Assignment 1

Guanzhi Wang

Exercise 1

(1)

The likelihood function is

$$L(\theta) = (1-\theta)^{x_1-1}\theta(1-\theta)^{x_2-1}\theta\ldots(1-\theta)^{x_n-1}\theta = \theta^n(1-\theta)^{\sum_1^n x_i - n}$$

Thus, taking log, we have:

$$lnL(\theta) = nln\theta + \left(\sum_1^n x_i - n\right) ln(1-\theta).$$

Let

$$\frac{d\left[lnL(\theta)\right]}{d\theta} = \frac{n}{\theta} - \frac{\left(\sum_1^n x_i - n\right)}{(1-\theta)} = 0,$$

we have

$$\theta = \frac{n}{\left(\sum_1^n x_i\right)}.$$

So the maximum likelihood estimator of θ is

$$\frac{1}{X}.$$

(2)

Let $b-a=\theta$, we have the likelihood function:

$$L(\theta)=\prod_{i=1}^{n}f(xi)=\prod_{i=1}^{n}\frac{1}{\theta}=\theta^{-n}.$$

Take log, we have

$$lnL(\theta)=-nln(\theta).$$

Same as (1), we will get

$$\frac{d}{d\theta}lnL(\theta)=\frac{-n}{\theta}=L(b-a)=\frac{-n}{b-a},$$

So

$$\frac{\partial \ln L(a,b)}{\partial a}=\frac{n}{b-a}$$
$$\frac{\partial \ln L(a,b)}{\partial b}=\frac{n}{a-b}.$$

The estimator for a,b are $min(x_i),\ max(x_i)$.

Exercise 2

(1)

The L2 loss function is

$$\sum_{i=1}^{n}(y_i-\hat{y_i})$$

For Normal distribution,

$$f(x_1,x_2,\ldots,x_n|\sigma,\mu)=L(\mu,\ \sigma)=(\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\ \sum_{i=1}^{n}(x_i-\mu)^2}$$

The negative log likelihood is

$$-lnL(\mu,\ \sigma)=-\frac{n}{2}\log(2\pi)-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Since we already know σ, the negative log likelihood can be written as

$$c-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

which is equivalent to the L2 loss function.

(2)

Easy to know the log likelihood function for $f(y) = \frac{1}{2b}\exp^{-\frac{|y-\mu|}{b}}$ is

$$l(y) = n\log\frac{1}{2b} - \frac{\sum |\ y_i - \mu|}{b}.$$

Take out the constant, we know it is equivalent to L1 loss function.

Exercise 3

(1)

Unbias means $E(\theta) - \theta = 0$. Which means Mse is no less than $Var(\hat{\theta})$. Thus, the mean is optimal decision rule for the MSE.

(2)

MAE=$\frac{1}{n}\sum |y_i - \hat{y_i}|$.

To minimize it, we'd like to minimize the sum of distance between each points. Thus, we find median is the optimal decision rule for the mean absolute error.

Exercise 4

(1)

Y={0,1}, $p \in [0, 1]$. Since the cross entropy loss function is second order differentiable, for

$$f(p) = -ylog(p),$$

$f''(p) = \frac{y}{p^2} \geq 0.$

For

$$g(p) = -(1-y)log(1-p),$$

$g''(p) = \frac{1-y}{(1-p)^2} \geq 0.$

So both two functions are convex, then $L(y,\ p)$ is also convex.

(2)

Like (1),

$$L''(y,\ p) = 2(y - p) \leq 0.$$

Thus, the mean squared error loss is not convex.

Exercise 5

(1)

According to the email, skipped.

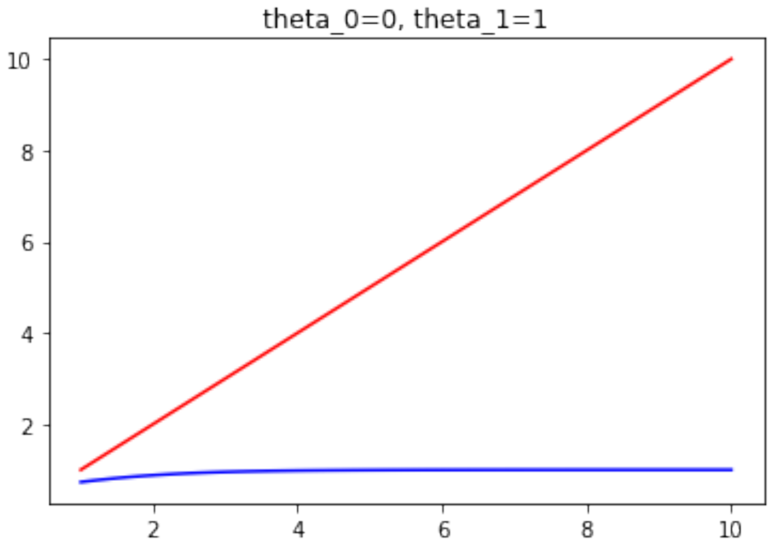
(2)

First, we have

$$logit(f_\theta(x)) = \theta_0 + \theta_1 x,$$

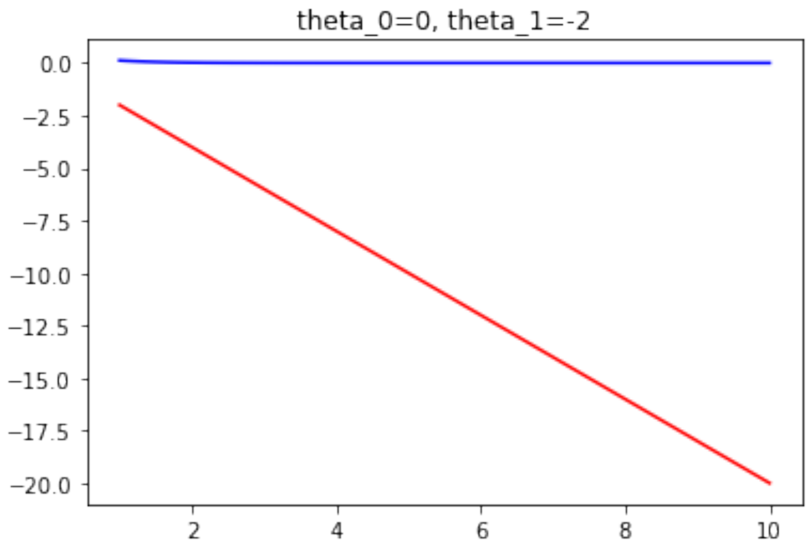
In [44]:
```
import matplotlib.pyplot as plt
import numpy as np

x = np.linspace(1, 10, 100)

theta_0 = 0
theta_1 = 1
theta_x=theta_0+theta_1*x
f_theta_x=1/(1+np.exp(-theta_x))
logit_f_theta_x=theta_x
plt.title("theta_0=0, theta_1=1")
plt.plot(x,logit_f_theta_x,"r")
plt.plot(x,f_theta_x,"b")
```

Out[44]: [<matplotlib.lines.Line2D at 0x113331ac8>]



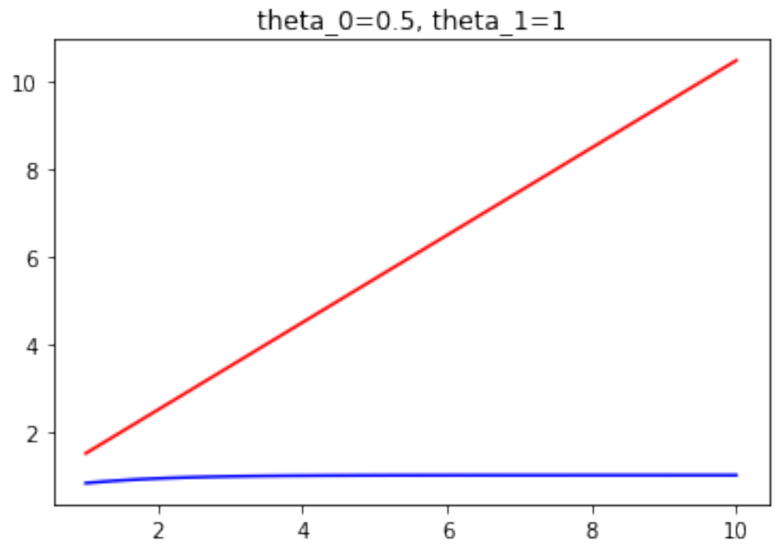
In [45]:
```
theta_0 = 0
theta_1 = -2
theta_x=theta_0+theta_1*x
f_theta_x=1/(1+np.exp(-theta_x))
logit_f_theta_x=theta_x
plt.title("theta_0="+str(theta_0)+", theta_1="+str(theta_1))
plt.plot(x,logit_f_theta_x,"r")
plt.plot(x,f_theta_x,"b")
```

Out[45]: [<matplotlib.lines.Line2D at 0x1133fc320>]



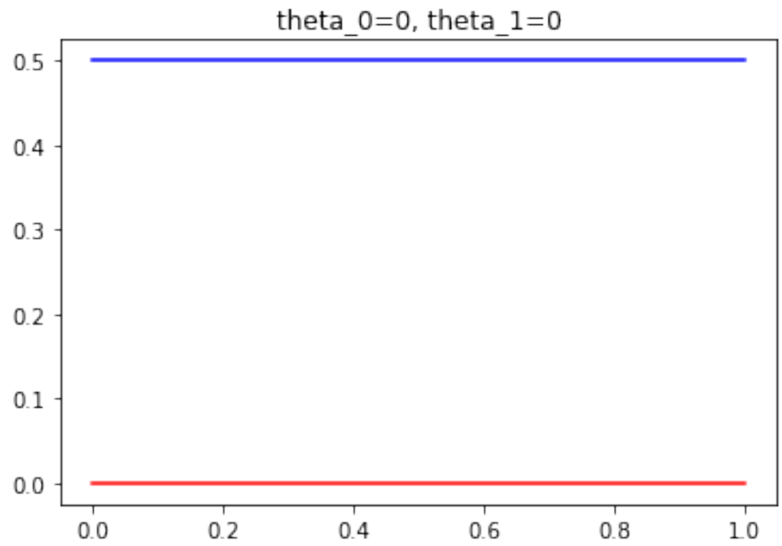
In [46]:
```
theta_0 = 0.5
theta_1 = 1
theta_x=theta_0+theta_1*x
f_theta_x=1/(1+np.exp(-theta_x))
logit_f_theta_x=theta_x
plt.title("theta_0="+str(theta_0)+", theta_1="+str(theta_1))
plt.plot(x,logit_f_theta_x,"r")
plt.plot(x,f_theta_x,"b")
```

Out[46]: [<matplotlib.lines.Line2D at 0x1134ccf28>]



In [42]:
```
theta_0 = 0
theta_1 = 0
theta_x=theta_0+theta_1*x
f_theta_x=1/(1+np.exp(-theta_x))
logit_f_theta_x=theta_x
plt.title("theta_0="+str(theta_0)+", theta_1="+str(theta_1))
plt.plot(x,logit_f_theta_x,"r")
plt.plot(x,f_theta_x,"b")
```

Out[42]:
```
[<matplotlib.lines.Line2D at 0x113207a90>]
```



Easy to find $f\theta(x) \sim logistic(\theta \cdot x) = \beta X$, which means θx defines a decision boundary. Also, the logit theta x function is monotonous, thus, $\theta_0 + \theta_1 * x$ is a linear seperating hyperplane.

Exercise 6

For Normal Distribution with unknown mean and known variance, we have

$$f(x_1, \ldots x_n|\mu) = (2\pi)^{\frac{-n}{2}} \sigma^{-n} exp(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}).$$

So let

$$\mu(x_1 \ldots x_n) = (2\pi)^{\frac{-n}{2}} \sigma^{-n} exp(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} x_i^2),$$

we could also transform the function to

$$v(k, \mu) = exp(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} k),$$

where

$$k = \sum_{i=1}^{n} x_i.$$

By the factorization theorem this shows that k is a sufficient statistic.

Exercise 7

(1) From the question we have:

1 ${X_i}_{i=1}^n$ be independent and identically distributed observations.

2 $X_1 = Z_1 + \theta, \ldots, X_n = Z_n + \theta.$

3 $min_i(X_i) = min_i(Z_i + \theta), max_i(X_i) = max_i(Z_i + \theta), where\ {Z_i}_{i=1}^n$ are independent and identically distributed observations from F(x).

So $R = max_i(X_i) - min_i(X_i) = max_i(Z_i) - min_i(Z_i)$ is irrelevant with local parameter. Which means R is ancilliary and not related with theta.

Exercise 8

First, The $N(\mu, \sigma^2) family\ with\ \theta = (\mu, \sigma^2)$ is a 2pef with

$$w(\theta) = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}), t(x) = (x, x^2).$$

In our case, $\mu = \sigma$.

Then $T(X) = (\sum n_{i=1} X_i, \sum n_{i=1} X_i^2)$ is the natural SS.

Then T(X) is a one-to-one function of $U(X) = (\bar{x}, s^2)$.

Let $g(x_1, x_2) = x_1^2 - x_2$, then

$$E(g(U)) = E(\bar{x}^2 - s^2) = \mu^2 - mu^2 = 0.$$

However, g(U) is not trivially 0. So $N(\mu, \sigma^2)$ has a sufficient statistic but is not complete.

Exercise 9

For Poisson distribution,

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$
$$p(x|\lambda) = \frac{1}{x!} exp(xlog\lambda - \lambda),$$

Where $\eta = log\lambda, T(x) = x, A(\eta) = \lambda = e^\eta, h(x) = \frac{1}{x!}$. Problem proved.

Exercise 10

From the property of regular exponential family we have

$$B(\eta) = log \int_x h(x)e^{\eta T(X)} dx.$$

Thus,

$$\frac{\partial B(\eta)}{\partial \eta_i} = E_i(T_i(X)) = \frac{\int_x T_i(X)h(x)e^{\eta T(X)} dx}{\int_x h(x)e^{\eta T(X)} dx}.$$

Also,

$$Cov_\eta(T_i(X), T_j(X)) = E_\eta[T_i(X)T_j(X)] - E_\eta[T_i(X)]E_\eta[T_j(X)]$$

$$= \frac{\int_x T_i(X)T_j(X)h(x)e^{\eta T(X)} dx}{\int_x h(x)e^{\eta T(X)} dx} - \frac{\frac{\partial}{\partial \eta_i} * \frac{\partial}{\partial \eta_j}}{(\int_x h(x)e^{\eta T(X)} dx)^2}.$$

Easy to find

$$\frac{\int_x T_i(X)T_j(X)h(x)e^{\eta T(X)} dx}{\int_x h(x)e^{\eta T(X)} dx} - \frac{\frac{\partial}{\partial \eta_i} * \frac{\partial}{\partial \eta_j}}{(\int_x h(x)e^{\eta T(X)} dx)^2} = \frac{\partial B(\eta)}{\partial \eta_i \partial \eta j}.$$

Problem proved.

Exercise 11

Easy to know,

$$X \sim \text{Bin}(n, p).$$

Thus,

$$E[X^2] = \text{Var}(X) + E[X]^2 = np(1-p) + n^2p^2 = np(1-p+np).$$

In this case,

$$E[\hat{p^2}] = \frac{1}{n^2}E[X^2] = \frac{p(1-p+np)}{n}.$$

By CLT and Delta Method, we have

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p)) \Rightarrow$$

$$\sqrt{n}(\widehat{p(1-p)} - p(1-p)) \xrightarrow{d} N(0, p(1-p)(1-2p)^2).$$

Problem proved.

Exercise 12

(1)
$H(X,Y) = -\sum_x \sum_y P(x,y) log P(x,y) = -2*(1/4*log2(1/4) + 1/6*log2(1/6) + 1/12*log2(1/1$

(2)

$$H(Y|X) = (-2*(1/4*log2(1/4) + 3/4*log2(3/4)) - 1*log2(1/2))/3 = 0.87$$

(3)

$$H(X) = -\sum p(x) log(p(x)) = 1.58.$$

So $H(X,Y) = H(X) + H(Y|X)$.

Exercise 13

$$\begin{aligned} H[x] = -\int_{-\infty}^{+\infty} N(x|\mu,\Sigma)\ln(N(x|\mu,\Sigma))dx \\ = -E[\ln(N(x|\mu,\Sigma))] \\ = -E[\ln((2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)})] \\ = \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}E[(x-\mu)^T\Sigma^{-1}(x-\mu)]. \end{aligned}$$

Simplify the formula, we have

$$H[x] = \frac{1}{2}\ln|\Sigma| + \frac{D}{2}(1+\ln(2\pi)),$$

where D is dimensionality of x.

Extra credit