MATH3911 Summary

Alexander Kwok

August 14, 2014

Lecture 1: The Subject of Statistical Inference

Statistical Models

In general, we can view the statistical model as the triplet ($\mathcal{X}$, $\mathcal{P}$, Θ) where:

- $\mathcal{X}$ is the sample space (i.e. the set of all possible realizations $\mathbf{X} = (X_1, X_2, \cdots, X_n)$)
- $\mathcal{P}$ is a family of model functions $P_\theta(\mathbf{X})$ that depend on the unknown parameter θ
- Θ is the set of possible θ- values, i.e. the parameter space indexing the model

Lecture 2: Sufficient statistic

Definition 1. (sufficient statistic) *$P(X = x|T = t)$ is a function of x and t only. (i.e. is not a function of θ).* The intuition behind the sufficient statistic concept is that it contains all the information necessary for estimating θ.

Definition 2. *Let $X_1, \cdots, X_n$ be iid RVs whose distribution is the pdf f_X, or the pmf p_{x_i}. The likelihood function is the product of the pdfs or pmfs*

$$\mathrm{L}(x_1, \cdots, x_n|\theta) = \begin{cases} \prod_{i=1}^n f_{X_i}(x_i) & \text{if } X_i \text{ is a continuous RV} \\ \prod_{i=1}^n p_{X_i}(x_i) & \text{if } X_i \text{ is a discrete RV} \end{cases}$$

The likelihood function is sometimes viewed as a function of $x_1, \cdots, x_n$ (fixing θ) and sometimes as a function of θ (fixing $x_1, \cdots, x_n$). In the latter case, the likelihood is sometimes denoted $\mathrm{L}(\theta)$.

Definition 3. (sufficiency principle) The sufficiency principle implies that if T is sufficient for θ, then if x and y are such that $T(x) = T(y)$, then inference about θ should be the same whether $X = x$ or $Y = y$ is observed.

Theorem 1 (Neyman Fisher Factorization Criterion) *T is a sufficient statistic for θ if the likelihood factorizes into the following form*

$$L(x_1, \cdots, x_n|\theta) = g(\theta, T(x_1, \cdots, x_n)) \cdot h(x_1, \cdots, x_n)$$

for some functions g,h.

Proof. (For the discrete case)

$$p(x_1, \cdots, x_n|T(x_1, \cdots, x_n)) = \frac{p(x_1, \cdots, x_n, T(x_1, \cdots, x_n))}{p(T(x_1, \cdots, x_n))} = \frac{p(x_1, \cdots, x_n)}{\sum_{y:T(y)=T(x)} p(y_1, \cdots, y_n)} = \frac{h(x_1, \cdots, x_n)}{\sum_{y:T(y)=T(x)} h(y_1, \cdots, y_n)}$$

which is not a function of θ. Conversely, assume that T is a sufficient statistic for θ. Then

$$L(x_1, \cdots, x_n|\theta) = p(x_1, \cdots, x_n|T(x_1, \cdots, x_n), \theta) = h(x_1, \cdots, x_n)g(T(x_1, \cdots, x_n), \theta)$$

Example (*normal population, unknown mean, known variance*) The joint density is

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

Since σ^2 is known, we can let

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

and

$$g(T(X),\mu)=\exp\left(-\frac{n\mu^2}{2\sigma^2}+\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i\right)$$

By the factorization theorem this shows that $\sum_{i=1}^{n}X_i$ is a sufficient statistic. It follows that the sample mean $\bar{X}_n$ is also a sufficient statistic.

Example (*Uniform distribution*) Suppose that X_i are uniformly distributed on [0, θ] where θ is unknown. Then the joint density is

$$f(x_1,\cdots,x_n|\theta)=\theta^{-n}1(x_i\leq\theta,i=1,2,\cdots,n)$$

Here 1(E) is an indicator function. It is 1 if the event E holds, 0 if it does not. Now $x_i\leq\theta$ for $i=1,2,\cdots,n$ if and only if $\max\{x_1,x_2,\cdots,x_n\}\leq\theta$. So we have

$$f(x_1,\cdots,x_n|\theta)=\theta^{-n}1(max\{x_1,x_2,\cdots,x_n\}\leq\theta)$$

By the factorization theorem this shows that

$$T=max\{X_1,X_2,\cdots,X_n\}$$

is a sufficient statistic.

Definition 4. (Minimal sufficient statistic) *A sufficient statistic allows the greatest data reduction without loss of information on θ.* A sufficient statistic is not uniquely defined. From the factorization theorem it is easy to see that (i) the identity function $T(x_1,\cdots,x_n)=(x_1,\cdots,x_n)$ is a sufficient statistic vector and (ii) if T is a sufficient statistic for θ then so any 1-1 function of T. A function that is not 1-1 of a sufficient static may or may not be a sufficient statistic. This leads to the notion of a minimal sufficient statistic.

Example: Since $T=(\sum X_i,\sum X_i^2)$ are jointly sufficient statistics for $\theta=(\mu,\sigma^2)$ for normally distributed data $X_1,\cdots,X_n\sum N(\mu,\sigma^2)$, then so are $(\bar{X},S^2)$ which are a 1-1 function of $(\sum X_i,\sum X_i^2)$.

Theorem 2. (Simplified version of Theorem by Lehmann- Scheffe) Suppose there exists a function **T(x)** such that, for two sample points **x** and **y**, the ratio $\frac{L(x,\theta)}{L(y,\theta)}$ is constant as a function of θ if and only if **T(x)=T(y)**. Then **T(X)** is a minimal sufficient statistic for θ.

Example: T=$(\sum X_i,\sum X_i^2)$ is a minimal sufficient statistic for the Normal distribution since the likelihood ratio is not a function of θ iff T(x)= T(y).

$$\frac{L(x_1,\cdots,x_n|\theta)}{L(y_1,\cdots,y_n|\theta)}=e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2-(y_i-\mu)^2}=e^{-\frac{1}{2\sigma^2}(\sum x_i^2-\sum y_i^2)+\frac{\mu}{\sigma^2}(\sum x_i-\sum y_i)}$$

When T(x)=T(y), it is not a function of θ nor zero function. Since ($\bar{X}$, S^2) is a function of T, it is minimal sufficient statistic as well.

Definition 5. (One parameter exponential family densities)

$$f(x,\theta)=a(\theta)b(x)\exp(c(\theta)d(x))$$

with $c(\theta)$ strictly monotone and $T=\sum_{i=1}^{n}d(X_i)$ is minimal sufficient.

Lecture 3: Maximum Likelihood Inference

Definition 1 (Likelihood principle) In the inference about θ, after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. In particular,

"Data sets with proportional likelihood functions should lead to identical conclusions"

Definition 2: (Likelihood function) The likelihood of a set of parameter values, θ, given some observed outcomes, x, is equal to the probability of those observed outcomes given those parameter values, i.e.

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

Definition 3: (Maximum Likelihood Estimation) The Maximum Likelihood Estimator (MLE) is defined as

$$\hat{\theta} = \arg[\sup_{\theta \in \Theta} L(x, \theta)]$$

Remark: It is more convenient to maximize not the function $L(x, \theta)$ but its logarithm. The function $\log L(x, \theta)$ is called the Log-likelihood. Since the logarithm is a monotone increasing function, maximization of $L(x, \theta)$ or of $\log L(x, \theta)$ is achieved for the same value of the argument $\hat{\theta}$.

Definition 4: (Normed likelihood) When comparing different models,

$$R(x, \theta) = \frac{L(x, \theta)}{L(x, \hat{\theta})}$$

which has a range in [0, 1]. An even more often used measure is the deviance $D(\theta)$ which is defined as

$$D(\theta) = -\log R(\theta) = -2[\log L(x, \theta) - \log L(x, \hat{\theta})]$$

The deviance is a non- negative number which can be attached to each model indexed by the parameter θ. The larger the deviance, the further the model from the "most likely" model.

Definition 5: (Fisher Information) can be defined as the variance of the score, or as the expected value of the observed information.

- We define $V(\mathbf{X}, \theta) = \frac{\partial}{\partial \theta} \log L(\mathbf{X}, \theta)$ to be the score function for non-i.i.d. random variable where $L(\mathbf{X}, \theta)$ is the joint density. This indicates how sensitively a likelihood function $L(\theta X)$ depends on its parameter θ.
- In the case where $\mathbf{X}= (X_1, X_2, \cdots, X_n)$ and the X_i are i.i.d. with a density $f(x, \theta)$ then $V(\mathbf{X}, \theta) = \sum_{i=1}^{n} \frac{\partial / \partial \theta f(X_i, \theta)}{f(X_i, \theta)}$

Properties:

- If $\hat{\theta}$ is the MLE then $V(x, \hat{\theta}) = 0$ holds. (This holds because $\hat{\theta}$ maximises $\log L(\mathbf{X}, \theta)$ with respect to θ.)
- $E_\theta(V(\mathbf{X}, \theta)) = 0$ holds under suitable regularity conditions.

Definition 6: (Expected Fisher Information) about θ contained in the vector $\mathbf{X}$:
It is denoted by $I_X(\theta)$ and is defined as

$$I_X(\theta) = Var_\theta(V(\mathbf{X}, \theta)) = E_\theta \left\{ \frac{\partial}{\partial \theta} \ln L(\mathbf{X}, \theta) \right\}^2$$

(where we utilized the fact that $E_\theta(V, \mathbf{X}, \theta)) = 0$)
Proof:

$$\mathbb{E}(V|\theta) = \int_{-\infty}^{\infty} \frac{\frac{\partial f(x;\theta)}{\partial \theta}}{f(x;\theta)} f(x;\theta) dx = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta} dx$$

If certain differentiability conditions are met, the integral may be rewritten as

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x;\theta) dx = \frac{\partial}{\partial \theta} 1 = 0$$