CSCI 5521 Spring 2017 Homework #2

Craig Ching
#1452647
chin0007@umn.edu

March 3, 2017

1 Problem 1

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

First, specify the likelihood as the joint density for X as:

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

Then simplify by calculating the product separately for $\frac{1}{\sqrt{2\pi}}$, $\frac{1}{\theta}$ and the exp term:

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

Now take the log of the likelihood function:

$$logL = -log(2\pi^{\frac{n}{2}}\theta^n) - \frac{x_i^2}{2\theta^2} \tag{3}$$

Apply log identities for multiplication and exponents

$$-\frac{n}{2} log 2\pi - n log\theta - \frac{\sum_{i=1}^{n} x_i^2}{2\theta^2} \tag{4}$$

Take the derivative of the log-likelihood function with respect to θ and set it equal to 0

$$\frac{\partial}{\partial\theta} logL = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i^2}{\theta^3} = 0 \tag{5}$$

Simplify by multiplying through by θ

$$\frac{\partial}{\partial\theta} logL = -n + \frac{\sum_{i=1}^{n} x_i^2}{\theta^2} = 0 \tag{6}$$

Add n to both sides:

$$\frac{\sum_{i=1}^{n} x_i^2}{\theta^2} = n \tag{7}$$

Multiply both sides by θ^2, divide both sides by n and solve for θ by taking the square root of both sides:

$$\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \theta \tag{8}$$

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

First, specify the likelihood as the joint density for X as:

$$L = \prod_{i=1}^{n} \frac{1}{\theta} exp\Big(- \frac{x_i}{\theta}\Big) \tag{9}$$

Carry out the product

$$L = \frac{1}{\theta^n} exp\left(- \frac{\sum\limits_{i=1}^{n} x_i}{\theta}\right) \tag{10}$$

Take the log of the likelihood function and simplify using the identity for log of a quotient and log of an exponent

$$logL = -nlog\theta - \frac{\sum\limits_{i=1}^{n} x_i}{\theta} \tag{11}$$

Take the derivative of the log-likelihood function with respect to θ and set it equal to 0

$$\frac{\partial L}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum\limits_{i=1}^{n} x_i}{\theta^2} = 0 \tag{12}$$

Simplify by multiplying through by θ, adding n to both sides, multiplying through by θ again, then dividing by n

$$\theta = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \bar{x} \tag{13}$$

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

First, specify the likelihood as the joint density for X as:

$$L = \prod_{i=1}^{n} \theta x_i^{\theta-1} \tag{14}$$

Carry out the product for θ

$$L = \theta^n \prod_{i=1}^{n} x_i^{\theta-1} \tag{15}$$

Take the log of the likelihood function

$$logL = log\theta^n \sum_{i=1}^{n} logx_i^{\theta-1} \tag{16}$$

Use the log identity for exponents to simplify

$$logL = nlog\theta + (\theta - 1)\sum_{i=1}^{n} logx_i \tag{17}$$

Take the derivative of the log-likelihood with respect to θ and set it equal to 0

$$\frac{\partial L}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} logx_i = 0 \tag{18}$$

Multiply through by θ, subtract n from both sides, then divide by the sum of $logx_i$

$$\theta = -\frac{n}{\sum\limits_{i=1}^{n} logx_i} \tag{19}$$

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

First, specify the likelihood as the joint density for X as:

$$L = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n} \tag{20}$$

Take the log of the likelihood and simplify using the exponent identity for *log*

$$logL = -nlog\theta \tag{21}$$

Take the derivative of the log-likelihood with respect to θ

$$\frac{\partial L}{\partial \theta} = \frac{-n}{\theta} \tag{22}$$

This shows us that the log-likelihood (and, hence, the likelihood) is a decreasing function and the way to maximize it is to minimize θ. Though the maximum may not occur in this interval, we can still maximize the likelihood within the interval. Given the constraints $0 \leq x \leq \theta, \theta > 0$, we can minimize θ by setting it equal to the maximum of the x_i values.

2 Problem 2

2.1 (a) Derive the maximum likelihood estimates for mean μ and covariance Σ based on the sample set X

Specify the likelihood function as the joint density for μ and Σ

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} exp\Big[-\frac{1}{2}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)\Big] \tag{23}$$

Carry out the multiplication

$$L = \frac{1}{(2\pi)^{\frac{nd}{2}}|\Sigma|^{\frac{n}{2}}} exp\Big[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)\Big] \tag{24}$$

Take the log of the likelihood and use log identities to simplify to

$$logL = -\frac{nd}{2}log2\pi - \frac{n}{2}log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) \tag{25}$$

Take the partial derivative of the log-likelihood function with respect to μ and, using **The Matrix Cookbook** equation number 86:

$$\frac{\partial}{\partial s} = (x - s)^T\mathbf{W}(x - s) = -2\mathbf{W}(x - s) \tag{26}$$

we arrive at:

$$\frac{\partial L}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{n}[-2\Sigma^{-1}(x_i - \mu)] = 0 \tag{27}$$

Then we simplify by using -2 to cancel the $-\frac{1}{2}$, pull Σ^{-1} out of the sum and divide both sides by Σ^{-1} we arrive at:

$$\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{28}$$

We can simplify this by expanding the sum:

$$\sum_{i=1}^{n} x_i - n\mu \tag{29}$$

Add $n\mu$ to both sides and then divide by n to solve for μ, we get:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{30}$$

To estimate Σ, we start with the log-likelihood function

$$logL = -\frac{nd}{2}log2\pi - \frac{n}{2}log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) \tag{31}$$

Recognizing that

$$|\Sigma| = \frac{1}{|\Sigma^{-1}|} \tag{32}$$

and using the log of the quotient, rewrite the log-likelihoood

$$logL = -\frac{nd}{2}log2\pi + \frac{n}{2}log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) \tag{33}$$

and take the derivative with respect to Σ^{-1} and, using **The Matrix Cookbook** equation number 57

$$\frac{\partial ln|det(\mathbf{X})|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T = (\mathbf{X}^T)^{-1} \tag{34}$$

and equation number 72

$$\frac{\partial a^T\mathbf{X}a}{\partial \mathbf{X}} = \frac{\partial a^T\mathbf{X}^Ta}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \tag{35}$$

we arrive at:

$$\frac{\partial L}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}[(x_i - \mu)(x_i - \mu)^T] = 0 \tag{36}$$

Now, add the sum to each side, multiply both sides by 2, and divide both sides by n to get

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu)(x_i - \mu)^T] \tag{37}$$

2.2 (b) Let $\hat{\mu}_n$ be the estimate of the mean. Is $\hat{\mu}_n$ a biased estimate of the true mean μ. Clearly justify your answer by computing $E[\hat{\mu}_n]$

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

$$\begin{aligned} E[\hat{\mu}] &= E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned} \tag{38}$$

2.3 (c) Let $\hat{\Sigma}_n$ be the estimate of the covariance. Is $\hat{\Sigma}_n$ a biased estimate of the true covariance Σ. Clearly justify your answer by computing $E[\hat{\Sigma}_n]$

$\hat{\Sigma}$ is a **biased** estimate of the true covariance matrix Σ

$$
\begin{aligned}
E[\hat{\Sigma}] &= E\left[\frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T]\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[(x_i-\mu)(x_i-\mu)^T] \\
&= \frac{1}{n}\sum_{i=1}^{n}E[x_i x_i^T] - nE[\mu\mu^T] \\
&= \frac{n-1}{n}\Sigma \\
&\neq \Sigma
\end{aligned}
\tag{39}
$$

This can be corrected by instead writing the covariance as

$$
E[\hat{\Sigma}] = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i-\mu)(x_i-\mu)^T \tag{40}
$$

and becomes less significant the more data you have.

3 Problem 3

3.1 Description

My Multivariate Gaussian classifier uses quadratic discriminants as described in section 5.5 of the Alpaydin text book. Specifically, the discriminant is

$$
g_i(x) = \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0} \tag{41}
$$

This is implemented as described with class-specific means $\mathbf{m}_i$, class-specific covariance matrices $\mathbf{S}_i$ and estimated priors $\hat{P}(C_i)$:

$$
\begin{aligned}
\mathbf{W}_i &= -\frac{1}{2}\mathbf{S}_i^{-1} \\
\mathbf{w}_i &= \mathbf{S}_i^{-1}\mathbf{m}_i \\
w_{i0} &= -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}log|\mathbf{S}_i| + log\hat{P}(C_i)
\end{aligned}
\tag{42}
$$

3.2 Results

NOTE: Gaussian noise $N(0, 0.001)$ was added to the Digits data set to avoid singular covariance matrices.

method: MultiGaussClassify
dataset: Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.2157	0.2475	0.2277	0.1584	0.1980	0.2095	0.0302

method: MultiGaussClassify
dataset: Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.2451	0.1584	0.3168	0.2673	0.2178	0.2411	0.0526

method: MultiGaussClassify
dataset: Digits

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.0667	0.0639	0.0279	0.0362	0.0418	0.0473	0.0154

method: LogisticRegression
dataset: Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.1275	0.1287	0.1683	0.1683	0.0891	0.1364	0.0297

method: LogisticRegression
dataset: Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.1176	0.0792	0.1386	0.0891	0.0693	0.0988	0.0256

method: LogisticRegression
dataset: Digits

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.0417	0.0417	0.0501	0.0251	0.0501	0.0417	0.0092

4 Appendix

Before correcting for the singular covariance matrix, I implemented a linear discriminant function as described in the book:

$$g_i(x) = \mathbf{w}_i^T x + w_{i0} \tag{43}$$

Even with a single, shared covariance matrix, I still had problems with singular covariance matrix. So I played around with first using the identity matrix, then a diagonal matrix with the σ^2 on the diagonal. I believe the latter gave me about a 9% error rate. Finally, once resolving the singular covariance matrix, I was able to implement the linear discriminant with a shared covariance matrix for the **Digits** dataset. Here are the results:
method: MultiGaussClassify
dataset: Digits

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	mean	std dev
0.0750	0.0722	0.0724	0.0669	0.0529	0.0679	0.0079

I only include this here because I went through the trouble and wanted to record it. Feel free to ignore if it's not interesting! If you do want to reproduce it, you can change the line in q3.py that has **MultiGaussClassify(linear=False)** to **True** and you will get the linear discriminant.