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

Lecture 6 — Jan 29, 2009

Lecturer: Nando de Freitas *Scribe: Nimalan Mahendran*

This lecture introduces parametric inference, including the Method of Moments estimator (MOME) and the Maximum Likelihood Estimator (MLE). The topics presented in these notes are discussed in more detail in chapter 9 of "All of Statistics" (Wasserman, 2004), freely available for authenticated users of the UBC network at http://www.myilibrary.com/?id=18966.

Parametric models are of the form

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$$

Parametric inference is the problem of estimating the parameter θ from the observed samples $x_1, x_2, ..., x_n$. [Was04]

Data Distribution Assumption

It is assumed that the data is distributed according to a distribution in the parametric model: $x \sim f(x|\theta_*)$.

6.1 Method of Moments Estimator (MOME)

Suppose that the parametric model has k parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$.

6.1.1 Moments and Sample Moments

The $\underline{j^{th} \text{ moment}}$ is defined as

$$\alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j P(dx|\theta)$$

where $P(dx)$, $P(x)dx$ and $dP(x)$ are equivalent.

The $\underline{j^{th} \text{ sample moment}}$ is defined as

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

where each x_i is a sample point.

6.1.2 Definition of Methods of Moments Estimator

The MOME estimator $\hat{\theta}_n$ of θ is the one that satisfies

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ &\vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

Each moment of the parametric model with the MOME estimator $\hat{\theta}_n$ will match the corresponding sample moment calculated from the sample points.

Example: MOME for the Normal Distribution

Let $x_{1:n} \sim \mathcal{N}(\mu, \sigma^2)$.
The parameters of the model are $\theta = (\mu, \sigma^2)$.
The MOME estimator $\hat{\theta}_n = (\hat{\mu}, \hat{\sigma}^2)$ matches the first and second moments.

Matching the First Moment

$$\begin{gathered} \alpha_1(\hat{\theta}_n) = \hat{\alpha}_1 \\ \Longleftrightarrow \mathbb{E}_{\hat{\theta}_n}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \Longleftrightarrow \hat{\mu} = \bar{x}_n \end{gathered}$$

Therefore, the MOME $\hat{\theta}_n = (\hat{\mu}, \hat{\sigma}^2)$ satisfies $\hat{\mu} = \bar{x}_n$.

Matching the Second Moment An expression is needed that relates $\hat{\sigma}^2$ to $\mathbb{E}_{\hat{\theta}_n}(X^2)$. Note that $\hat{\sigma}^2 = \int (x-\mu)^2 p(x)dx$, where $p(x)$ is written as shorthand for $p(x|\hat{\theta}_n)$.

$$\begin{aligned}
\int (x-\mu)^2 p(x)dx &= \int (x^2 - 2x\mu + \mu^2)p(x)dx \\
&= \int x^2 p(x)dx - 2\mu \int xp(x)dx + \mu^2 \int p(x)dx \\
&= \int x^2 p(x)dx - 2\mu^2 + \mu^2 \\
&= \int x^2 p(x)dx - \mu^2 \\
\therefore \int x^2 p(x)dx &= \int (x-\mu)^2 p(x)dx + \mu^2 \\
\therefore \mathbb{E}_{\theta_n}(X^2) &= \sigma^2 + \mu^2
\end{aligned}$$

Now $\hat{\sigma}^2$ is derived:

$$\begin{aligned}
\alpha_2(\hat{\theta}_n) &= \hat{\alpha_2} \\
\iff \mathbb{E}_{\hat{\theta}_n}(X^2) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\
\iff \hat{\sigma}^2 + \hat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\
\iff \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2 \\
\iff \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \\
\iff \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^2
\end{aligned}$$

Therefore, the MOME $\hat{\theta}_n = (\hat{\mu}, \hat{\sigma}^2)$ estimates σ^2 by $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$.

Theorem 6.1. *(Theorem 9.6 in [Was04])*
Let $\hat{\theta}_n$ be the MOME.

1. *$\hat{\theta}_n \xrightarrow{p} \theta_*$, where θ_* is the true parameter.*
2. *$\sqrt{n}(\hat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \Sigma)$, or equivalently, $(\hat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \frac{\Sigma}{n})$ where*

$$\Sigma = \left[\frac{d\alpha_1^{-1}(\theta)}{d\theta} \cdots \frac{d\alpha_k^{-1}(\theta)}{d\theta}\right] \mathbb{E}_\theta \left(\begin{bmatrix} X \\ X^2 \\ \vdots \\ X^k \end{bmatrix} [X X^2 \ldots X^k] \right) \begin{bmatrix} \frac{d}{d\theta}\alpha_1^{-1}(\theta)^T \\ \vdots \\ \frac{d}{d\theta}\alpha_k^{-1}(\theta)^T \end{bmatrix}$$

$$\frac{d\alpha_i^{-1}(\theta)}{d\theta} = \begin{bmatrix} \frac{d\alpha_i^{-1}(\theta)}{d\theta_1} \\ \vdots \\ \frac{d\alpha_i^{-1}(\theta)}{d\theta_k} \end{bmatrix}$$

6.2 Maximum Likelihood

The likelihood function is given by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- $\mathcal{L}_n(\theta) : \Theta \to [0, \infty)$ for continuous domains
- $\mathcal{L}_n(\theta) : \Theta \to [0, 1]$ for discrete domains

The maximum likelihood estimate (MLE) $\hat{\theta}_n$ is the θ that aximizes $\mathcal{L}_n(\theta)$. Equivalently, The MLE $\hat{\theta}_n$ can also be found by maximizing the log likelihood.

6.2.1 Log Likelihood

The log likelihood function is given by

$$l_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

6.2.2 Kullback-Leibler (KL) Divergence

$$\begin{aligned} KL(f,g) &= \log \frac{f(x)}{g(x)} dF(x) \\ &= \log \frac{f(x)}{g(x)} f(x) dx \end{aligned}$$

- $KL(f,f) = 0$
- $KL(f,g) \geq 0$

Kullback-Leibler (KL) divergence gives the distance between two probability distributions, although it is not a proper measure of distance because it is not symmetric (it could be the case that $KL(f,g) \neq KL(g,f)$).

KL divergence is used in the proof of MLE convergence in the next lecture.

6.2.3 Identifiability Assumption

Model $\mathcal{F}$ is identifiable if $\theta \neq \psi$ implies $KL(f_\theta, f_\psi) > 0$. If different parameter values θ and ψ correspond to different distributions f_θ and f_ψ in $\mathcal{F}$, then $\mathcal{F}$ is identifiable.

It is assumed that $\mathcal{F}$ is identifiable in the proof of MLE convergence.

Example of Unidentifiablilty

$P(x|\theta) = \lambda_1 \mathcal{N}(\mu_1, 1) + \lambda_2 \mathcal{N}(\mu_2, 1)$, $\theta = \{\lambda_1, \lambda_2, \mu_1, \mu_2\}$.

Let the true parameters for the Gaussian mixture model be $\lambda_1^* = 0.2$, $\lambda_2^* = 0.8$, $\mu_1^* = 1$, $\mu_2^* = 10$.

The following two models correspond to the same distribution:

λ_1	λ_2	μ_1	μ_2	
0.2	0.8	1	10	$= 0.2\mathcal{N}(1,1) + 0.8\mathcal{N}(10,1)$
0.8	0.2	10	1	$= 0.8\mathcal{N}(10,1) + 0.2\mathcal{N}(1,1)$

Therefore, the Gaussian mixture model is not identifiable.

Bibliography

[Was04] Larry Wasserman. *All of Statistics : A Concise Course in Statistical Inference (Springer Texts in Statistics)*. Springer, September 2004.