COMP6229 Machine Learning

Week 5: Introduction to Estimation

Mahesan Niranjan

School of Electronics and Computer Science
University of Southampton

Autumn Semester 2017/18

Estimation

- We have data $\boldsymbol{x}_k$
- 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

- $p\left(\boldsymbol{x} \mid \omega_j\right)$
- Parametric form $p\left(\boldsymbol{x} \mid \omega_j, \boldsymbol{\theta}_j\right)$
  For example
$$p\left(\boldsymbol{x} \mid \omega_j, \boldsymbol{\theta}_j\right) = \mathcal{N}\left(\boldsymbol{m}_j, \boldsymbol{C}_j\right)$$
- Dataset $\mathcal{D} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ 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!)
$$p\left(\boldsymbol{x}_k \middle| \boldsymbol{\theta}\right)$$
- Likelihood of the set of data (independent draws)
$$p\left(\mathcal{D} \middle| \boldsymbol{\theta}\right) = \prod_{i=1}^{n} p\left(\boldsymbol{x}_k \middle| \boldsymbol{\theta}\right)$$
- Log likelihood
$$l\left(\boldsymbol{\theta}\right) = \ln\, p\left(\mathcal{D} \middle| \boldsymbol{\theta}\right)$$

Maximum Likelihood

- Maximum likelihood
$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \, l\left(\boldsymbol{\theta}\right)$$
- Maximize by taking derivative
$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \vdots \\ \frac{\partial}{\partial\theta_n} \end{bmatrix}$$
$$\boldsymbol{\nabla}_{\boldsymbol{\theta}}\, l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} p\left(\mathcal{D} \middle| \boldsymbol{\theta}\right)$$
  ... and equating to zero
$$\boldsymbol{\nabla}_{\boldsymbol{\theta}}\, l = \boldsymbol{0}.$$
  ... and solve for the unknown parameter values.

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

Mean unknown, (Covariance known)

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

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

- ... the derivative

$$\boldsymbol{\nabla}_{\boldsymbol{m}} \ln p(\boldsymbol{x}_k|\boldsymbol{m}) = C^{-1}(\boldsymbol{x}_k - \boldsymbol{m})$$

- Given n data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, the derivative we equate to zero is sum over all data:

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

- and the solution is...

$$\widehat{\boldsymbol{m}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$$

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

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

- Given n data $x_1, x_2, ..., x_n$, and considering the full log likelihood

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

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

Example

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

- ... after some algebra

$$\widehat{m} = \frac{1}{n}\sum_{k=1}^{n} x_k$$
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \widehat{m})^2$$

- 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**):

$$\widehat{\boldsymbol{m}} =$$
$$\widehat{C} =$$

Bayesian Estimation

Illustrate the idea through univariate Gaussian, only mean unknown

- Data: $\mathcal{D}: \; x_1, \ldots x_n$
- Likelihood (as seen before): $p(x|m) \sim \mathcal{N}(m, \sigma^2)$
- Prior uncertainty over parameters (*i.e.* mean): $p(m) \sim \mathcal{N}(m_0, \sigma_0^2)$
  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

(Only the mean is unknown)

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

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

- 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...

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

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

Bayesian Estimation: Univariate Gaussian (cont'd)

- Finally...

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

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

- We now have an estimate that combines *prior* information about the parameter $(p(m) = m_0)$ with data $(x_1, ..., x_k)$ 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.