

Maheesan Niranjana

School of Electronics and Computer Science
University of Southampton



Autumn Semester 2016/17

Estimation

- We have data \mathbf{x}_k
- We have a model: e.g. the data came from a Gaussian density
- We have parameters relating to the model: e.g. mean of the Gaussian
- Our task is to estimate the parameters given the data
- Frequentist thought
 - The given data is a particular realization of the underlying system
 - Repeated experiments will give different estimates
 - If each experiment uses a lot of data, the variation may be small
 - We define a probabilistic model and maximize likelihood
 - Bias and Variance in estimation
- Bayesian thought
 - We are interested in the uncertainty in parameters
 - We have a prior uncertainty
 - There is some information in the data
 - We combine these to get a posterior uncertainty

Likelihood & Log likelihood

- $p(\mathbf{x} | \omega_j)$
- Parametric form $p(\mathbf{x} | \omega_j, \boldsymbol{\theta}_j)$

For example

$$p(\mathbf{x} | \omega_j, \boldsymbol{\theta}_j) = \mathcal{N}(\mathbf{m}_j, \mathbf{C}_j)$$

- Dataset $\mathcal{D} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ of identical and independently distributed samples (iid)
 - All samples were drawn from this distribution
 - Independent draws (previous value does not affect the next draw)
- Likelihood of an item of data (function of the parameter!)

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

- Likelihood of the set of data

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

- Log likelihood

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

Maximum Likelihood

- Maximum likelihood

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

- Maximize by taking derivative

$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_n} \end{bmatrix}$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^n \nabla_{\boldsymbol{\theta}} p(\mathcal{D} | \boldsymbol{\theta})$$

... and equating to zero

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

... and solve for the unknown parameter values.

Example: Multivariate Gaussian $\mathcal{N}(\mathbf{m}, C)$

Mean unknown, (Covariance known)

- ... product of Gaussians; taking log removes exp and turns \prod into \sum
- ... write it out for a single data point

$$\ln p(\mathbf{x}_k | \mathbf{m}) = \frac{1}{2} \ln(2\pi)^d \det C - \frac{1}{2} (\mathbf{x}_k - \mathbf{m})^T C^{-1} (\mathbf{x}_k - \mathbf{m})$$

- ... the derivative

$$\nabla_{\mathbf{m}} \ln p(\mathbf{x}_k | \mathbf{m}) = C^{-1} (\mathbf{x}_k - \mathbf{m})$$

- Given n data $\mathbf{x}_1, \dots, \mathbf{x}_n$

$$\sum_{k=1}^n C^{-1} (\mathbf{x}_k - \hat{\mathbf{m}}) = \mathbf{0}$$

- and the solution is...

$$\hat{\mathbf{m}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$$

Example: Univariate Gaussian $\mathcal{N}(m, \sigma^2)$

(Mean and variance unknown)

- Two parameters: $\theta_1 = m$ and $\theta_2 = \sigma^2$
- Log likelihood of a single data:

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

- Derivative of the log likelihood

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

- Given n data x_1, x_2, \dots, x_n , and considering the full log likelihood

$$\begin{aligned} \sum_{k=1}^n \frac{1}{\theta_2} (x_k - \hat{\theta}_1) &= 0 \\ -\sum_{k=1}^n \frac{1}{\theta_2} + \sum_{k=1}^n \frac{1}{\theta_2^2} (x_k - \hat{\theta}_1)^2 &= 0 \end{aligned}$$

Example

Univariate Gaussian, unknown mean and variance (cont'd)

- ... after some algebra

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

- If we did this estimation from several datasets...

$$E \left[\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \right] = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

expected value of estimate is not the same as the true value!

For the multivariate Gaussian, we give the results (slides of **L1**):

$$\begin{aligned}\hat{\mathbf{m}} &= \\ \hat{\mathbf{C}} &= \end{aligned}$$

Bayesian Estimation

Illustrate the idea through univariate Gaussian, only mean unknown

- Data: $\mathcal{D} : x_1, \dots, x_n$
- Likelihood (as seen before) $p(x|m) \sim \mathcal{N}(m, \sigma^2)$
- Prior uncertainty over parameters: $p(m) \sim \mathcal{N}(m_0, \sigma_0^2)$
 m_0 and σ_0^2 are known.
- Posterior via Bayes' formula

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\int p(\mathcal{D}|m)p(m) dm}$$

Denominator is a constant, so we deal with

$$p(m|\mathcal{D}) = \alpha \prod_{k=1}^n p(x_k|m) p(m)$$

- Two ways forward from here
 - Maximum *a posteriori* estimation
 - Inference by integrating out parameters

Bayesian Estimation: Univariate Gaussian

- Data: $\mathcal{D} : x_1, \dots, x_n$
- Likelihood (as seen before) $p(x|m) \sim \mathcal{N}(m, \sigma^2)$
- Prior uncertainty over parameters: $p(m) \sim \mathcal{N}(m_0, \sigma_0^2)$
 m_0 and σ_0^2 are known.
- Substituting gives the posterior as a product of Gaussians

$$p(m|\mathcal{D}) = \alpha \prod_{k=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_k - m}{\sigma} \right)^2 \right] \times \frac{1}{\sigma_0 \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_k - m_0}{\sigma_0} \right)^2 \right]$$

- Which can be reduced to...

$$p(m|\mathcal{D}) = \alpha_2 \exp \left\{ -\frac{1}{2} \left\{ \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) m^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{m_0}{\sigma_0^2} \right) m \right\} \right\}$$

- But then...

$$p(m|\mathcal{D}) = \frac{1}{\sigma_n} \exp \left\{ -\frac{1}{2} \left(\frac{m - m_n}{\sigma_n} \right)^2 \right\}$$

- Matching terms...

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\frac{m_n}{\sigma_n^2} = \frac{n}{\sigma^2} \hat{m}_n + \frac{m_0}{\sigma_0^2}, \quad \text{where, } \hat{m}_n = \frac{1}{n} \sum_{k=1}^n x_k.$$

Bayesian Estimation: Univariate Gaussian (cont'd)

- Finally...

$$m_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{m}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right) m_0$$
$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

- We now have an estimate that combines *prior* information about the parameter ($p(m) = m_0$) with data (x_1, \dots, x_k) to quantify uncertainty about the parameter.
 - Before seeing any data, we have a belief
 - As we see more and more data, our belief is taken over by what the data tells us.