

Large Sample Maximum Likelihood

Lecture #11: Tuesday, 8 February 2005
Lecturer: Prof. Charles Elkan
Scribe: Stephen Krotosky

1 Large Sample Maximum Likelihood: Properties of Score Function

We have n i.i.d. training examples x_i , for $i = 1 \dots n$. We define $x = (x_1, \dots, x_i, \dots, x_n)$ as the collection of the i.i.d. training examples. Each $x_i \sim p^*(x_i, \theta)$ and $x \sim \prod_i p^*(x_i, \theta)$.

Lemma: For every θ , $s(x_i, \theta) = \frac{\partial}{\partial \theta} \log p^*(x_i, \theta)$ has zero mean.

Corollary: $\text{var}[s(x_i, \theta)] = E[s(x_i, \theta)^2] = I$ where I is the Fisher information of a *single* example x_i .

Lemma: $I = E[-\frac{\partial}{\partial \theta} s(x_i, \theta)]$

Proof:

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \log p^*(x_i, \theta) &= \frac{\partial^2}{\partial \theta^2} \log p \\ &= \frac{\partial}{\partial \theta} \left[\frac{1}{p} \frac{\partial p}{\partial \theta} \right] \\ &= - \left[\frac{1}{p} \frac{\partial p}{\partial \theta} \right]^2 + \frac{1}{p} \frac{\partial^2 p}{\partial \theta^2} \\ &= A + B \end{aligned}$$

Let's Consider A and B separately:

$$A = - \left[\frac{1}{p} \frac{\partial p}{\partial \theta} \right]^2 = - \left[\frac{\partial}{\partial \theta} \log p \right]^2 = -s(x_i, \theta)^2$$

We want $E[B] = 0$

$$E[B] = \int_x p \frac{\partial^2 p}{\partial \theta^2} dx = \int_x \frac{\partial^2 p}{\partial \theta^2} dx = \frac{\partial^2}{\partial \theta^2} \int_x p dx = \frac{\partial^2}{\partial \theta^2} 1 = 0$$

So

$$E[-\frac{\partial}{\partial \theta} s(x_i, \theta)] = E[-(A + B)] = E[s(x_i, \theta)^2]$$

2 Weak law of large numbers

Informally stated, when n is large, for most i.i.d samples $(x_1 \dots x_n)$, the observed average of $f(x_i)$ is close to $E[f(x)]$. This is true for any function f with *finite variance*.

Theorem: Let $x_1 \dots x_n$ be i.i.d. random variables with mean μ and finite variance $\sigma^2 > 0$. Let $S_n = \sum_{i=1}^n x_i$. Then for any $\epsilon > 0$,

$$p\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

More precisely: For every $\epsilon > 0$ for every $\delta > 0$, there exists an m such that for every $n \geq m$

$$p\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \geq 1 - \delta$$

This is a formalization of informal "for large n with high probability". This is only one possible formulation. If we were to prove this rigorously, we need to carry ϵ , δ , m , and n through the proofs. This makes the proofs too tedious and difficult for the scope of this lecture.

Proof: Omitted but assumed true.

Lemma: For large n with high probability, the MLE $\hat{\theta}$ is near the true value θ_0 .

Proof: Consider $f(x_i) = \log p^*(x_i, \theta)$ for any arbitrary θ .

$$E[f(x_i)] = z(\theta) = E[\log p^*(x_i, \theta) | x_i \sim \theta_0]$$

The Weak Law of Large Numbers says that for large n , with high probability:

$$\frac{1}{n} \sum \log p^*(x_i, \theta) \text{ is near } z(\theta)$$

To obtain this lemma, we need to make more assumptions:

Suppose the above is uniformly true for all θ i.e. we have the same $m(\epsilon, \delta)$ for all θ . Let $n \geq m(\epsilon, \delta)$.

$$\text{Suppose } \left| \frac{1}{n} \sum \log p^*(x_i, \theta) - z(\theta) \right| \leq \epsilon \text{ for all } \theta$$

$$\text{Let } \hat{\theta} = g(x) = \operatorname{argmax}_{\theta} \frac{1}{n} \sum \log p^*(x_i, \theta)$$

Consider $|Z(\hat{\theta}) - z(\theta_0)| \leq 2\epsilon$.

This is true by the Lemma:

$$|y(\theta) - z(\theta)| < \epsilon$$

$$y(\hat{\theta}) \geq y(\theta_0)$$

$$|y(\theta_0) - z(\theta_0)| < \epsilon$$

$$z(\theta_0) \geq z(\theta)$$

$$|z(\hat{\theta}) - z(\theta_0)| \leq 2\epsilon \text{ for all } \theta \neq \hat{\theta}$$

Lemma: If $z(\hat{\theta})$ is close to $z(\theta_0)$, then $\hat{\theta}$ is close to θ_0 .

$$z(\theta) = E[\log p^*(x_i, \theta)]$$

We claim that

$$z(\theta_0) \geq z(\text{any other } \theta)$$

The proof of this will be in the next lecture.

3 Central Limit Theorem

Let $Y_1 \dots Y_n$ be i.i.d. random variables with mean μ and variance $\sigma^2 > 0$.

Let $S_n = \sum_{i=1}^n Y_i$, then :

$$\lim_{n \rightarrow \infty} p\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq z\right) = \Phi(z) \text{ for all } z$$

Where $\Phi(z)$ is the cdf of a standard Gaussian distribution.

This means as $n \rightarrow \infty$, this transformation of S_n has a Gaussian Distribution.

Informally, $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ tends to an $N(0, 1)$ distributions.

S_n tends to $N(n\mu, n\sigma^2)$.

Application: Consider $\frac{1}{\sqrt{n}} \sum s(x_i, \theta)$. $\mu = E[s(x_i, \theta)] = 0$ and $\sigma^2 = \text{var}[s(x_i, \theta)] = I$.

Then as $n \rightarrow \infty$, $\frac{1}{\sqrt{n}} \sum s(x_i, \theta) \rightarrow N(0, n\sigma^2 \frac{1}{n}) = N(0, I)$.

4 Taylor expansion

Suppose $f(x)$ has continuous derivatives up to $(n+1)^{\text{th}}$ order. Then the Taylor expansion around a is:

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)(x - a)^2}{2!} + \dots + \frac{f^{(n)}(a)(x - a)^n}{n!} + R_n$$

When $|x - a| < 1$, $(x - a)^n \rightarrow 0$ as $n \rightarrow \infty$ and higher order terms become negligible.

If $|x - a| \ll 1$, $f(x) \approx f(a) + f'(a)(x - a)$.

We wish to use the above approximation to do a Taylor expansion of the score function $s(x, \theta)$ around the true θ_0 and apply this to $\hat{\theta}$.

$$s(x, \hat{\theta}) = s(x, \theta_0) + (\hat{\theta} - \theta_0) \frac{\partial}{\partial \theta} s(x, \theta_0) + R$$

where R involves $(\hat{\theta} - \theta_0)^2$.