Probabilities / Statistics (10701 Recitation 2)

Part II: Estimation

Bernoulli (θ)

$$(1, 0, 1, 1, 1, 0, 1, 1) = S$$

$$P(S;\theta) = \theta^6 (1-\theta)^2 = \theta \cdot (1-\theta) \cdot \theta\,\theta\,\theta\,(1-\theta)\,\theta\,\theta$$

$$\ell\ell(\theta) = \log P(S;\theta) = 6 \log \theta + 2 \log (1-\theta)$$

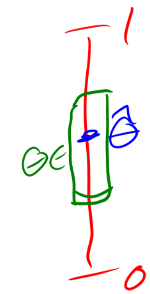
$$\frac{\partial}{\partial \theta} \ell\ell(\theta) = \frac{6}{\theta} - \frac{2}{1-\theta} = 0$$

$$6(1-\theta) = 2\theta$$

$$\frac{6}{8} = \hat{\theta}_{MLE}$$

$$P(|\hat{\theta} - \theta^*| \ge \varepsilon) \le 2e^{-2N\varepsilon^2} = \delta$$

$$P\left(|\hat{\theta} - \theta^*| \ge \sqrt{\frac{\ln\frac{2}{\delta}}{2N}}\right) \le \delta$$

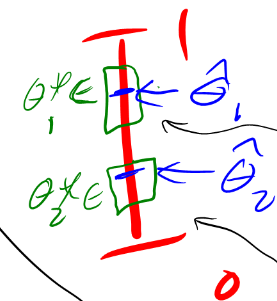


given 2 coins, with biases θ_1^* & θ_2^*

Proto-learning theory:

say $\theta_1^* \neq \theta_2^*$

which one is smaller?



If we pick N large enough, these confidence intervals will be small enough to tell which one is better.

so we are pretty sure $\theta_1^* > \theta_2^*$

Multinomial $(1, \theta_1, \theta_2, \theta_3, \ldots, \theta_K)$

$$\forall i \in 1,2,\ldots,K \quad \theta_i \in [0,1]$$

$$\sum_{i=1}^{K} \theta_i = 1$$

Die:

$p(\text{lands w/ 1 face up}) = \theta_1$

$p(\text{lands w/ } i \text{ face up}) = \theta_i$ $\leftarrow$

Toss die N times

3 6 1 3 6 6 5 6 4 2 2 6 4 5 5 6 2 6 4 2

$N_1 = \#$ w/ 1 face up

$\vdots$

$N_i = \#$ w/ i face up

$$\sum_{i=1}^{K} N_i = N$$

$$\prod_{i=1}^{K} \theta_i^{N_i}$$

$$(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K)$$

$$\ell\ell(\vec{\theta}) = \sum_{i=1}^{K} N_i \log \theta_i$$

$$ll(\vec{\theta}) = \sum_{i=1}^{K} N_i \log \theta_i$$

$$\frac{\partial}{\partial \theta_i} ll(\vec{\theta}) = \frac{N_i}{\theta_i} = 0$$

:(didn't work!

forgot about → $$\sum_{i=1}^{K} \theta_i = 1$$

Lagrange to the rescue!

$$\frac{\partial}{\partial \theta_i}\left(ll(\vec{\theta}) - \lambda\left(\sum_{i=1}^{K} \theta_i - 1\right)\right) = \frac{N_i}{\theta_i} - \lambda = 0$$

$$\hat{\theta}_i = \frac{N_i}{\lambda}$$

$$\sum_{i=1}^{K} \hat{\theta}_i = 1$$

$$\sum_{i=1}^{K} \frac{N_i}{\lambda} = \frac{N}{\lambda} = 1$$

$$\lambda = N$$

$$\hat{\theta}_i = \frac{N_i}{N}$$

So, for the sequence above where $N=20$,
$\hat{\theta}_1 = \frac{1}{20}$, $\hat{\theta}_2 = \frac{1}{5}$, $\hat{\theta}_3 = \frac{1}{10}$, $\hat{\theta}_4 = \frac{3}{20}$, $\hat{\theta}_5 = \frac{3}{20}$, $\hat{\theta}_6 = \frac{7}{20}$

Bayesian Inference

(The prior makes a bigger difference when S is small)

Given
- likelihood $P(S \mid \theta)$ (small S)
- Prior $P(\theta)$

Find Posterior
$$P(\theta \mid S) = \frac{P(S \mid \theta) P(\theta)}{P(S)}$$

Conjugacy

$$\frac{P(S \mid \theta) f(\theta; \alpha)}{P(S)} = f(\theta; \tilde{\alpha})$$

posterior has same form as prior just different parameter value.

Beta / Bernoulli

$\alpha_H = \#$ heads
$\alpha_T = \#$ tails

$$P(S \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

$$f(\theta; \beta_H, \beta_T) = \underbrace{\frac{\Gamma(\beta_H + \beta_T)}{\Gamma(\beta_H)\Gamma(\beta_T)}}_{\frac{1}{B(\beta_H, \beta_T)}} \theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}$$

$$P(S \mid \theta) f(\theta; \beta_H, \beta_T) \propto \frac{1}{B(\alpha_H + \beta_H, \alpha_T + \beta_T)} \theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\beta_T + \alpha_T - 1}$$

$P(\theta \mid S)$

Mean $E[\theta \mid S]$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

$$\hat{\theta}_{Mean} = \frac{\alpha_H + \beta_H}{\alpha_H + \beta_H + \alpha_T + \beta_T}$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

derived earlier

$$P(S \mid \theta)\underbrace{P(\theta)}$$

$(1, 0, 1, 1, 1, 0, 1, 1)$

$\alpha_H = 6$ $\quad \beta_H = 10$

$\alpha_T = 2$ $\quad \beta_T = 10$

(also see the demo on the website)

$\frac{3}{4}$ $\quad \frac{4}{7}$ $\quad \frac{16}{28}$ $\quad \frac{21}{28}$

$\hat{\theta}_{Mean}$ $\quad \hat{\theta}_{MLE}$

from first page.

Gaussian

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$\in \mathbb{R}$

the Central Limit Theorem explains why the Gaussian comes up so often.

CLT

$Y_1 + Y_2 + Y_3 + \cdots$

$$S = (x_1, x_2, \ldots, x_N)$$

$$\frac{\partial}{\partial\mu} \log P(S|\mu,\sigma^2)$$

$$= \frac{\partial}{\partial\mu}\left(\sum_{i=1}^{N} \cancel{\log\frac{1}{\sqrt{2\pi}\sigma}} - \frac{(x_i-\mu)^2}{2\sigma^2}\right) = 0$$

$$\sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i = \hat{\mu}_{MLE}$$