ANLY-601 HW1

Norman Hong

January 29, 2020

1.)

a.) What is the maximum likelihood estimate for θ when $X_i \sim Geometric(\theta)$?

Let X_i be a geometric random variable with parameter θ. Then X_i is equal to the number of trials until first success.

$$P(X = x) = (1-p)^{x-1}p$$

$$L(\theta; p) = P(\theta; X = x) = (1-p)^{x-1}p = \frac{(1-p)^x}{(1-p)}p$$

$$\frac{dL}{dp} = (1-p)^{x-1} + p(1-p)^{x-2}(x-1)(-1)$$

$$0 = (1-p)^{x-1} - p(x-1)(1-p)^{x-2}$$

$$p(x-1)(1-p)^{x-2} = (1-p)^{x-1}$$

$$\frac{p(x-1)(1-p)^x}{(1-p)^2} = \frac{(1-p)}{1-p}$$

$$p(x-1)(1-p)^x = (1-p)^x(1-p)$$

$$px - p = 1 - p$$

$$p = 1/x$$

Since $\theta = p$, the mle for θ is $1/x$

b.) What is the maximum likelihood estimate for α and β when $X_i \sim Unif(\alpha, \beta)$?

Since, X_i is uniform distributed, the probability density function is:

$$P(X = x) = \begin{cases} \frac{1}{b-a} & x \leq 0 \\ 0 & otherwise \end{cases}$$

Then the loss function $L(\theta; X)$ is

$$L(\theta; X) = (b-a)^{-1}$$

Since there is no stationary point, can't take the derivative to find Maximum likelihood estimate for α and β. Note that the goal is to select α and β to maximize the likelihood function. The smallest possible value of $\beta - \alpha$ is the range the covers the data. Therefore, $\beta = max(X_i)$ and $\alpha = min(X_i)$.

2.)

a.) Show that squared error loss (L2 loss) is equivalent to the negative log likelihood of a $Y \sim N(\mu, \sigma)$ where $sigma$ is known.

The mean squared error, mse, is defined as

$$mse = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}$$

. The pdf of normal distribution is

$$f(x;\mu,b) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

. Assume have a collection of n observations that are i.i.d. Then, the likelihood function $L(Y;\mu,b)$ is

$$L(Y;\mu,b) = (\frac{1}{\sigma\sqrt{2\pi}})^n(e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2})^n$$

. Then the negative log likelihood is

$$-\log(L(Y;\mu,b)) = -\log[(\frac{1}{\sigma\sqrt{2\pi}})^n(e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2})^n]$$

$$-\log(L(Y;\mu,b)) = -\log[(\frac{1}{\sigma\sqrt{2\pi}})^n] - \log[(e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2})^n]$$

$$-\log(L(Y;\mu,b)) = -n\log[(\frac{1}{\sigma\sqrt{2\pi}})] - \log[(e^{\frac{-1}{2}(\frac{\sum_{i=1}^n x_i-\mu}{\sigma})^2})]$$

$$-\log(L(Y;\mu,b)) = -n\log[(\frac{1}{\sigma\sqrt{2\pi}})] - \frac{-1}{2}(\frac{\sum_{i=1}^n x_i-\mu}{\sigma})^2$$

$$-\log(L(Y;\mu,b)) = -n\log[(\frac{1}{\sigma\sqrt{2\pi}})] + \frac{1}{2}(\frac{\sum_{i=1}^n x_i-\mu}{\sigma})^2$$

$$-\log(L(Y;\mu,b)) = -n\log[(\frac{1}{\sigma\sqrt{2\pi}})] + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}$$

Note that $\sum_{i=1}^n(y_i-\hat{y}_i) = \sum_{i=1}^n(y_i-\bar{y})$ where $\bar{y}$ is the mean of $\hat{y}$. This can be shown by a few algebraic steps Since σ, n, $-n\log[(\frac{1}{\sigma\sqrt{2\pi}})]$, and μ are constants, it follows that the mean squared error is a simple linear transformation of the negative log likelihood of the normal distribution.

b.) Show that the mean absolute error (L1 loss) is equivalent to the negative log likelihood of a $Y \sim LaPlace(\theta)$.

The mean absolute error, mae, is defined as

$$mae = \frac{\sum_{i=1}^n |y_i-\hat{y}_i|}{n}$$

. The pdf of laplace distribution is

$$f(x;\mu,b) = \frac{1}{2b}e^{\frac{-|x-\mu|}{b}}$$

. Assume have a collection of n observations that are i.i.d. Then, the likelihood function $L(Y;\mu,b)$ is

$$L(Y;\mu,b) = (\frac{1}{2b})^n(e^{\frac{-|x-\mu|}{b}})^n$$

$$L(Y;\mu,b) = (\frac{1}{2b})^n e^{\sum_{i=1}^n \frac{-|x_i-\mu|}{b}}$$

Then the negative log likelihood is

$$-\log(L(Y;\mu,b)) = -\log[(\frac{1}{2b})^n e^{\sum_{i=1}^n \frac{-|x_i-\mu|}{b}}]$$

$$-\log(L(Y;\mu,b)) = -\log[(\frac{1}{2b})^n] - \log[e^{\sum_{i=1}^n \frac{-|x_i-\mu|}{b}}]$$

$$-\log(L(Y;\mu,b)) = -n\log[(\frac{1}{2b})] - \sum_{i=1}^{n}\frac{-|x_i-\mu|}{b}$$

$$-\log(L(Y;\mu,b)) = -n\log[(\frac{1}{2b})] + \sum_{i=1}^{n}\frac{|x_i-\mu|}{b}$$

Note that $\sum_{i=1}^{n}(y_i - \hat{y_i}) = \sum_{i=1}^{n}(y_i - \bar{y})$ where $\bar{y}$ is the mean of $\hat{y}$. This can be shown by a few algebraic steps Since b, n, $-n\log(\frac{1}{2b})$, and μ are constants, it follows that the mean absolute error is a simple linear transformation of the negative log likelihood of the laplace distribution.

3.) Suppose that X has mean μ and variance $\sigma^2 < \infty$.

a.) Show that the mean is optimal decision rule for the mean squared error when the decision rule is unbiased.

Let $L(\theta;\delta(x))$ denote the mean squared error loss function where θ is the parameter of interest and $\delta(x)$ is the decision function that estimates θ. Therefore, the loss function has the following form

$$L(\theta;\delta(x)) = Var(\delta(x)) + Bias^2$$

Unbiased $\delta(x)$ implies $E(\delta(x)) = \theta$. Since $\delta(x)$ is unbiased, the loss function becomes

$$L(\theta;\delta(x)) = Var(\delta(x))$$

Assume collection of n estimates from decision function. Then the variance of the decision function is

$$Var(\delta(x)) = E[(\delta(x) - E[\delta(x)])^2]$$

This equation means how far, on average, is the collection of estimates from decision function from the expected value of the decision function. Lets say the decision function follows some unknown distribution and $E[\delta(x)]$ is the expected value of this distribution. It follows that the value of $\delta(x)$ that minimizes the variance the most is the sample mean μ. The reason is that the sample mean on average will equal the expected value of this distribution as long as the decision function is unbiased. Therefore, the variance will approach 0 and msean squared error will approach 0.

b.) Show the median is the optimal decision rule for the mean absolute error. Do this by minimizing Risk.

Let the loss function be the absolute error where θ is the true value of parameter of interest. Let $R(\delta(x))$ denote the risk function for decision function δ. Therefore,

$$L(\theta,\delta(x)) = |\theta - \delta(x)|$$

By definition, Risk is the expected value of the loss function. For n estimates of the decision function, risk is defined as

$$R(\delta(x)) = E[|\theta - \delta(x)|] = \sum_{i=1}^{n}\frac{|\theta - \delta(x)_i|}{n}$$

The above equation shows that the mean absolute error measures the distance an estimate is from the true value. Let m denote the median of a random variable. Consider the case when the random variable is continous. Then

$$P(X \le m) = 1/2$$

and

$$P(X \ge m) = 1/2$$

are true. This implies that half of estimates will fall below the median and half the estimates will be above the median. This indicates that, on average, the shortest distance to any point is from the median. It follows that the median will be the optimal decision rule because it gives the minimum risk on average.

4.) Suppose $Y \sim Bernoulli(p)$ where $p = \frac{1}{1+e^{-\beta x}}$.

a.) For a fixed x show that the cross entropy loss $L(y,p) = -(y\log(p)+(1-y)\log(1-p))$ is convex with respect to β.

By definition, a bernoulli distribution is part of the exponential family. Therefore, the natural parameter space T is convex. Let $\eta \in T$. By definition,

$$B(\eta) = p = \frac{1}{1+e^{\eta}}$$

is also convex. If fix x, then β can be thought of as a scaled value of η, which means *beta* is convex. Looking at the loss function, the function is always doing a log transformation of p or $1-p$ then multiplying by y, which is either 1 or 0. Since the log function is monotone and has a 1 to 1 mapping, this means that the loss function is convex with respect to *beta*. By definition, any monotone transformation of a convex function is also a convex function.

b.) Show that the mean squared error loss $L(y,p) = (y-p)^2$ is not convex in *beta*.

The loss function is not monotone nor does it have a 1 to 1 mapping. Therefore, the argument that was used for part a does not apply here. To prove that the loss function is not convex, the second partial derivative is used to show that the function is not always positive. In other words, holding all else constant, the loss function is not convex in β.

$$L(y,\beta,x) = (y-(1+e^{-\beta x})^{-1})^2$$

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

$$\frac{\partial L}{\partial \beta} = -2xe^{-\beta x}y(1+e^{-\beta x})^{-2} + 2xe^{-\beta x}(1+e^{-\beta x})^{-3}$$

$$\frac{\partial^2 L}{\partial \beta^2} = 2x^2e^{-\beta x}[y(1+e^{-\beta x})^{-2} - (1+e^{-\beta x})^{-3}] + 4x^2e^{-2\beta x}y(1+e^{-\beta x})^{-3} + 6x^2e^{-2\beta x}(1+e^{-\beta x})^{-4}$$

Lets fix x and y by setting $x=1$ and $y=0$. The second derivative then becomes

$$\frac{\partial^2 L}{\partial \beta^2} = \frac{-2}{e^{\beta}(1+e^{-\beta})^3} + \frac{6}{e^{2\beta}(1+e^{-\beta})^4}$$

If β is small, then $\frac{6}{e^{2\beta}(1+e^{-\beta})^4}$ dominates so the function is convex. If β is large, then $\frac{6}{e^{2\beta}(1+e^{-\beta})^4}$ approaches 0 faster than $\frac{-2}{e^{\beta}(1+e^{-\beta})^3}$, so $\frac{-2}{e^{\beta}(1+e^{-\beta})^3}$ dominates, which means the loss function is concave down.

6. Suppose $\{X_i\}_{i=1}^n \sim N(\mu,\sigma); \sigma^2 < \infty$ and σ^2 is known. Show that the sample mean $T(X) = \bar{x}$ is a sufficient statistic for μ.

This can be shown using the factorization theorem, which states that if a pdf $f(x|\theta)$ can be broken up into two functions g and h.

$$f(x|\theta) = g(T(x)|\theta)h(x)$$

where g is a function that depends on X only through the sufficient statistic and h is a function that does not depend on any parameters. Let $T(X)$ denote the sufficient statistic for the parameter θ. Since this is a normal distribution with n iid observations and known sigma,

$$f(x|\mu) = (\frac{1}{\sigma\sqrt{2\pi}})^n(e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2})^n$$

Note that $e^{\sum_{i=1}^{n} x_i^2-2x_i\mu+\mu^2} = e^{\sum_{i=1}^{n}(x_i-\mu)^2}$. Therefore, the pdf can be broken up into

$$f(x|\mu) = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{\sum_{i=1}^{n} \frac{-1}{2\sigma^2}(x_i^2-2x_i\mu+\mu^2)}$$

$$f(x|\mu) = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{\sum_{i=1}^{n} \frac{-x_i^2}{2\sigma^2}} e^{\sum_{i=1}^{n} \frac{2\mu x_i}{2\sigma^2}} e^{\frac{-n\mu^2}{2\sigma^2}}$$

Therefore,

$$g(T(x)|\mu) = e^{\frac{-n\mu^2}{2\sigma^2}} e^{\sum_{i=1}^{n} \frac{2\mu x_i}{2\sigma^2}}$$

and

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

Looking at $g(T(x)|\mu)$, we can see that the sufficient statistic is $T(x) = \sum_{i=1}^{n} x_i$ because if we replace $\sum_{i=1}^{n} x_i$ with $T(x)$, can see that g only depends on X through the sufficient statistic.

7. Let $\{X_i\}_{i=1}^n$ be iid observations from a location parameter family with cumulative distribution function $F(x-\theta)$, $-\infty < \theta < \infty$. Show that range of the distribution of $R = max(X_i) - min(X_i)$ does not depend on the parameter θ. Use the fact that $X_1 = Z_1+\theta$, $X_2 = Z_2+\theta$, . . . , $X_n = Z_n+\theta$ and $min(X_i) = min(Z_i+\theta)$ and $max(X_i) = max(Z_i+\theta)$, where $\{Z_i\}_{i=1}^n$ are iid observations from $F(x)$. In other words, show R is ancillary.

Note that $Z_i = X_i - \theta$, which implies that Z has cumulative distribution function $F(x-0)$ or simply $F(z)$. Let $F_R(r|\theta)$ denote the joint distribution of R and θ. Using the definition of a cumulative distribution function, $F_R(r|\theta) = P(R \leq r)$.

$$F_R(r|\theta) = P(max(X_i) - min(X_i) \leq r)$$

$$P(max(X_i) - min(X_i) \leq r) = P(max(Z_i+\theta) - min(Z_i+\theta) \leq r)$$

Since θ is always constant,

$$= P(max(Z_i) - min(Z_i) + \theta - \theta \leq r)$$

$$= P(max(Z_i) - min(Z_i) \leq r)$$

Since Z_i does not depend on θ because cdf does not depend on θ, $R = max(X_i) - min(X_i)$ does not depend on θ. Intuitively, the subtraction of 2 functions that don't depend on θ does not add a θ term into the resulting function.

8.) Show that $N(\mu, \mu^2)$ has a sufficient statistic but is not complete. Find a linear combination that is not trivially 0 for g(T).

Assume there are n observations each iid $\sim N(\mu, \mu^2)$. Then it can be shown that the joint pdf can be factored into the following

$$f(x|\mu) = (\frac{1}{\mu\sqrt{2\pi}})^n e^{\sum_{i=1}^{n} \frac{-x_i^2}{2\mu^2}} e^{\sum_{i=1}^{n} \frac{2\mu x_i}{2\mu^2}} e^{\frac{-n\mu^2}{2\mu^2}}$$

$$f(x|\mu) = (\frac{1}{\mu\sqrt{2\pi}})^n e^{\sum_{i=1}^{n} \frac{-x_i^2}{2\mu^2}} e^{\sum_{i=1}^{n} \frac{x_i}{\mu}} e^{\frac{-n}{2}}$$

The above equation can be broken up into two functions, $g(T(x)|\theta)$ and $h(x)$

$$h(x) = e^{-n/2}$$

$$g(T(X)|\theta) = (\frac{1}{\mu\sqrt{2\pi}})^n e^{\sum_{i=1}^n \frac{-x_i^2}{2\mu^2}} e^{\sum_{i=1}^n \frac{x_i}{\mu}}$$

Looking at $g(T(X)|\theta)$, can see that the sufficient statistic for μ has 2 components. Note that only 1 sufficent statistic is required because 1 parameter, μ, describes both the mean and variance of the normal distribution. Therefore, $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is the sufficient statsitic. The normal distribution is part of the exponential family. This implies that the sufficient statistic found is a complete statistic because each component is the sum of all X_i or transformed X_i and the parameter space is open. In order to show that a sufficient statistic exists but is not complete, a new sufficent statistic is created by adding redunant information. Let $T'(x) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, max(x_i))$ denote the new sufficient statistic. This is a sufficient statistic because it is a function of $T(X)$. Note that

$$(\sum_{i=1}^n a_i)^2 = \sum_{i=1} na_i^2 + 2\sum_{a=1}^n \sum_{b=1}^{a-1} a_a a_b \tag{1}$$

Let $y(T(X))$ denote some transformation of $T(X)$.

$$y(T(X)) = \frac{(-(\sum_{i=1}^n x_i)^2 + \sum_{i=1} nx_i^2 + 2\sum_{a=1}^n \sum_{b=1}^{a-1} x_a x_b)^{1/2}}{n} - max(x_i)$$

Notice that $y(T(X))$ depends on the sufficient statistic, and $2\sum_{a=1}^n \sum_{b=1}^{a-1} x_a x_b$ can be obtained using equation 1. However, equation 1 can be used to turn the expression into

$$y(T(X)) = \frac{[(\sum_{i=1}^n x_i)^2]^{1/2}}{n} - max(x_i) = \frac{\sum_{i=1}^n x_i}{n} - max(x_i)$$

The definition of a complete statistic states that for a given $T(X)$ and a $X \sim f(t|\theta)$, the statistic is complete if $E_\theta(g(T)) = 0$ for any arbitrary θ, then $P(g(T) = 0; \theta) = 1$ for any arbitrary θ.

$$E[y(T(X))] = \frac{1}{n} E(\sum_{i=1}^n x_i) - E(max(x_i))$$

$$E[y(T(X))] = \frac{n\mu}{n} - \mu = 0$$

For $y(T(X))$, the conditional is met. However, $P(y(T) = 0; \theta) = 1$ is not true because in a colletion of n observations, we're not guaranteed that each observation is the same nor the sum of each observation is the same nor the max value in the collection is the expected value.

9.) Show that the poisson distribution is part of the regular exponential family.

Let $\{X_i\} \sim pois(\lambda)$ for $i = 1, 2, 3, ..., n$ be a collection of random variables from poisson distribution. Then, the probabiltiy density function has the form $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, and joint probability density function has the form $P(X = k) = \frac{\lambda^{nk} e^{-\lambda n}}{k^n!}$. A distribution is part of the exponential family if it has the following form

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

for $h(x) > 0$.

$$P(X = k) = \frac{e^{\log \lambda^{nk}} e^{-\lambda n}}{k^n!}$$

$$P(X=k)=\frac{e^{nk\cdot\log\lambda}e^{-\lambda n}}{k^n!}$$
$$P(X=k)=\frac{e^{nk\cdot\log\lambda-\lambda n}}{k^n!}$$

From the above equation, $\psi(\theta)=\psi(\lambda)=\log\lambda$, $A(\theta)=A(\lambda)=n\lambda$, $T(X)=T(k)=nk$, and $h(x)=\frac{1}{k!}>0$. Let $\eta=\log\lambda$. Therefore, the conical form of the collection of poisson random variables is

$$P(X=k)=\frac{e^{nk\eta-ne^{\eta}}}{k^n!}$$

For a regular exponential family, the natural parameter space has to be an open set. Since $\eta=\log\lambda$, the natural parameter space is equal to the range of $\log x$. Since a log function is bounded by $(-infty, infty)$, the parameter space is open because infinity is not included in the range.

11.) Suppose we want to estimate the variance of the Bernoulli distribution $\tau(p)=p(1-p)$ the MLE of this variance is given by $\hat{\tau}(p)=\hat{p}(1-\hat{p})$ where $\hat{p}=\bar{X}$. Using the Delta method, find the approximate distribution $\hat{\tau}$.

Let $\tau(p)=p(1-p)=Var(X_i)$ and $X_i\sim Bernoulli(p)$. Let maximum likelihood estimate of $\tau(p)$ be $\hat{\tau}(p)=\hat{p}(1-\hat{p})$ and $\hat{p}=\bar{X}$. By the central limit theorem, $\sqrt{n}(\bar{X}-p)\sim N(0,\sigma^2)$. Also, $E(X_i)=p$ and $Var(X_i)=p(1-p)$.

$g(p)=p(1-p)=\tau(p)$ $g(\bar{X})=\hat{\tau}(p)=\hat{p}(1-\hat{p})=\bar{X}(1-\bar{X})$ By the delta method, $\sqrt{n}(g(\bar{X})-g(p))\sim N(0,\sigma^2 g'(p)^2$. Note that $g'(p)=(1-p)-1=1-2p$ and $\sigma^2=p(1-p)$. Therefore, $\sqrt{n}(g(\bar{X})-p(1-p))\sim N(0,p(1-p)(1-2p)^2$. Therefore, $Var(\hat{\tau}(p))=p(1-p)(1-2p)^2$, $E(g(\hat{\tau}(p)))=p(1-p)$, and $\hat{\tau}(p)\sim N(p(1-p),p(1-p)(1-2p)^2)$

13.) Find the differential entropy of a multivariate normal distribution.

$H(X)=-\int_{-\infty}^{\infty}N(x;\mu,\Sigma)\ln(N(x;\mu,\Sigma)dx=-E[\ln(N(x;\mu,\Sigma))]=-E[\ln[\frac{1}{\sqrt{2^d\pi^d\det(\Sigma)}}e^{\frac{-1}{2}(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)}]]$ where d is the dimension of the distribution and Σ is a positive semi definite covariance matrix. $=-\ln(\frac{1}{\sqrt{2^d\pi^d\det(\Sigma)}})-E[\ln(e^{\frac{-1}{2}(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)})]=-\ln(1)+ln(\sqrt{2^d\pi^d\det(\Sigma)})-E[\frac{-1}{2}(x-\mu)^{\tau}\Sigma^{-1}(x-\mu))]$ $=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}E[(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)]$ Note that $\frac{1}{2}E[(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)]=\frac{1}{2}E[Trace[(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)]$ because $(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)$ is scalar. It is the mahalanobis distance. $=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}E[Trace[(x-\mu)^{\tau}\Sigma^{-1}(x-\mu)]=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}E[Trace[\Sigma^{-1}(x-\mu)^{\tau}(x-\mu)]$ $=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}Trace[E[\Sigma^{-1}(x-\mu)^{\tau}(x-\mu)]]$ Since Σ^{-1} is constant, $=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}Trace[\Sigma^{-1}E[(x-\mu)^{\tau}(x-\mu)]]=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}Trace[\Sigma^{-1}\Sigma]=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{1}{2}Trace[I]$ $H(X)=\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln(det(\Sigma))+\frac{d}{2}$

Application of sufficiency

Suppose we were constructing the running average with no buffer in a stream of data.

1. What would be the minimal amount of information required to reconstruct the set of data assuming it came from a Normal distribution?

Factorization Theorem extended to 2 parameter: Let $X_1, X_2, ..., X_n$ denote random variables with a joint pdf or pmf $f(x_1, x_2, x_3, ..., x_n; \theta_1, \theta_2)$, which depends on the parameters θ_1 and θ_2. Then, the statistic $Y_1 = u_1(X_1, X_2, ..., X_n)$ and $Y_2 = u_2(X_1, X_2, ..., X_n)$ are joint sufficient statistics for θ_1 and θ_2 if and only if

$$f(x_1, x_2, x_3, ..., x_n; \theta_1, \theta_2) = \phi[u_1(X_1, X_2, ..., X_n), u_2(X_1, X_2, ..., X_n); \theta_1, \theta_2] \cdot h(x_1, x_2, ..., x_n)$$

where ϕ is a function that depends on $x_1, x_2, x_3, ..., x_n$ only through the functions $u_1(X_1, X_2, ..., X_n)$ and $u_2(X_1, X_2, ..., X_n)$, and the function $h(x_1, x_2, ..., x_n)$ does not depend on either parameters θ_1 nor θ_2.

$$f(x|\mu\sigma^2) = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{\sum_{i=1}^n \frac{-x_i^2}{2\sigma^2}} e^{\sum_{i=1}^n \frac{\mu x_i}{\sigma^2}} e^{\frac{-n\mu^2}{2\sigma^2}}$$

Let $T(X)$ denote the joint sufficient statistic for μ and σ^2. From the above equation, $T(X) = (\sum_i X_i^2, \sum_i X_i, n)$. The reason is included is because n is not constant in this situation. The data set is constantly growing as new data is being streamed. Knowing the sufficient statistic is all the information needed to reconstruct the distribution.

2. What is the complete and sufficient statistic for the distribution?

The normal distribution is part of the exponential family. $T(X)$ can be rewritten as $T(X) = (\sum_i X_i^2, \sum_i X_i, \sum_i X_i/X_i)$ because $n = \sum_i X_i/X_i$. Theorem 6.2.25, from Casella and Berger Statsitical Inference textbook, states that a sufficient statistic from an exponential family is complete as long as the parameter space is open and $T(X)$ is of the form $T(X) = (\sum_i t_1(X_i), \sum_i t_2(X_i), \sum_i t_3(X_i))$. The parameter space for is an open set because $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$. Therefore, the sufficient statistic found in part a is complete.