JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Applied and Computational Mathematics

Data Mining

625.740

A Quick Review of Statistics

Mike Weisman

email: `data.mining.625.740@gmail.com`

Parametric Models

Parametric models are models of the form

$$\mathscr{F} = \{f(x;\theta) : \theta \in \Theta\}$$

where $\Theta \subset \Re^k$ is the parameter space and $\theta = (\vartheta_1, \vartheta_2, \cdots, \vartheta_k)^T$ is the parameter.
We seek an estimate of θ.

The Method of Moments

Suppose the parameter $\theta = (\vartheta_1, \vartheta_2, \cdots, \vartheta_k)^T$ has k components. For $1 \leq j \leq k$, define the j^{th} moment

$$\alpha_j = \alpha_j(\theta) = E_\theta(X^j) = \int x^j dF_\theta(x).$$

and the j^{th} sample moment is

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

Method of Moments Estimator

Definition: The method of moments estimator $\hat{\theta}_n$ is defined to be the value of $\vec{\theta}$ such that

$$
\begin{aligned}
\alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\
\alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\
\vdots \quad & \vdots \quad \vdots \\
\alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k.
\end{aligned}
$$

The Bernoulli Distribution

Let X be a binary coin flip.

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p, \quad (p \in [0, 1]).$$

We say X has a Bernoulli distribution. The probability function is

$$f(x) = p^x (1 - p)^{1-x}, \quad \text{for} \quad x \in \{0, 1\}.$$

Example I

Let $X_1, \ldots, X_n \sim$ Bernouill(p).

$$\alpha_1 = \alpha_1(p) = E_p(X) = \sum_{x \in \{0,1\}} x\, f(x) = p.$$

$$\hat{\alpha}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \alpha_1(\hat{p}_n) = \hat{p}_n.$$

Thus, by the method of moments:

$$\alpha_1(\hat{p}_n) = \hat{\alpha}_1 \quad \Longrightarrow \quad \hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example II

Let $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then

$$\begin{aligned}
\alpha_1 &= E_\theta(X_1) = \mu, \\
\alpha_2 &= E_\theta(X_1^2) = V_\theta(X_1) + (E_\theta(X_1))^2 \\
&= \sigma^2 + \mu^2.
\end{aligned}$$

We have

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

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

The solution is

$$\hat{\mu} = \overline{X}_n \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Likelihood

Let $X_1, \ldots, X_n$ be i.i.d. with pdf $f(x;\theta)$.
Definition: The likelihood function is defined to be

$$\mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

The log-likelihood function is defined to be

$$\ell_n(\theta) = \log \mathscr{L}_n(\theta) = \log\left[\prod_{i=1}^{n} f(x_i;\theta)\right] = \sum_{i=1}^{n} \log\left[f(x_i;\theta)\right].$$

Maximum Likelihood

Definition:

The maximum likelihood estimate (MLE), $\hat{\theta}_n$ is the value of θ that maximizes $\mathscr{L}_n(\theta)$ (or equivalently $\ell_n(\theta)$).

Example III

Let $X_1, \ldots, X_n \sim$ Bernouill(p). Then $f(x) = p^x(1-p)^{1-x}, \quad \text{for} \quad x \in \{0,1\}$.

$$\mathscr{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}, \quad \text{where } S = \sum_i X_i.$$

$$\ell_n(p) = S \log p + (n-S) \log(1-p).$$

To find MLE,

$$0 = \frac{d\ell_n(p)}{dp} = \frac{S}{p} - \frac{n-S}{1-p}.$$

$$\implies \text{MLE is } \hat{p}_n = \frac{S}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Example IV

Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$.
The parameter vector we are interested in estimating is $\hat{\theta} = (\mu, \sigma)^T$.

$$\mathscr{L}_n(\mu, \sigma) = \text{const.} \cdot \prod_{i=1}^{n} \frac{1}{\sigma} \exp\{-\frac{1}{2\sigma^2}(X_i - \mu)^2\}$$

$$= \frac{\text{const.}}{\sigma^n} \exp\{-\frac{n\zeta^2}{2\sigma^2} \sum_{i=1}^{n} (\overline{X} - \mu)^2\},$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \zeta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

Example IV (continued)

$$\ell_n(\mu, \sigma) = \log(\text{const.}) - n \log \sigma - \frac{n\zeta^2}{2\sigma^2} - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}.$$

$$0 = \frac{\partial \ell_n}{\partial \mu} = \frac{n(\mu - \overline{X})}{2\sigma^2} \implies \hat{\mu} = \overline{X}.$$

$$0 = \frac{\partial \ell_n}{\partial \sigma} = \frac{n\zeta^2}{\sigma^3} - \frac{n}{\sigma} - \frac{n(\mu - \overline{X})^2}{\sigma^3} \implies \hat{\sigma} = \zeta.$$

Example V

Let $X_1, \ldots, X_n \sim$ Uniform $(0, \theta)$.

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Consider a fixed value of θ. Suppose $\exists X_i \ni \theta < X_i$.
Then $f(X_i;\theta) = 0 \implies \mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) = 0$.
Therefore, $\mathscr{L}_n(\theta) = 0$ if $\theta < X_{(n)} = \max\{X_1, \ldots, X_n\}$.
Now consider any $\theta \geq X_{(n)}$. $\forall X_i, f(X_i;\theta) = \frac{1}{\theta}$ so $\mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) = \frac{1}{\theta}$.

$$\mathscr{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & \theta \geq X_{(n)}, \\ 0\ , & \theta < X_{(n)}. \end{cases}$$

$$\mathscr{L}_n(\theta) \text{ is strictly decreasing on } [X_{(n)}, \infty) \quad \implies \quad \hat{\theta}_n = X_{(n)}.$$

Example V (continued)

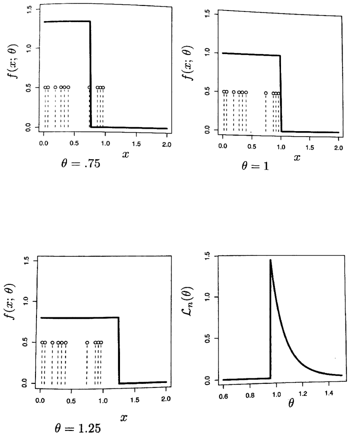

Figure: From Wasserman, **All of Statistics**.

Kullback-Leibler Divergence

Definition: If $f(x)$ and $g(x)$ are probability distributions, then the Kullback-Leibler divergence of g from f is

$$D(f,g) = \int f(x) \log\left[\frac{f(x)}{g(x)}\right] dx \quad .$$

Kullback-Leibler Divergence (continued)

Let $h(x) = f(x)/g(x)$, $\quad D(f,g) = \int g(x)h(x)\log h(x)\,dx$. Let $d\mu = g(x)\,dx$.

$$D(f,g) = \int h(x) log\, h(x) d\mu(x).$$

Now set $\varphi(t) = t\log t$. Since $0 < h(x) < \infty$,

$$\varphi\left(h(x)\right) = \varphi(1) + (h(x) - 1)\varphi'(1) + \frac{1}{2}\left(h(x) - 1\right)^2 \varphi''\left(m(x)\right),$$

where $m(x)$ lies between $h(x)$ and 1, so that $0 < m(x) < \infty$.

Kullback-Leibler Divergence (continued)

We have $\varphi(1) = 0, \varphi'(1) = 1,$ and $\int h(x) d\mu x = \int f(x) dx = 1$. So,

$$\varphi(h(x)) = \frac{1}{2} \int (h(x) - 1)^2 \varphi''(m(x)) . \quad (*)$$

$\varphi'' = \frac{1}{t} > 0$ for $t > 0$. From $(*)$,

$$\int h(x) \log h(x)\, d\mu = \frac{1}{2} \int \left(\frac{f}{g} - 1 \right)^2 \cdot \frac{g}{f}\, d\mu \geq 0$$

and equal to zero iff $h = \frac{f}{g} = 1$ a.s.

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Applied and Computational Mathematics

Data Mining

625.740

A Quick Review of Statistics