CSCI 5521 Denis Rybkin
Section 2 rybki001@umn.edu

Homework 2
Due Friday, October 23

Note: Every time we use log in this homework, we refer to natural log.

Problem 1. Derive the maximum likelihood estimate of θ based on the samples $\mathcal{X}$:

(a) $p(x|\theta) = \frac{1}{\sqrt{2\pi}\theta}\exp(-\frac{(x-2)^2}{2\theta^2}), \theta > 0$:

We know that the likelihood function for $\mathcal{X}$ is

$$\mathcal{L}(\theta|\mathcal{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta}\exp(-\frac{(x_i-2)^2}{2\theta^2})$$

We can simplify it by calculating the product for each part:

$$\mathcal{L}(\theta|\mathcal{X}) = (\frac{1}{2\pi})^{\frac{n}{2}}\frac{1}{\theta^n}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-2)^2}{2\theta^2}\right)$$

Take the log of the function:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = -\log(2\pi^{\frac{n}{2}}\theta^n) - \frac{\sum_{i=1}^{n}(x_i-2)^2}{2\theta^2}$$

Or equivalently:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = -\frac{n}{2}\log(2\pi) - n\log(\theta) - \frac{\sum_{i=1}^{n}(x_i-2)^2}{2\theta^2}$$

Now, we take the derivative of the function with respect to θ and set it equal to 0:

$$\frac{\partial}{\partial\theta}\log(\mathcal{L}(\theta|\mathcal{X})) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n}(x_i-2)^2}{\theta^3} = 0$$

Now, let's multiply both sides by θ and add n to get:

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

From this point, we can rearrange the equation by multiplying both sides by $\frac{\theta^2}{n}$ and taking the square root:

$$\theta = \sqrt{\frac{\sum_{i=1}^{n}(x_i - 2)^2}{n}}$$

Hence, we derived the maximum likelihood estimate of θ in (a).

(b) $p(x|\theta) = \frac{1}{\theta}exp(-\frac{x}{\theta}), 0 \leq x < \infty, \theta > 0$:

We know that the likelihood function for $\mathcal{X}$ is

$$\mathcal{L}(\theta|\mathcal{X}) = \prod_{i=1}^{n} \frac{1}{\theta} \exp\Big(- \frac{x_i}{\theta}\Big)$$

Applying the product for each term:

$$\mathcal{L}(\theta|\mathcal{X}) = \frac{1}{\theta^n} \exp\left(- \frac{\sum_{i=1}^{n} x_i}{\theta}\right)$$

Let's take the log of the likelihood function:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = -nlog(\theta) - \frac{\sum_{i=1}^{n} x_i}{\theta}$$

Now, take the derivative of the function with respect to θ and set it equal to 0:

$$\frac{\partial}{\partial\theta} \log(\mathcal{L}(\theta|\mathcal{X})) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0$$

Multiply both sides by θ, add then add n. Then, multiply by $\frac{\theta}{n}$ to get:

$$\theta = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Hence, we derived the maximum likelihood estimate of θ in (b).

(c) $p(x|\theta) = \frac{1}{2\theta^3} x^2 \exp(-\frac{x}{\theta}), 0 \leq x < \infty, \theta > 0$:

We know that the likelihood function for $\mathcal{X}$ is

$$\mathcal{L}(\theta|\mathcal{X}) = \prod_{i=1}^{n} \frac{1}{2\theta^3} x_i^2 \exp\Big(-\frac{x_i}{\theta}\Big)$$

Applying the product for each term:

$$\mathcal{L}(\theta|\mathcal{X}) = \frac{1}{2^n \theta^{3n}} \prod_{i=1}^{n} x_i^2 \ \cdot \ \exp\left(-\frac{\sum\limits_{i=1}^{n} x_i}{\theta}\right)$$

Let's take the log of the likelihood function:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = -3n\log(\theta) - n\log(2) + \log(\prod_{i=1}^{n} x_i^2) - \frac{\sum\limits_{i=1}^{n} x_i}{\theta}$$

Simplify by using properties of logarithm, we get:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = -3n\log(\theta) - n\log(2) + 2\sum_{i=1}^{n} \log(x_i) - \frac{\sum\limits_{i=1}^{n} x_i}{\theta}$$

Now, take the derivative of the function with respect to θ and set it equal to 0:

$$\frac{\partial}{\partial\theta} \log(\mathcal{L}(\theta|\mathcal{X})) = -\frac{-3n}{\theta} + \frac{\sum\limits_{i=1}^{n} x_i}{\theta^2} = 0$$

Now we multiply both sides by θ, add $3n$ and rearrange to get:

$$\theta = \frac{\sum\limits_{i=1}^{n} x_i}{3n}$$

Hence, we derived the maximum likelihood estimate of θ in (c).

(d) $p(x|\theta) = \theta x^{\theta-1}, 0 \leq x \leq 1, 0 < \theta < \infty$:

We know that the likelihood function for $\mathcal{X}$ is

$$\mathcal{L}(\theta|\mathcal{X}) = \prod_{i=1}^{n} \theta x_i^{\theta-1}$$

Applying the product for each term:

$$\mathcal{L}(\theta|\mathcal{X}) = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

Let's take the log of the likelihood function:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = \log(\theta^n) \sum_{i=1}^{n} \log(x_i^{\theta-1})$$

Simplify by using the log identity:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(x_i)$$

Now, take the derivative of the function with respect to θ and set it equal to 0:

$$\frac{\partial}{\partial \theta} \log(\mathcal{L}(\theta|\mathcal{X})) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) = 0$$

Then we multiply both sides by θ, subtract n and rearrange to get equation for θ:

$$\theta = -\frac{n}{\sum\limits_{i=1}^{n} log(x_i)}$$

Hence, we derived the maximum likelihood estimate of θ in (d).

(e) $p(x|\theta) = \frac{1}{\theta}, 0 \leq x \leq \theta, \theta > 0$:

We know that the likelihood function for $\mathcal{X}$ is

$$\mathcal{L}(\theta|\mathcal{X}) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}$$

Let's take the log of the likelihood function:

$$\log(\mathcal{L}(\theta|\mathcal{X})) = -n\log(\theta)$$

Now, take the derivative of the function with respect to θ:

$$\frac{\partial}{\partial\theta}\log(\mathcal{L}(\theta|\mathcal{X})) = \frac{-n}{\theta}$$

From the equation above, it follows that the log-likelihood and the likelihood is a decreasing function. In order to maximize it, we would need to minimize θ. Given the constraints $0 \leq x \leq \theta, \theta > 0$, we should be able to minimize θ by setting it to the maximum of the x_i values. Although the maximum may not occur in this interval, we can still maximize the likelihood within the interval.

Hence, we showed how to get the maximum likelihood estimate of θ in (e).

Problem 2.

(a) Derive the maximum likelihood estimates for the mean μ and covariance Σ based on the sample set $\mathcal{X}$.

First, we will need the following three formulas from The Matrix Cookbook:

$$\frac{\partial}{\partial s} = (x-s)^T\mathbf{W}(x-s) = -2\mathbf{W}(x-s) \tag{1}$$

$$\frac{\partial ln|det(\mathbf{X})|}{\partial\mathbf{X}} = (\mathbf{X}^{-1})^T = (\mathbf{X}^T)^{-1} \tag{2}$$

$$\frac{\partial a^T\mathbf{X}a}{\partial\mathbf{X}} = \frac{\partial a^T\mathbf{X}^Ta}{\partial\mathbf{X}} = \mathbf{a}\mathbf{a}^T \tag{3}$$

(In the book, these are equations 86, 57 and 72 respectively. I listed them here to avoid references to the book.)

We have the likelihood function as the joint density for μ and Σ:

$$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} exp\bigg(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \bigg)$$

Apply the product for each term:

$$\mathcal{L} = \frac{1}{(2\pi)^{\frac{nd}{2}} |\Sigma|^{\frac{n}{2}}} exp\bigg(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \bigg)$$

Now we take the log and then simplify the equation using log properties to get:

$$\log(\mathcal{L}) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)$$

First, we find the maximum likelihood estimate for μ. For that, we will take the partial derivative with respect to μ and set it to 0. After that, we can apply the equation (1) from above to simplify:

$$\frac{\partial}{\partial \mu}\log(\mathcal{L}(\mu|\mathcal{X})) = -\frac{1}{2}\sum_{i=1}^{n}[-2\Sigma^{-1}(x_i - \mu)] = 0$$

Let's further simplify the equation by dividing both sides by Σ^{-1} and rewriting it:

$$\frac{\partial}{\partial \theta}\log(\mathcal{L}(\mu|\mathcal{X})) = \sum_{i=1}^{n}(x_i - \mu) = 0$$

Now we can expand the sum and add $n\mu$ to both sides. Then we will divide by n to obtain μ:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Similarly, we now consider the log-likelihood function for Σ:

$$\frac{\partial}{\partial \theta}\log(\mathcal{L}(\Sigma|\mathcal{X})) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)$$

Using properties of log, we can rewrite the equation as:

$$\frac{\partial}{\partial\theta}\log(\mathcal{L}(\Sigma|\mathcal{X})) = -\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log(|\Sigma^{-1}|) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

Now, we take the derivative with respect to Σ^{-1} and use equations (2) and (3) to arrive at:

$$\frac{\partial}{\partial\Sigma^{-1}}\log(\mathcal{L}(\Sigma|\mathcal{X})) = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T] = 0$$

From here, we can rewrite the equation by multiplying each side by 2, adding the sum to both sides and divide by n to get the equation for Σ:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T]$$

Hence, we derived the maximum likelihood estimates for the mean μ and covariance Σ based on the sample set $\mathcal{X}$.

(b) Let $\hat{\mu}_n$ be the estimate of the mean.

Let's compute $E[\hat{\mu}_n]$:

$$E[\hat{\mu}_n] = E\left[\frac{\sum\limits_{i=1}^{n} x_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}E[x_i] = \frac{n\mu}{n} = \mu$$

Hence, $\hat{\mu}_n$ is an **unbiased** estimate of the true mean μ.

(c) Let $\hat{\Sigma}_n$ be the estimate of the covariance.

Let's compute $E[\hat{\Sigma}_n]$:

$$
\begin{aligned}
E[\hat{\Sigma}] &= E\left[\frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T]\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[(x_i-\mu)(x_i-\mu)^T] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[x_i x_i^T] - nE[\mu\mu^T] \\
&= \frac{n-1}{n}\Sigma \\
&\neq \Sigma
\end{aligned} \tag{4}
$$

Therefore, $\hat{\Sigma}_n$ is a **biased** estimate of the true covariance matrix Σ. However, from the formula, it is evident that the greater n we have, the more accurate the estimate is. In other words, as $n \to \infty$, $\hat{\Sigma}_n \to \Sigma$.

Problem 3.

Note: I added some random noise to Digits dataset in order to avoid singular covariance matrices. The value of epsilon is 10^{-6}, so it shouldn't affect the output in any significant way. The implementation of it can be found in *datasets.py* file.

Summary of results:

MultiGaussClassify with full covariance matrix on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.1569	0.2178	0.2178	0.1584	0.2574	0.2017	0.0388

MultiGaussClassify with full covariance matrix on Boston25						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.1176	0.0990	0.1386	0.0693	0.1287	0.1107	0.0245

MultiGaussClassify with full covariance matrix on Digits						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.1111	0.1556	0.1003	0.0780	0.0808	0.1051	0.0280

MultiGaussClassify with diagonal covariance matrix on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.1275	0.1881	0.2772	0.1980	0.2178	0.2017	0.0483

MultiGaussClassify with diagonal covariance matrix on Boston25						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.2059	0.1386	0.1584	0.0990	0.1089	0.1422	0.0382

MultiGaussClassify with diagonal covariance matrix on Digits						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.5556	0.5083	0.4513	0.3928	0.4568	0.4729	0.0552

Logistic Regression on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.0686	0.1881	0.1386	0.1287	0.1980	0.1444	0.0465

Logistic Regression on Boston25						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.1078	0.0990	0.1485	0.1287	0.0792	0.1127	0.0240

Logistic Regression on Digits						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.0444	0.0306	0.0474	0.0306	0.0306	0.0367	0.0075