CPSC 550: Machine Learning II **2008/9 Term 2**

Lecture 5 — Feb 3, 2009

Lecturer: Nando de Freitas *Scribe: David Duvenaud*

This lecture proves the consistency of the maximum likelihood estimator (MLE), and also introduces the Lebesgue Integral.

Consistency of the MLE Estimator

Proof outline:

First, we will show that using the MLE will cause the data to have the same likelihood as under the true parameter. Then we will show that if the model is identifiable, then an estimate that gives the same likelihood as the true parameter must be the true parameter. Throughout this, we assume that the data has been generated by the distribution $f(x|\theta_0)$ for some true parameter θ_0.

The MLE estimate is defined as

$$\begin{aligned}
\hat{\theta}_n &= \operatorname*{argmax}_\theta \prod_{i=1}^{n} f(x_i|\theta) && (5.1)\\
&= \operatorname*{argmax}_\theta \sum_{i=1}^{n} \log f(x_i|\theta) && (5.2)\\
&= \operatorname*{argmax}_\theta \tfrac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta) && (5.3)\\
&= \operatorname*{argmax}_\theta \tfrac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta) - \tfrac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta_0) && (5.4)\\
&= \operatorname*{argmax}_\theta \tfrac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i|\theta)}{f(x_i|\theta_0)} && (5.5)\\
&= \operatorname*{argmax}_\theta M_n(\theta) && (5.6)\\
& && (5.7)
\end{aligned}$$

where θ_0 is the true parameter, and

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(x_i|\theta)}{f(x_i|\theta_0)}$$

Mean Likelihood Converges to KL-Divergence

The intuition behind convergence is that, as $n \to \infty$,

$$\begin{aligned}
M_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(x_i|\theta)}{f(x_i|\theta_0)} && (5.8)\\
&\to \mathbb{E}_{\theta_0}[\log \frac{f(x_i|\theta)}{f(x_i|\theta_0)}] && (5.9)\\
&= \int \log \frac{f(x_i|\theta)}{f(x_i|\theta_0)} f(x|\theta_0)dx && (5.10)\\
&= -\int \log \frac{f(x_i|\theta_0)}{f(x_i|\theta)} f(x|\theta_0)dx && (5.11)\\
&= -KL[f(x_i|\theta_0)||f(x_i|\theta)] && (5.12)\\
&= -KL[\theta_0, \theta] && (5.13)\\
& && (5.14)
\end{aligned}$$

So that, by the weak law of large numbers,

$$M_n(\theta) \xrightarrow{p} M(\theta) = -KL(\theta_0, \theta)$$

And therefore when we take an argmax over $M_n(\theta)$, we will be minimizing KL-divergence, which is at a minimum when we pick $\theta = \theta_0$ because $KL(\theta_0, \theta_0) = 0$.

Max-Likelihood gives Minimum Divergence

First we will show that using the MLE $\hat{\theta}_n$ will cause the likelihood to asymptotically converge to the likelihood under the true parameter θ_0.

Theorem 9.13

Suppose that

1. $\sup_\theta |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$

   or equivalently:

   For all θ, $M_n(\theta) \xrightarrow{p} M(\theta)$

2. For every $\epsilon > 0$, $\sup_{\theta:|\theta-\theta_0|\geq\epsilon} M(\theta) < M(\theta_0)$

   (Note: this condition is equivalent to identifiability)

   To put this condition another way:

   For any $\epsilon > 0$, with $|\theta - \theta_0| \geq \epsilon$, there exists $\delta > 0$ such that $M(\theta) < M(\theta_0) - \delta$

Given these two conditions, we have that for the MLE:

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

or equivalently,

$$\text{plim}(\hat{\theta}_n) = \theta_0$$

or equivalently,

$$\lim_{n\longrightarrow\infty} P(|\hat{\theta}_n - \theta_0| > \alpha) \rightarrow 0$$

for every α.

Proof:

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &= M_n(\theta_0) - M(\hat{\theta}_n) + M(\theta_0) - M_n(\theta_0) && (5.15)\\ &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + M(\theta_0) - M_n(\theta_0) && (5.16)\end{aligned}$$

(The above uses the fact that
$M_n(\theta_0) \leq M_n(\hat{\theta}_n)$, since $\hat{\theta}_n = \text{argmax}_\theta\, M_n(\theta)$).
Using condition 1, and because all terms go to zero,

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq \sup_\theta |M_n(\theta) - M(\theta)| + M(\theta_0) - M_n(\theta_0) && (5.17)\\ &\xrightarrow{p} 0 && (5.18)\end{aligned}$$

□

Now that we have proven that the MLE will converge to the same likelihood as the true parameter, all that remains is to show that, if we have identifiability, this implies that the MLE is converging to the true parameter.

Proof: For $\delta > 0$:

$$P(|M(\theta_0) - M(\hat{\theta}_n)|) > \delta \xrightarrow{p} 0 \tag{5.19}$$

$$P(M(\hat{\theta}_n) < M(\theta_0) - \delta) \xrightarrow{p} 0 \tag{5.20}$$

$$\tag{5.21}$$

Using condition 2 above,

$$P(|\hat{\theta}_n - \theta_0|) \geq \epsilon) \leq P[M(\hat{\theta}_n) < M(\theta_0) - \delta] \xrightarrow{p} 0$$

$$\therefore \hat{\theta}_n \xrightarrow{p} \theta_0$$

□

The Riemann Integral

The Riemann integral over the interval $[a, b]$,

$$I = \int_a^b f(x)dx$$

is defined as the limit of taking smaller and smaller vertical slices of f(x), and summing their area. It exists only if

$$\overline{I} \triangleq \sum_{i=1}^{n} \max(f(t_i : t_{i+1}))\Delta t_i = \sum_{i=1}^{n} \min(f(t_i : t_{i+1}))\Delta t_i \triangleq \underline{I}$$

That is, if the upper integral converges to the lower integral in the limit as we take smaller and smaller slices.

There exist functions for which this integral is not defined. For instance, let

$$f(x) = I_{\mathbb{Q}}(x) = \left\{ \begin{array}{ll} 1 & \text{if x is rational} \\ 0 & \text{if x is irrational} \end{array} \right\}$$

In this case, each slice of $f(x)$ will contain both a rational and an irrational number, so over any interval $\max(f(x)) = 1$ and $\min(f(x)) = 0$. Thus the $\overline{I} \neq \underline{I}$ and the integral does not exist.

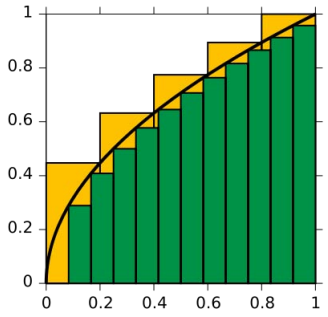

Figure 5.1. A Riemann Integral, obtained by summing the area of many vertical slices. Source: Wikipedia

The Lebesgue Integral

Instead of defining the integral as a sum over vertical slices along the x axis, we can do the following: First, construct intervals covering the range of f the y axis. Then for each interval, find the measure of all the points in x such that $f(x)$ is in that interval. Then, add to our sum that measure times the height of that interval. This procedure will exhaust all the area under a curve.

If we are in a σ-field $\mathfrak{F}$, we can define the Lebesgue integral as follows:

$$I = \int f(x) d\nu = \sum_n f_n \nu(A_n)$$

Where f_n is the height of f at vertical slice n, A_n is the set of all points in x such that $f(x) \cong f_n$, and $\nu(A_n)$ is the measure of the set A_n. These concepts will be outlined more formally in the next class.

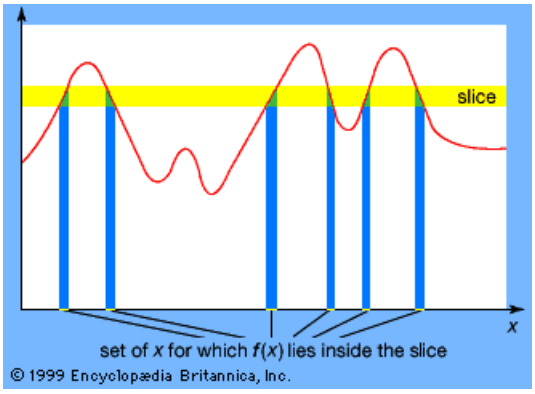

Figure 5.2. The Lebesgue integral is obtained by summing the measure of many horizontal slices. Source: Britannica Online Encyclopedia `http://www.britannica.com/EBchecked/topic/22486/analysis`