Yunjia Zeng, collaborate with Shaoyu Feng, Mengtong Zhang ANLY 601
yz682 **Assignment 1** January 30, 2020

Exercise 1

1. Since the geometric distribution is $f(n;p) = (1-p)^{n-1}p$, and $n = 1, 2, 3...$, therefore the likelihood function is

$$\begin{aligned} L(p) &= (1-p)^{x_1-1}p(1-p)^{x_2-1}...(1-p)^{x_n-1}p \\ &= p^n(1-p)^{\sum_1^n x_i - n} \end{aligned}$$

By taking log, and make the first derivative of the log likelihood function equals to 0, $lnL(p) = nlnp + (\sum_1^n x_i - n)ln(1-p) \implies p = \frac{n}{(\sum_1^n x_i)} = \frac{1}{X}$. Which is the maximum likelihood estimate for θ.

2. For $X_i \sim Unif(a,b)$, the likelihood function is $L(x_1, ..., x_n|a,b) = (\frac{1}{b-a})^n$.

Take the derivative with respect to a and b we have:

$$\begin{aligned} \frac{d}{da}ln(L(a,b)) &= \frac{n}{b-a} \\ \frac{d}{db}ln(L(a,b)) &= -\frac{n}{b-a} \end{aligned}$$

Therefore, to maximize this we must minimize the value of $b-a$, yet we must keep all samples with in the range. Thus $\hat{a} = min(X_i), \hat{b} = max(X_i)$.

Exercise 2

1. The log likelihood function for Gaussian is given by

$$\begin{aligned} LL &= \sum_{n=1}^{N} log(N(x_n|\mu, \sigma^2) \\ &= \sum_{n=1}^{N} log(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}}\frac{(x_n-\mu)^2}{\sigma^2}) \end{aligned}$$

Therefore,

$$\begin{aligned} LL &= \sum_{n=1}^{N}(-log(\sqrt{2\pi\sigma^2}) + (-0.5)\frac{(x_n-\mu)^2}{\sigma^2}) \\ LL &= -\frac{N}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n-1} N(x_n-\mu)^2 \end{aligned}$$

Thus, according to the results above the negative log likelihood of a Gaussian is same as L2 Loss (given σ is a constant).

2. The log likelihood function for LaPlace is given by

$$\begin{aligned} LL &= \sum_{n=1} Nlog(L(x_n|\mu, \theta)) \\ &= \sum_{n=1}^{N} \frac{1}{2\theta} e^{(} - \frac{|x-\mu|}{\theta})) \quad = -Nlog(2\theta) - \frac{1}{b} \sum_{n=1}^{N} |x - \mu| \end{aligned}$$

Thus, we conclude L1 loss is equivalent to the negative log likelihood of LaPlace(θ).

Exercise 3

1. Let T be the optimal estimator so that it minimize $MSE(T) = E((T-x)^2)$ for every $x \in X$. Since if T is such that $E(T^2)$ is finite, then $E((T-c)^2) = Var(T) + (E(T)-c)^2$. This is minimized by taking $c = E(T)$ since

$$\begin{aligned} E((T-c)^2) &= E((T - E(T) + E(T) - c)^2) \\ &= E((T - E(T))^2) + 2E(T - E(T))(E(T) - c) + (E(T) - c)^2 \\ &= Var(T) + (E(T) - c)^2 \end{aligned}$$

because $E(T - E(T)) = E(T) - E(T) = 0$. As $(E(T) - c)^2) \geq 0$, and $Var(T)$ does not depend on c, the value is minimized by taking $c = E(T)$. Therefore the estimator comes closest on average to its mean makes optimal unbiased estimation, therefore mean is the optimal decision rule for the MSE when the decision rule is unbiased.

2. The mean absolute error is $MAE = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x_i}|$, for every real valued random variable X,

$$E(|X - c|) = \int_{-\inf}^{c} P(X \leq t)dt + \int_{c}^{+\inf} P(X \geq t)dt$$

hence the function $u : c \rightarrow E(|X - c|)$ is differentiable almost everywhere and, where $u'(c)$ exists, $u'(c) = P(X \leq c) = P(X \geq c)$. Hence $u'(c) \leq 0$ if c is smaller than every median, $u'(c) = 0$ is a median, and $u'(c) \geq 0$ if c is greater than every median. Therefore, for every x and c,

$$|x - c| = \int_{-\inf}^{c} [x \leq t]dt + \int_{c}^{+\inf} [x > t]dt$$

hence, for every median m, $E(|X - c|) = E(|X - m|) + \int_m^c v(t)dt$ with $v(t) = P(X \leq t) - P(X > t) = 2P(X \leq t) - 1$. Then v is non-decreasing and $v(m) \geq 0$ hence, for every $c > m$, $v \geq 0$ on (m, c), which implies $E(|X - c|) \geq E(|X - m|)$. Likewise for $c < m$.

Exercise 4

1. The first order derivative of the cross entropy loss function is

$$\frac{dL}{d\beta} = \frac{dL}{dp} \cdot \frac{dp}{d\beta}$$
$$= -(\frac{y}{p} - \frac{1-y}{1-p}) \cdot (\frac{\beta e^{-\beta x}}{(e^{-\beta x}+1)^2})$$

and the second order derivative is

$$\frac{d^2L}{d\beta^2} = \frac{x^2 e^{\beta x}}{(e^{\beta x}+1)^2}$$

Therefore, according to the equation above, the results are always greater than or equal to 0 for a fixed x, which satisfies the second order theorem of convexity. Thus we conclude that the cross entropy loss function of Bernoulli is convex with respect to β.

2. The first order derivative of the mean squared error loss function is

$$\frac{dL}{d\beta} = \frac{dL}{dp} \cdot \frac{dp}{d\beta}$$
$$= -2(y-p)p(1-p)x$$

The second order derivative is

$$\frac{d^2L}{d\beta^2} = -2[y - 2yp - 2p + 3p^2]x^2p(1-p)$$

The above equation does not satisfy second order convexity since when $y = 0$, the second order derivative is positive only when $p \in [0, 2/3]$. Thus the mean squared loss function of Bernoulli does not convex.

Exercise 5

1. If $\theta = 0$, then

$$f_\theta(0) = \frac{1}{1 + exp(-\beta_0 + \beta_1 x)} = \frac{1}{2},$$

therefore, the decision threshold to classify a point to either class A or B is $f_\theta(x) = 0.5$. If $\beta_0 = 100$, then

$$f_\theta(0) = \frac{1}{1 + exp(100)} = 0,$$

therefore all points will be classified to class A.

2. Since

$$\begin{aligned} logit(f_\theta(x)) &= log(\frac{f_\theta(x)}{1-f_\theta(x)}) \\ &= log(\frac{1}{exp(-\beta x}), \end{aligned}$$

given that logit function is monotonous, therefore the we can rewrite the function by $\beta x = logit^{-1}(c)$, by supposing c as decision threshold. Therefore, $\theta \cdot x = \theta_0 + \theta_1 x$ is a linear separable hyperplane.

Exercise 6

Let $f(x|\theta)$ denote the joint pdf or pmf of the sample X. According to Factorization Theorem, a statistic is a sufficient statistic for θ if and only if there exists a factorization of the function $f(x|\theta)$ into two functions, $h(x)$ and $g(t|\theta)$ for all sample points x and all parameter points θ, which means $f(x|\theta) = g(T(x)|\theta)h(x)$.

Therefore, since the samples follows normal distribution,

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

Since σ^2 is known, then we have

$$h(x) = 2\pi)^{-n/2}\sigma^{-n}exp^{(\frac{-1}{2\sigma^2}\sum_{i=1}^n x_i^2)}$$

and

$$g(r(x_1,x_2,...,x_n),\mu) = exp^{\frac{\mu}{\sigma^2}r(x_1,x_2,...,x_n)-\frac{n\mu^2}{2\sigma^2}}$$

where

$$r(x_1,x_2,...,x_n) = \sum_{i=1}^n x_i$$

Therefore, by factorization theorem $\sum_{i=1}^n x_i$ is a sufficient statistics. Thus the sample mean is also a sufficient statistic.

Exercise 7

Suppose $Z_{i\,i=1}^{\,n}$ is identically and independent distributed random variables with cdf $F(x)$. And we have $X_i = Z_i + \theta$. Therefore, $R = max(X_i) - min(X_i) = max(Z_i+\theta) - min(Z_i+\theta) = max(Z_i) - min(Z_i)$. R is a function of Z_i, thus it does not depend on θ.

Exercise 8

$$
\begin{aligned}
f(x_1, x_2, \dots, x_n|\mu) &= (2\pi|\mu|)^{-\frac{n}{2}} exp(\frac{1}{2\pi^2}\sum_i (x_i - \mu)^2) \\
&= (2\pi|\mu|)^{-\frac{n}{2}} exp(-\frac{n}{2|\mu|}(\bar{x} - \mu)^2) exp(-\frac{1}{2|\mu|^2} s^2)
\end{aligned}
$$

Easy to know $(\bar{x}, s^2)$ is a sufficient statistic for $N(\mu, \mu^2)$. Let $T = (\bar{x}, s^2)$, $h(T) = \bar{x}^2 - \frac{n+1}{n} s^2$, therefore

$$
E(h(T)) = E((\bar{x}))^2 + Var(\bar{x}) - \frac{n+1}{n} E(s^2) \quad = \mu^2 + \frac{\mu^2}{n} - \frac{n+1}{n}\mu^2 = 0
$$

but $h(T)$ is not trivially 0.

Exercise 9

To show that Poisson distribution is part of the regular exponential family, we first to be clear that if f_θ follows

$$
f(x|\theta) = h(x) e^{\psi(\theta)T(X) - A(\theta)}
$$

where $A(\theta)$ is the cumulant, and $T(X)$ is the sufficient statistics for the parameter, so that it is in exponential family. We can also rewrite the canonical form as

$$
f(X = x|\eta) = h(x)^{\eta T(X) - B(\eta)}
$$

The pmf of Poisson distribution is given by

$$
\begin{aligned}
p(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \frac{1}{x!} e^{x log\lambda - \lambda}
\end{aligned}
$$

In such a case, we show that Poisson distribution is part of the regular exponential family with $\eta = log\lambda$ and $T(x) = x$ and $B(\eta) = \lambda$ and $h(x) = \frac{1}{x!}$

Exercise 10

Since

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

By differentiating the equation with respect to η_i we have

$$\begin{aligned}\frac{\delta}{\delta\eta_i}B(\eta) &= \frac{\int_x T_i(x)h(x)e^{\eta T(X)}dx}{\int_x h(x)e^{\eta T(X)}dx} \\ &= E_i[T_i(X)]\end{aligned}$$

By differentiating $Z(\eta) = \int_x T_i(x)h(x)e^{\eta T(X)}dx$ with respect to η_j we now have

$$\begin{aligned}\frac{\delta^2}{\delta\eta_i\eta_j}B(\eta) &= \frac{\int_x T_i(x)T_j(x)h(x)e^{\eta T(X)}dx}{Z(\eta)} - \frac{(\frac{\delta}{\delta\eta_i})(\frac{\delta}{\delta\eta_j})}{Z(\eta)^2} \\ &= E_\eta[T_i(X)T_j(X)] - E_\eta[T_i(X)]E_\eta[T_j(X)] \\ &= Cov_\eta[T_i(X), T_j(X)]\end{aligned}$$

Exercise 11

Since $X_i \sim Bernoulli(p)$,

$$n\bar{X} = X_1 + X_2 + \ldots + X_n \sim Binomial(n, p)$$

and we know that

$$E(\bar{X}) = p$$

and

$$Var(\bar{X}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

From Delta Method, we obtain

$$\begin{aligned}Var(\bar{X}(1-\bar{X})) &= (1-2\hat{p})^2 Var(\bar{X}) \\ &= (1-2\hat{p})^2\frac{p(1-p)}{n} \\ &= \frac{(1-2p)^2(1-p)p}{n^2}\end{aligned}$$

Thus, we have

$$\sqrt{n}(\hat{p}(1-\hat{p}) - p(1-p)) \sim N(0, (1-2p)^2(1-p)p),$$

which is the approximate distribution for τ.

Exercise 12

1. The joint entropy $H(X, Y)$ of X and Y can be computed as

$$\begin{aligned} H(X,Y) &= \sum_x \sum_y P(x,y) log_2(P(X,y)) \\ &= P(0,0)log_2(P(0,0)) + P(1,0)log_2(P(1,0)) + P(2,0)log_2(P(2,0)) \\ &+ P(0,1)log_2(P(P(0,1)) + P(1,1)log_2(P(1,1)) + P(2,1)log_2(P(2,1)) \\ &= 2 \cdot \frac{1}{4} log_2 \frac{1}{4} + 2 \cdot \frac{1}{6} log_2 \frac{1}{6} + 2 \cdot \frac{1}{12} log_2 \frac{1}{12} \\ &= -2.45 \end{aligned}$$

2. The marginal distribution of X are as follows

$$P(X = 0) = P(X = 0|Y = 0) + P(X = 0|Y = 1) = \frac{1}{3}$$

$$P(X = 1) = P(X = 1|Y = 0) + P(X = 1|Y = 1) = \frac{1}{3}$$

$$P(X = 2) = P(X = 2|Y = 0) + P(X = 2|Y = 1) = \frac{1}{3}$$

The conditional entropy can be computed as

$$P(Y = 0|X = 0) = \frac{P(X = 0, Y = 0)}{P(X = 0)} = 3/4$$

$$P(Y = 0|X = 1) = \frac{P(X = 1, Y = 0)}{P(X = 0)} = 1/4$$

$$P(Y = 0|X = 2) = \frac{P(X = 2, Y = 0)}{P(X = 0)} = 1/2$$

$$P(Y = 1|X = 0) = \frac{P(X = 0, Y = 1)}{P(X = 1)} = 1/4$$

$$P(Y = 1|X = 1)\frac{P(X = 1, Y = 1)}{P(X = 1)} = 3/4$$

$$P(Y = 1|X = 2)\frac{P(X = 2, Y = 1)}{P(X = 1)} = 1/2$$

Thus we have

$$H(Y|X = 0) = \frac{3}{4} \cdot log_2(\frac{3}{4}) + \frac{1}{4} \cdot log_2(\frac{1}{4}) = -0.81$$

$$H(Y|X = 1) = \frac{1}{4} \cdot log_2(\frac{1}{4}) + \frac{3}{4} \cdot log_2(\frac{3}{4}) = -0.81$$

$$H(Y|X=2) = log_2(\frac{1}{2}) = -1$$

$$H(Y|X) = \frac{1}{3}\cdot(-0.81\cdot 2 - 1) = -0.87$$

Exercise 12 - Differential Entropy

The differential entropy of the multivariate normal distribution can be computed as

$$\begin{aligned}
-\int_{-\infty}^{+\infty} N(x|\mu,\sum)ln(N(x|\mu,\sum))dx &= -E[ln(N(x|\mu,\sum))] \\
&= -E[ln((2\pi)^{-\frac{D}{2}}|\sum|^{-0.5}e^{-0.5(x-\mu)^T\sum^{-1}(x-\mu)})] \\
&= \frac{D}{2}ln(2\pi) + \frac{1}{2}ln|\sum| + \frac{1}{2}E[(x-\mu)^T\overset{-1}{\sum}(x-\mu)]
\end{aligned}$$

Then we have

$$\begin{aligned}
E[(x-\mu)^T\overset{-1}{\sum}(x-\mu)] &= E[tr((x-\mu)^T\overset{-1}{\sum}(x-\mu))] \\
&= tr(E[\overset{-1}{\sum}(x-\mu)(x-\mu)^T]) \\
&= tr(\overset{-}{\sum}1\sum) \\
&= tr(I) \\
&= D,
\end{aligned}$$

therefore, the differential entropy is

$$\frac{D}{2}ln(2\pi) + \frac{1}{2}ln|\sum| + D,$$

where D is the number of dimensions.