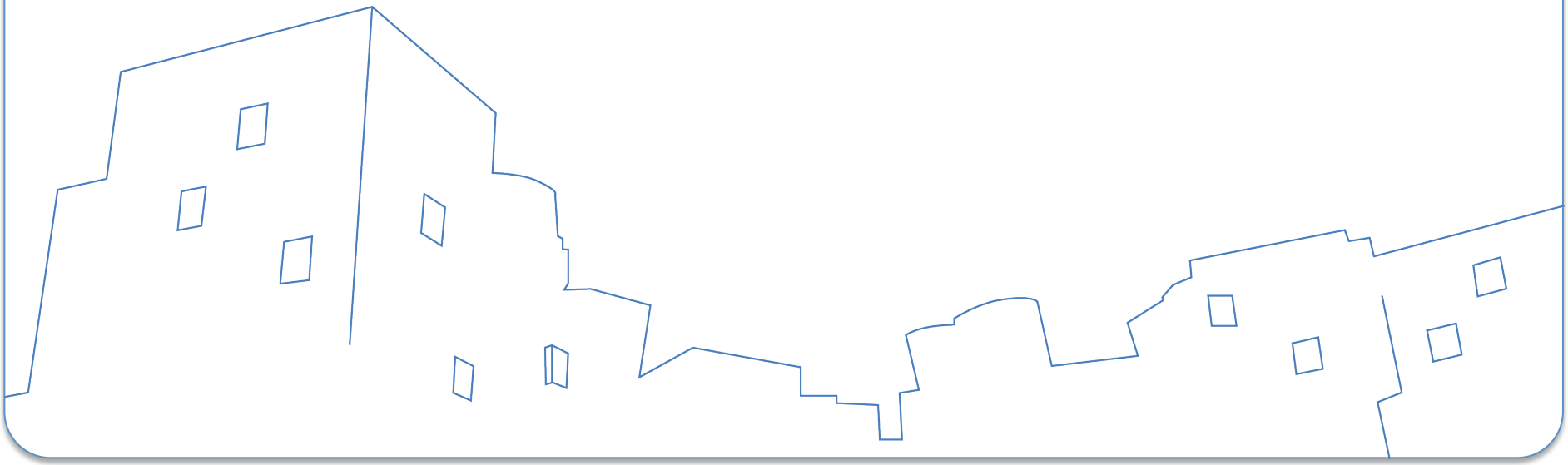
LIDS

6.434/16.391 Statistics for Engineers and Scientists

Lecture 11 10/24/2011

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology



Lecture 11 10/17/2012

BAYES ESTIMATORS

Background

- So far θ is unknown, but fixed
- When θ is a quantity whose variation can be described by a probability distribution, which is called the prior distribution
- A sample is take from a population indexed by θ. The prior distribution is updated with this sample information
- The updated prior is called the posterior distribution and the updated process is done using Bayes' rule

Bayes' rule

- Let $\pi(\theta)$ be the prior distribution, $f(\mathbf{x}|\theta)$ be the sampling distribution, and $\pi(\theta|\mathbf{x})$ be the posterior distribution, which is the conditional distribution of θ given the sample $\mathbf{x}$
- Bayes' rule is given as follows

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x},\theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

  where $m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$

- Posterior distribution is used to infer θ, which is considered as a random quantity
  - For example, mean of the posterior distribution can be used as a point estimate of θ

Example 1: Binomial Bayes estimation

- (Example 7.2.14, Casella & Berger) Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\text{Bernoulli}(p)$. Then $Y = \sum X_i$ is $\text{binomial}(n, p)$. We assume the prior distribution on p is $\text{beta}(\alpha, \beta)$. The joint distribution of Y and p is

$$\begin{aligned} f(y,p) &= f(y|p)\pi(p) \\ &= \left[\binom{n}{y} p^y (1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \end{aligned}$$

and the marginal distribution of Y is

$$f(y) = \int_0^1 f(y,p)dp = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

Example 1: Binomial Bayes estimation

- Therefore,

$$\pi(p|y) = \frac{f(y,p)}{f(y)}$$
$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

  Note that this is $\text{beta}(y+\alpha, n-y+\beta)$

- One way to estimate p is to calculate the mean of the posterior distribution,

$$\hat{p} = \mathbb{E}\{P|Y=y\} = \frac{y+\alpha}{\alpha+\beta+n}$$
$$= \underbrace{\frac{n}{\alpha+\beta+n}\frac{y}{n}}_{A} + \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta}}_{B}$$

  where A and B are contributions from sample and priori knowledge, respectively

Example 2: Normal Bayes estimators

- (Example 7.2.16, Casella & Berger) Let $X \sim \mathrm{n}\left(\theta, \sigma^2\right)$, and suppose that the prior distribution on θ is $\mathrm{n}\left(\mu, \tau^2\right)$. (Here we assume that σ^2, μ, and τ^2 are all known.) Find the posterior distribution $\pi(\theta|x)$
- Following Bayes' rule, $\pi(\theta|x)$ is also normal, with mean and variance given by

$$\mathbb{E}\left(\theta|x\right) = \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu$$

and

$$\mathbb{V}\left(\theta|x\right) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$$

Conjugate family

- Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$

Lecture 11 10/24/2011

MEAN SQUARED ERROR

Mean squared error and bias

- Definition: The mean squared error (MSE) of an estimator W of a parameter θ is the function of θ defined by $\mathbb{E}_\theta\{(W-\theta)^2\}$

- Remark:
  - Measure the average squared difference between the estimator W and parameter θ
  - Reasonable measure for the performance of point estimates
  - In general, any metric that increases with $|W-\theta|$ would be reasonable metric to measure goodness of the estimator
  - MSE is analytical tractable

Mean squared error and bias

- Note that

$$\mathbb{E}_\theta\left\{(W-\theta)^2\right\} = \mathbb{E}_\theta\left\{(W-\mathbb{E}_\theta\{W\}+\mathbb{E}_\theta\{W\}-\theta)^2\right\}$$
$$= \underbrace{\mathbb{V}_\theta\{W\}}_{\text{variability}} + \underbrace{(\mathbb{B}_\theta\{W\})^2}_{\text{bias}}$$

- MSE incorporate variability (precision) and bias (accuracy).
- To find estimator with good MSE, we need one with small variance and small bias.
- Unbiased estimator is a good choices since $\mathbb{B}_\theta\{W\}$ is already zero.

$$\mathbb{E}_\theta\left\{(W-\theta)^2\right\}\mathbb{V}_\theta\{W\}$$

- For unbiased estimator MSE is equal to its variance.

Comments on bias and MSE

- Many unbiased estimators are also reasonable from the standpoint of MSE

  However, controlling bias does not guarantee that MSEs controlled.

  A tradeoff can occur between variance and bias in such a way that a small increase in bias can result in a larger decrease in variance, resulting in an improvement in MSE

- We will illustrate this in the following

Example 1

- (Example 7.3.3, Casella & Berger, Normal MSE)
- Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}\left(\mu, \sigma^2\right)$. The statistics $\bar{X}$ and S^2 are both unbiased estimators since

$$\mathbb{E}\{\bar{X}\} = \mu, \;\; \mathbb{E}\{S^2\} = \sigma^2, \;\; \text{for all } \mu \text{ and } \sigma^2$$

  (Recall that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$)

- The MSEs of these estimators are given by

$$\mathbb{E}\left(\bar{X} - \mu\right)^2 = \mathbb{V}\{\bar{X}\} = \frac{\sigma^2}{n}$$

$$\mathbb{V}\left(S^2 - \sigma^2\right)^2 = \mathbb{V}\{S^2\} = \frac{2\sigma^4}{n-1}$$

  – Without the normality assumption, the MSE of $\bar{X}$ remains the same but the expression for the MSE of S^2 will change

Example 2

- (Example 7.3.4, Casella & Berger, Normal MSE) Continuation of previous example
- An alternative estimator for σ^2 is the maximum likelihood estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$. It is straightforward to calculate

$$\mathbb{E}\{\hat{\sigma}^2\} = \mathbb{E}\left\{\frac{n-1}{n}S^2\right\} = \frac{n-1}{n}\sigma^2$$

so $\hat{\sigma}^2$ is a biased estimator of σ^2. The variance of $\hat{\sigma}^2$ can also be calculated as

$$\mathbb{V}\{\hat{\sigma}^2\} = \mathbb{V}\left\{\frac{n-1}{n}S^2\right\} = \left(\frac{n-1}{n}\right)^2 \mathbb{V}\{S^2\} = \frac{2(n-1)\sigma^4}{n^2}$$

Example 2

- Hence, the MSE of $\hat{\sigma}^2$ is given by

$$\mathbb{E}\left\{\left(\hat{\sigma}^2-\sigma^2\right)^2\right\}=\frac{2(n-1)\sigma^4}{n^2}+\left(\frac{n-1}{n}\sigma^2-\sigma^2\right)^2=\left(\frac{2n-1}{n^2}\right)\sigma^4$$

We thus have

$$\mathbb{E}\left\{\left(\hat{\sigma}^2-\sigma^2\right)^2\right\}=\left(\frac{2n-1}{n^2}\right)\sigma^4<\left(\frac{2}{n-1}\right)\sigma^4=\mathbb{E}\left\{\left(S^2-\sigma^2\right)^2\right\}$$

This result shows that $\hat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved

Example 2

- Remark:
  - Does not imply that S^2 should be discarded
  - It says that on the average S^2 is close to σ^2 in MS sense
  - However $\hat{\sigma}^2$ on the average under estimate σ^2

- In general MSE
  - penalizes equally for over and under estimation
  - is reasonable for location estimate

    scale estimation (lower bound is 0) is not symmetric so

Lecture 11 10/24/2011

BEST UNBIASED ESTIMATOR

Motivation

- There is no one "best MSE" estimator, since the class of all estimators is too large
- We can restrict the class of estimators, for example, all unbiased estimators, i.e., $\mathcal{C}_\theta \{W : \mathbb{E}\{W\} = \theta\}$
- If W_1 and W_2 are both unbiased estimators of parameter θ, i.e.,

$$\mathbb{E}\{W_1\} = \mathbb{E}\{W_2\} = \theta$$

  When comparing W_1 and W_2, we should choose the one with smaller variance

Motivation

- Consider $\mathcal{C}_\tau = \{W : \mathbb{E}\{W\} = \tau(\theta)\}$. If $W_1, W_2 \in \mathcal{C}_\tau$, then

$$\mathbb{E}\{W_1\} = \mathbb{E}\{W_2\} = \tau \text{ and } \mathbb{B}\{W_1\} = \mathbb{B}\{W_2\}$$

$$\begin{aligned}\mathbb{E}\left\{(W_1-\theta)^2\right\} - \mathbb{E}\left\{(W_2-\theta)^2\right\} &= \mathbb{V}\{W_1\} + (\mathbb{B}\{W_1\})^2 \\ &\quad - \mathbb{V}\{W_2\} - (\mathbb{B}\{W_2\})^2 \\ &= \mathbb{V}\{W_1\} - \mathbb{V}\{W_2\}\end{aligned}$$

  i.e., the comparison of MSE within the class $\mathcal{C}_\tau$ can be based on variance alone

UMVUE

- Definition 7.3.7: An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $\mathbb{E}_\theta\{W^*\} = \tau(\theta)$ for all θ *and* for any other estimator W with $\mathbb{E}_\theta\{W\} = \tau(\theta)$, we have $\mathbb{V}_\theta\{W^*\} \leq \mathbb{V}_\theta\{W\}$ for all θ. W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$