Lecture 7 (Page 1)

Estimators : Bias and Variance

Note Title 3/26/2008

Evaluating an Estimator : Bias & Variance

X is a sample from population with parameter θ
Let $d = d(X)$ be an estimator of θ

To evaluate it, measure how much it differs from θ — e.g. $(d(X) - \theta)^2$
but this depends on the sample X.

To evaluate $d(x)$, we must average over the possible samples.

$r(d,\theta) = E[(d(X) - \theta)^2]$ mean square error!

$b_\theta(d) = E[d(X)] - \theta$, bias of estimator.

If $b_\theta(d) = 0$ for all θ,
then d is an unbiased estimator of θ.

E.G. $E[m] = E\left[\frac{\sum_t x^t}{N}\right] = \frac{1}{N}\sum_t E[x^t] = \frac{N\mu}{N} = \mu.$

So the MLE m is an unbiased estimator of μ.

On a particular sample, m may not be μ. But averaged over all samples, it will be

(2)

m is also a <u>consistent</u> estimator.

Var(m) → 0 as N → ∞

$$\mathrm{Var}(m) = \mathrm{Var}\left(\frac{\sum_t x^t}{N}\right) = \frac{1}{N^2}\sum_t \mathrm{Var}(x^t) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

As the number of samples get large,

$m(x) \to \mu$

expected error $\sim \frac{1}{N}$.

<u>Now</u> s^2 is the MLE of σ^2.

$$s^2 = \sum_t \frac{(x^t - m)^2}{N} = \sum_t \frac{(x^t)^2 - Nm^2}{N}.$$

$$E[s^2] = \sum_t \frac{E[(x^t)]^2 - N.E[m^2]}{N}.$$

After some algebra $(\mathrm{Var}(x) = E(x^2) - (E[x])^2)$

$$E[s^2] = \left(\frac{N-1}{N}\right)\sigma^2 \neq \sigma^2.$$

Hence s^2 is a biased estimator of σ^2.

But for large N, the difference is negligible

<u>Asymptotically</u> <u>unbiased</u> <u>estimator</u>

(3)

Bias & Variance

Can also express the mean square error as:

$$
\begin{aligned}
r(d,\theta) &= E[(d-\theta)^2] \\
&= E[(d-E[d]+E[d]-\theta)^2] \\
&= \underbrace{E[(d-E[d])^2]}_{\text{variance of estimator}} + \underbrace{(E[d]-\theta)^2}_{\text{bias}}
\end{aligned}
$$

variance of estimator – measure how much d varies between datasets.

bias – measures how much d differs from θ.

$$r(d,\theta) = Var(d) + (b_\theta(d))^2$$

(we will discuss the bias/variance dilemma in next lecture).

(4) <u>Bayes Estimator</u>.

Sometimes we have prior knowledge about θ

— e.g. value range.

This should be used, particularly if the sample size is small.

<u>Prior density</u> $p(\theta)$ (before data)

<u>posterior density</u> $p(\theta|\mathcal{X}) = \dfrac{p(\mathcal{X}|\theta)\,p(\theta)}{p(\mathcal{X})}$

$$p(\mathcal{X}) = \int p(\mathcal{X}|\theta')\,p(\theta')\,p(\theta')$$

To estimate the density at x.

$$p(x|\mathcal{X}) = \int p(x,\theta|\mathcal{X})\,d\theta$$
$$= \int p(x|\theta,\mathcal{X})\,p(\theta|\mathcal{X})\,d\theta$$
$$= \int p(x|\theta)\,p(\theta|\mathcal{X})\,d\theta.$$

Takes an average over predictions using all values of θ, weighted by ~~their~~ probabilities.

If prediction of form $y = g(x|\theta)$ (regression)

then $y = \int g(x|\theta)\,p(\theta|\mathcal{X})\,d\theta$

(5)

Bayes Estimator (Cont)

$$y = \int g(x|\theta)\, p(\theta|\mathcal{X})\, d\theta.$$

if evaluating the integral is hard, then use a maximum a posterior (MAP) estimate.

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{X})$$

$$p(x|\mathcal{X}) = p(x|\theta_{MAP})$$

$$y_{MAP} = g(x|\theta_{MAP})$$

If the prior is flat - then we get MLE as before

Another possibility
Bayes estimator

$$\theta_{Bayes} = E[\theta|\mathcal{X}] = \int \theta\, p(\theta|\mathcal{X})\, d\theta$$

This is the best estimate if we want to minimize the expected square loss

$$E[(\theta-c)^2] = E[(\theta-\mu)^2] + (\mu-c)^2.$$

Example: if $x^t \sim N(\theta, \sigma_0^2)$, $\theta \sim N(\mu, \sigma^2)$

$$p(\mathcal{X}|\theta) = \frac{1}{(2\pi)^{N/2}\sigma_0^N} \exp\left\{-\sum_t \frac{(x^t-\theta)^2}{2\sigma_0^2}\right\}$$

$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\theta-\mu)^2/2\sigma^2}$. Then $E[\theta|\mathcal{X}] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2}\mu.$

weighted average of prior mean μ & sample mean m.