Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2016 : Lecture 13

Avinash Sharma

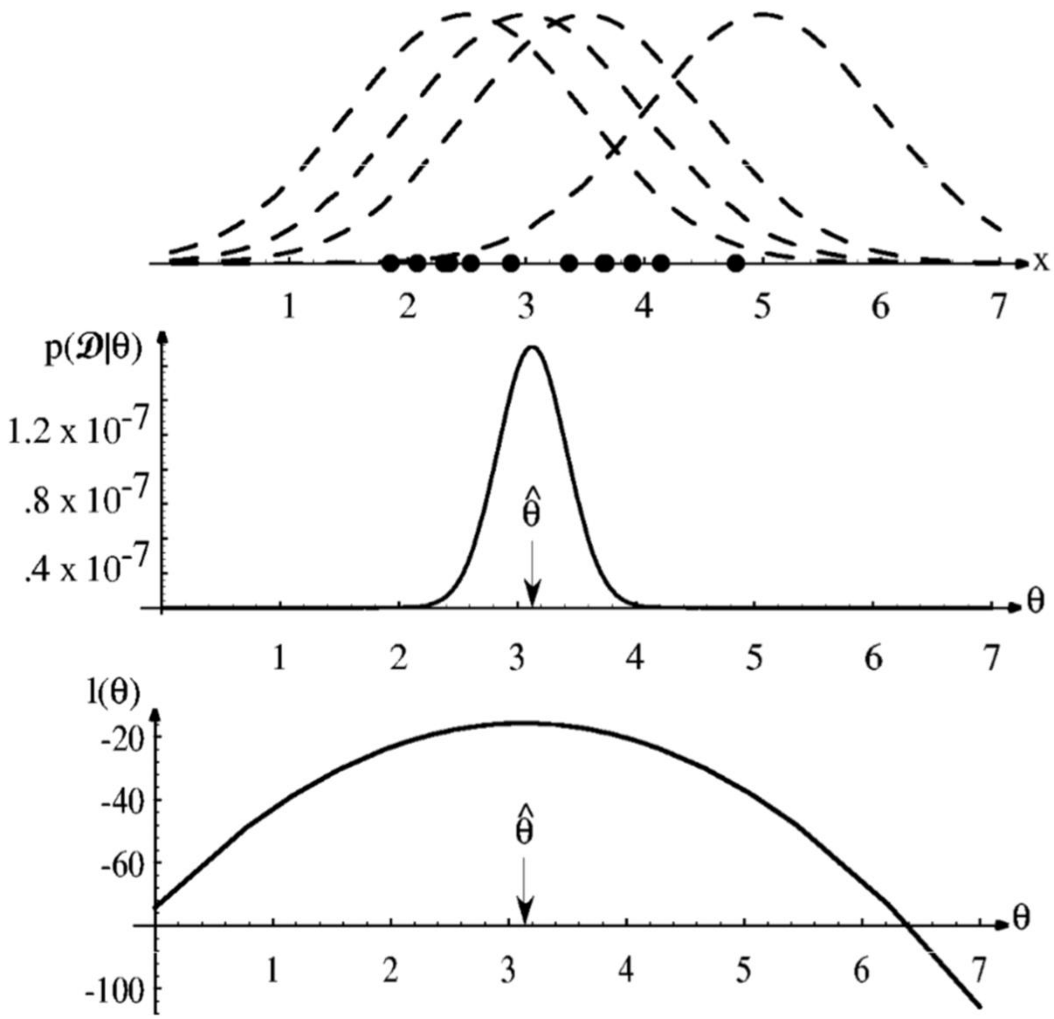
CVIT, IIIT Hyderabad

Lecture Plan

- Revision from Previous Lecture
- Parameter Estimation
  - Maximum Likelihood Estimation (MLE)
  - Bayesian Parameter Estimation (BPE)
    - Univariate Gaussian Case
    - General Theory
- MLE v/s BPE
- Problems of Dimensionality
- Component Analysis (3.8 in the next class)

Maximum Likelihood Estimation

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$



Maximum Likelihood Estimation

$$l(\boldsymbol{\theta}) \equiv \ln\, p(\mathcal{D}|\boldsymbol{\theta})$$

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}),$$

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}}\, \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}).$$

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0}.$$

Maximum Likelihood Estimation

The Gaussian Case: Unknown μ and $\mathbf{\Sigma} = \sigma^2$ (Univariate)

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0,$$

$$\boxed{\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{k=1}^{n} x_k \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2. \end{aligned}}$$

Bayesian Parameter Estimation (BPE)

- Class Posterior probabilities can be estimated by 1) assuming certain functional form for class conditional probabilities and 2) using the samples for parameter estimation.

$$P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D}_j)P(\omega_j)}$$

- The class conditional densities can further be written as:

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

- If $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply around some $\hat{\boldsymbol{\theta}}$ then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}})$$

BPE: Univariate Gaussian Case

- Let, $p(x|\mu) \sim N(\mu, \sigma^2)$, and $p(\mu) \sim N(\mu_0, \sigma_0^2)$.

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\ d\mu} \\
&= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu) \\
&= \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)} \\
&= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right] \\
&= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]
\end{aligned}
$$

BPE: Univariate Gaussian Case

- Let, $p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$
- This yields: $\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma_2}\bar{x}_n + \frac{\mu_0}{\sigma_0^2}, \qquad \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$
- Or, $\boxed{\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}.}$

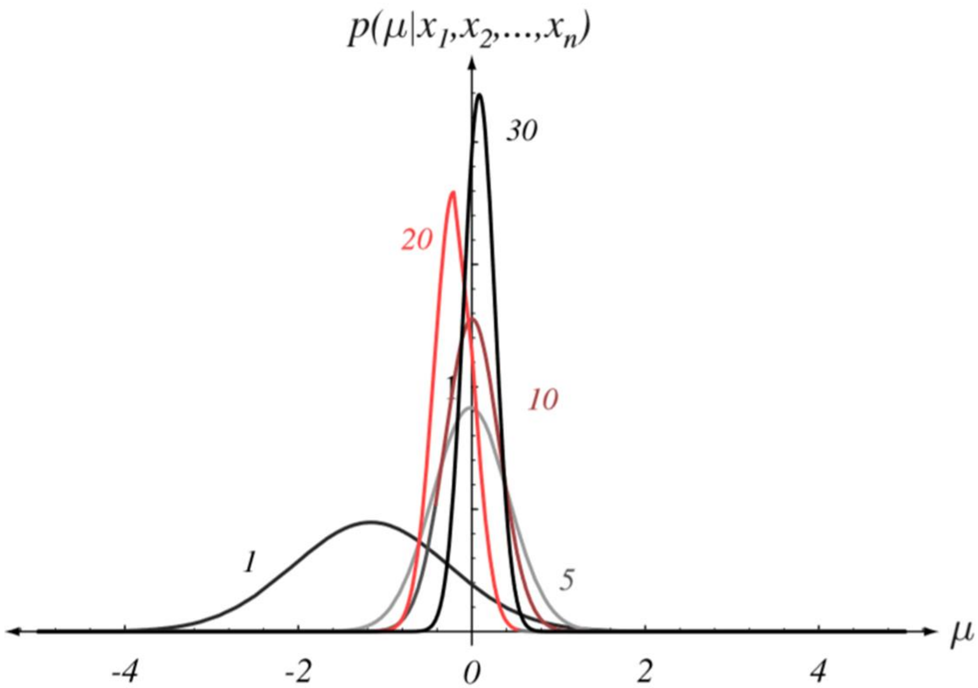
❖ ${\sigma^2}_n$ decreases monotonically as n increases to a large value, establishing the notion of ***Bayesian Learning***.

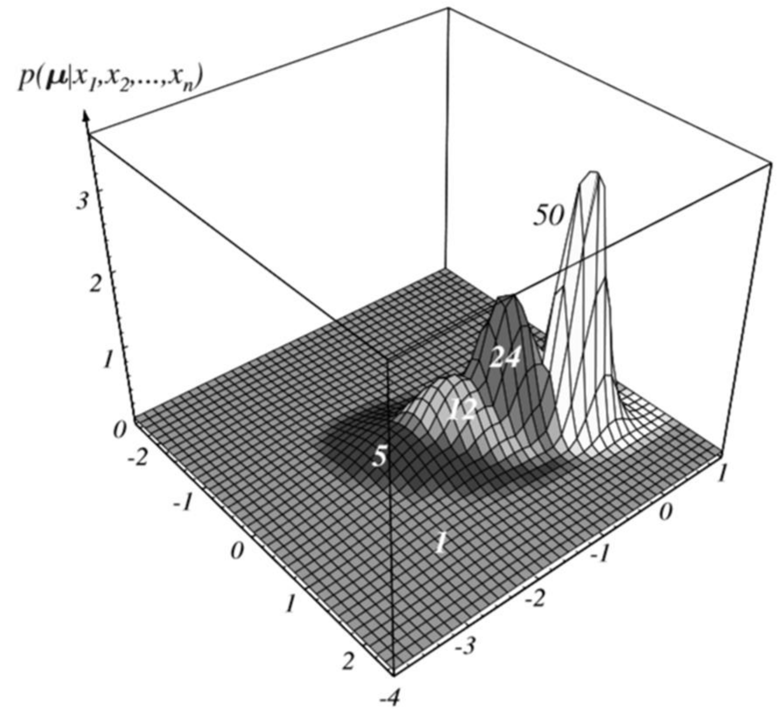
❖ As n increases to a large value, μ_n converge to sample mean.

❖ If $\sigma_0 = 0$, then $\mu_n = \mu_0$.

❖ If $\sigma_0 \gg \sigma$ then μ_n is only sample mean.

BPE: Univariate Gaussian Case





BPE: Univariate Gaussian Case

$$\begin{aligned} p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\, d\mu \\ &= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\ &= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n), \end{aligned}$$

$$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu.$$

$$p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2+\sigma_n^2).$$

- Conditional mean μ_n is treated as true mean of the class conditional density.
- Additional uncertainty in x due to lack of our knowledge about exact μ is modelled by increase in variance.

BPE: General Theory

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}},$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}).$$

- Recursive Bayes Learning

$$p(\mathcal{D}^n|\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta}).$$

$$p(\boldsymbol{\theta}|\mathcal{D}^n) = \frac{p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{n-1})}{\int p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{n-1})\ d\boldsymbol{\theta}}$$

MLE v/s BPE

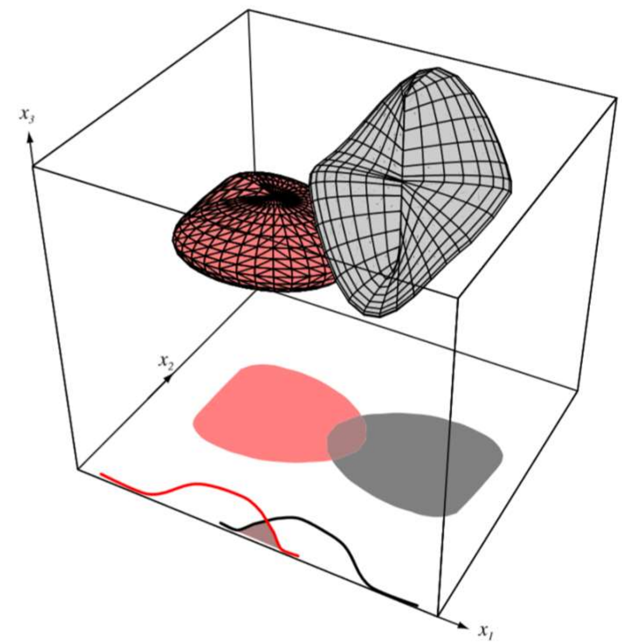
- Computational Complexity
  - MLE is preferred over BPE as it can employ simple iterative minimization techniques
- Interpretation
  - MLE returns single value as compare to weighted average of models(parameters) returned by BPE
- Prior Information
  - Bayes formulation incorporate more information than MLE if the prior information is at all reliable

Error in Bayesian Classification

- Bayes or Indistinguishability Error
  - Due to overlapping class conditional densities
- Model Error
  - Due to disparity in true v/s assumed distribution
- Estimation Error
  - Due to small data size

Problems of Dimensionality

- Dimensions
  - More the Merrier (?)
    - Wrong Model Choice
    - Small Training Sample Size



- Computational Complexity
  - Order of operations

- Overfitting