ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 4: Sufficient Statistics

Consider a random variable X whose distribution p is parametrized by $\theta \in \Theta$ where θ is a scalar or a vector. Denote this distribution as $p_X(x|\theta)$ or $p(x|\theta)$, for short. In many signal processing applications we need to make some decision about θ from observations of X, where the density of X can be one of many in a family of distributions, $\{p(x|\theta)\}_{\theta \in \Theta}$, indexed by different choices of the parameter θ.

More generally, suppose we make n independent observations of X: $X_1, X_2, \ldots, X_n$ where $p(x_1 \ldots x_n|\theta) = \prod_{i=1}^n p(x_i|\theta)$. These observations can be used to infer or estimate the correct value for θ. This problem can be posed as follows. Let $x = [x_1, x_2, \ldots, x_n]$ be a vector containing the n observations.

Question: Is there a lower dimensional function of x, say $t(x)$, that alone carries all the relevant information about θ? For example, if θ is a scalar parameter, then one might suppose that all relevant information in the observations can be summarized in a scalar statistic.

Goal: Given a family of distributions $\{p(x|\theta)\}_{\theta \in \Theta}$ and one or more observations from a particular distribution $p(x|\theta^*)$ in this family, find a data compression strategy that preserves all information pertaining to θ^*. The function identified by such strategyis called a *sufficient statistic*.

1 Sufficient Statistics

Example 1 (Binary Source) *Suppose X is a 0/1 - valued variable with $\mathbb{P}(X = 1) = \theta$ and $\mathbb{P}(X = 0) = 1 - \theta$. That is $X \sim p(x|\theta) = \theta^x(1-\theta)^{1-x}, (x \in [0, 1])$.*

We observe n independent realizations of X: $x_1, \ldots, x_n$ with $p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^k(1-\theta)^{n-k}$; $k = \sum_{i=1}^n x_i$ (number of 1's).
Note that $K = \sum_{i=1}^n X_i$ is a random variable with values in $\{0, 1 \ldots, n\}$

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \text{ a binomial distribution with } \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

The joint probability mass function of $(X_1, \ldots, X_n)$ and K is

$$p(x_1, \ldots, x_n, k|\theta) = \begin{cases} p(x_1, \ldots, x_n|\theta); & \text{if } k = \sum x_i \\ 0; & \text{otherwise} \end{cases}$$

$$\Rightarrow p(x_1, \ldots, x_n|k, \theta) = \frac{p(x, k|\theta)}{p(k|\theta)}$$

$$= \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \frac{1}{\binom{n}{k}}$$

$\Rightarrow$ conditional prob of $X_1, \ldots, X_n$ given $\sum x_i$ is uniformly distributed over the $\binom{n}{k}$ sequences that have exactly k 1's. In other words, the condition distribution of $X_1, \ldots, X_n$ given k is independent of θ. So k carries all relevant info about θ!

Note: $k = \sum x_i$ compresses $\{0, 1\}^n$ (n bits) to $\{0, \ldots, n\}$ ($\log n$ bits).

Definition 1 *Let X denote a random variable whose distributon is parametrized by $\theta \in \Theta$. Let $p(x|\theta)$ denote the density of mass function. A statistic $t(X)$ is* sufficient *for θ if the distribution of X given $t(X)$ is independent of θ; i.e., $p(x|t, \theta) = p(x|t)$*

Theorem 1 (Fisher-Neyman Factorization) *Let X be a random variable with density $P(x|\theta)$ for some $\theta \in \Theta$. The statistic $t(X)$ is sufficient for θ iff the density can be factorized into a function $a(x)$ and a function $b(t,\theta)$, a function of θ but only depending on x through the $t(x)$; i.e.,*

$$p(x|\theta) = a(x)b(t,\theta)$$

Proof: (if/sufficiency) Assume $p(x|\theta) = a(x)b(t|\theta)$

$$p(t|\theta) = \int_{x:t(x)=t} p(x|\theta)dx = \left(\int_{x:t(x)=t} a(x)dx\right) b(t,\theta)$$

$$\begin{aligned} p(x|t,\theta) &= \frac{p(x,t|\theta)}{p(t|\theta)} = \frac{p(x|\theta)}{p(t|\theta)} \\ &= \frac{a(x)}{\int_{x:t(x)=t} a(x)dx} \text{ independent of } \theta \\ &\Rightarrow t(x) \text{ is a sufficient statistic} \end{aligned}$$

(only if/necessity) If $p(x|t,\theta) = p(x|t)$ independent of θ then $p(x|\theta) = p(x|t,\theta)p(t|\theta) = \underbrace{p(x|t)}_{a(x)}\underbrace{p(t|\theta)}_{b(t,\theta)}$

Example 2 (Binary Sourse) $p(x|\theta) = \theta^k(1-\theta)^{n-k} = \underbrace{\frac{1}{\binom{n}{k}}}_{a(x)}\underbrace{\binom{n}{k}\theta^k(1-\theta)^{n-k}}_{b(k,\theta)} \Rightarrow k$ *is sufficient for θ.*

Example 3 (Poisson) *Let λ be an average number of packets/sec sent over a network. Let X be a random variable representing number of packets seen in 1 second. Assume $\mathbb{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!} =: p(x|\lambda)$.*
Given $X_1, \ldots, X_n$,

$$p(x_1,\ldots,x_n|\lambda) = \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = \underbrace{\prod_{i=1}^n \frac{1}{x_i}}_{a(x)} \underbrace{e^{-n\lambda}\lambda^{\sum x_i}}_{b(\sum x_i,\lambda)} .$$

So $\sum_{i=1}^n x_i$ is a sufficient statistic for λ.

Example 4 (Gaussian) *$X \sim \mathcal{N}(\mu,\Sigma)$ is d-dimensional.*
$X_1,\ldots,X_n \overset{iid}{\sim} \mathcal{N}(\mu,\Sigma)$; $\theta = (\mu,\Sigma)$

$$\begin{aligned} p(x_1,\ldots,x_n|\theta) = \prod_{i=1}^n p(x_i;\theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi^d|\Sigma|}} e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)} \\ &= 2\pi^{-nd/2}|\Sigma|^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)} \end{aligned}$$

Define sample mean

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

and sample covariance

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

$$\begin{aligned}
\exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)) &= \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu}+\hat{\mu}-\mu)^T\Sigma^{-1}(x_i-\hat{\mu}+\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu})^T\Sigma^{-1}(x_i-\hat{\mu}) - \sum_{i=1}^{n}(x_i-\hat{\mu})^T\Sigma^{-1}(\hat{\mu}-\mu) - \frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu})^T\Sigma^{-1}(x_i-\hat{\mu}))\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}tr(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T))\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)) \\
&= \exp(-\frac{1}{2}tr(\Sigma^{-1}(n\hat{\Sigma})))\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu))
\end{aligned}$$

Note that the second term on the second line is zero because $\frac{1}{n}\sum_i x_i = \widehat{\mu}$. *For any matrix* B, $tr(B)$ *is the sum of the diagonal elements. On the fourth line above we use the* trace *property,* $tr(AB) = tr(BA)$.

$$p(x_1,\ldots,x_n|\theta) = \underbrace{2\pi^{-nd/2}|\Sigma|^{-n/2}\exp(-\frac{1}{2}\sum_{i=1}^{n}(\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu))\exp(-\frac{1}{2}tr(\Sigma^{-1}n\hat{\Sigma}))}_{b(\hat{\mu},\hat{\Sigma},\theta)}\cdot\underbrace{1}_{a(x_1,\ldots,x_n)}$$

2 Minimal Sufficient Statistic

Definition 2 *A sufficient statistic is* minimal *if the dimension of* $T(X)$ *cannot be further reduced and still be sufficient.*

Example 5 $X \sim \mathcal{N}(0,1)$ *and* $X_1,\ldots,X_n \overset{iid}{\sim} \mathcal{N}(0,1)$

$$u(x_1,\ldots,x_n) = [x_1+x_2,\ldots,x_{n-1}+x_n]^T \quad u \text{ is a } n/2\text{-dimensional statistic}$$

$$T(x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i \quad \text{a 1-dimensional statistic}$$

T *is sufficient, and* $T = \sum_{i=1}^{n/2} u_i \Rightarrow u$ *is sufficient.*

3 Rao-Blackwell Theroem

Theorem 2 *Assume* $X \sim p(x|\theta)$, $\theta \in \mathbb{R}$, *and* $t(X)$ *is a sufficient statistic for* θ. *Let* $f(x)$ *be an estimator of* θ *and consider the mean square error* $\mathbb{E}[(f(x)-\theta)^2]$. *Define* $g(t(X)) = \mathbb{E}[f(X)|t(X)]$.

Then $\mathbb{E}[(g(t(X))-\theta)^2] \leq \mathbb{E}[(f(X)-\theta)^2]$, *with equality iff* $f(X) = g(t(X))$ *with probability 1; i.e., if the function* f *is equal to* g *composed with* t.

Proof: First note that because $t(X)$ is a sufficient statistic for θ, it follows that $g(t(X)) = \mathbb{E}[f(X)|t(X)]$ does not depend on θ, and so it too is a valid estimator (i.e., if $t(X)$ were not sufficient, then $g(t(X))$ might be a function of $t(X)$ and θ and therefore not computable from the data alone).

Next recall the following basic facts about conditional expectation. Suppose X and Y are random variables. Then

$$\mathbb{E}[X|Y] = \int xp(x|y)dx$$

In the present context

$$\mathbb{E}[f(X)|t(X)] = \int f(x)p(x|t)dx$$

where $p(x|t)$ is conditional density of X given $t(X) = t$. Furthermore, for any random variables X and Y

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y)]] &= \int \underbrace{\mathbb{E}[X|Y=y]}_{h(y)} p(y)dy \\ &= \int \left(\int xp(x|y)dx \right) p(y)dy \\ &= \int x \left(\int p(x|y)p(y)dy \right) dx \\ &= \int xp(x)dx = \mathbb{E}[X]\end{aligned}$$

This is sometimes called the *smoothing* property.

Now consider the conditional expectation

$$\mathbb{E}[f(X) - \theta|t(X)] = g(t(X)) - \theta$$

Also

$$(\mathbb{E}[f(X) - \theta|t(X)])^2 \leq \mathbb{E}[(f(X) - \theta)^2|t(X)] \text{ by Jensen's inequality}$$

Jensen's inequality (see general statement below) implies that the expectation of a squared random variable is greater or equal to than the square of its expected value. So

$$(g(t(X)) - \theta)^2 \leq \mathbb{E}[(f(X) - \theta)^2|t(X)]$$

Take expectation of both sides (recall the smoothing property above) yields

$$\mathbb{E}[(g(t(X)) - \theta)^2] \leq \mathbb{E}[(f(X) - \theta)^2]$$

4 Jensen's Inequality

Suppose that ϕ is a convex function; $\lambda\phi(x) + (1-\lambda)\phi(y) \geq \phi(\lambda x + (1-\lambda)y)$.
Then

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$$

average of convex functions $\geq$ convex function of average

Example 6

$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$$

$$mean^2 + var \geq mean^2$$