents over 9 days is 79, so $\lambda = \frac{79}{9}$. We can also confirm that the value we find is actually the maximum and not the minimum by taking the second derivative of the $ll(\lambda)$ function.

$$\frac{\partial^2 ll}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

This value is always negative. So the optimal value for λ we found is maximum.

(c)

Similar to part (b), the input data is $\mathcal{D}$, the sum of data is 79 and we have 9 instances of data entries. We can find the maximum likelihood estimation for λ in the following:

$$\lambda_{MAP} = \arg\max_{\lambda}\{p(\mathcal{D}|\lambda)p(\lambda)\}$$

Since we have prior knowledge of $p(\lambda) = \theta e^{-\theta\lambda}$, we can rewrite $p(\mathcal{D}|\lambda)p(\lambda)$ in the following:

$$p(\mathcal{D}|\lambda)p(\lambda) = \prod_{i=1}^{n} p(x_i|\lambda)\theta e^{-\theta\lambda}$$

Now we rewrite $p(\mathcal{D}|\lambda)p(\lambda)$ with the knowledge that each sample is from a Poisson distribution and the prior is known to be an exponential distribution:

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

Now, in order to maximize (optimize) this function, we need to take its derivative with respect to λ. However, to make these calculations simpler we can first take *log* of this function and then take the derivative.

$$log(p(\mathcal{D}|\lambda)p(\lambda)) = log(\frac{\lambda^{\sum_{i=1}^{n} e^{-n\lambda}}}{\prod_{i=1}^{n} x_i!}\theta e^{-\theta\lambda}) = log(\lambda)\sum_{i=1}^{n} x_i - (n+\theta)\lambda - \sum_{i=1}^{n} log(x_i!) + log(\theta)$$

Now we can take the derivative of this function in the following with respect to λ.

$$\frac{\partial log(p(\mathcal{D}|\lambda)p(\lambda))}{\partial\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - (n+\theta) = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n+\theta}$$

For our problem setting, we know that the sum of number of accidents over 9 days is 79 and the θ is believed to be $\frac{1}{2}$, so $\lambda = \frac{79}{9+\frac{1}{2}}$. The second derivative is the same as the maximum likelihood estimation, so we know the value we found is maximum.

(d)

The prediction of MLE and MAP estimation will be the $\arg\max_k \lambda_{ML}(k)$ and $\arg\max_k \lambda_{MAP}(k)$. Since these are the Poisson distributions, the *argmax* will be the mode of each distributions, which is $\lfloor\lambda\rfloor$ or $\lceil\lambda\rceil - 1$. We can check the values around this point just to make sure, in our case 8 is the best value for both.

(e)

In MAP estimation, we can use our prior knowledge to impact the final estimation. Without prior, the estimated parameter is only based on the samples we get. By using prior we break this strong relation by putting our prior knowledge in the model. The prior, in fact, going to be less important when the number of samples increases. On the other hand, with small number of samples, prior can be very effective in helping the model to find a better estimation. It is important to note that, we need to come up with a prior and put into the system.

(f)

Since we want to reflect the sharp decrease in the number of accidents, we can increase the parameter θ in the prior. To shed light the matter, we can look at this in two ways. First, by increasing parameter θ in our prior which is in fact an exponential distribution, the mean of the distribution will decrease so the estimated λ from this distribution will be smaller. Another way to look at this problem, as the λ from the MAP estimate is $\frac{\sum_{i=1}^{n} x_i}{n+\theta}$ by increasing θ, estimated λ would decrease.

Question 5

Honestly, I thought about the question not in the framework of the Question 5 and I think Naive Bayes can be a possible estimate of the problem. It is also possible to apply this for part(c). Then it is going to be the joint distribution.

(a)

We can consider our samples be in the form of (x_i, y_i) in which x_i is indicating whether we have sunny or not sunny weather and y_i is the corresponding table state (free or not free). We want to find the maximum likelihood estimation for

$$\lambda_{ML} = \arg\max_{\lambda} p(\mathcal{D}|\lambda)$$

One possible solution is to find two distributions based on variable x. If $x = sunny$ we find a Bernoulli distribution with parameter θ_1 and for the case $x = not\_sunny$ we find anther Bernoulli distribution with parameter θ_2.

(b)

To find the maximum likelihood of each distribution, we can use the corresponding data. To clarify, we use the subset of data in which x is what we want. When it comes to predication we just use the corresponding distribution.

(c)

We can now have 6 functions, three possible values for day time and two possible values sunny or not sunny. Then we can find have different estimation for each (day time, weather state).