MSPR 4: Appendix: Maximum Likelihood Parameter Estimation for Gaussians

Dr. Hendrik Purwins

AAU CPH

September 27, 2015

This lecture is based on presentations made in my course 'Advanced Topics in in Music Technology' at Music Technology Group, Universitat Pompeu Fabra between 2007-2011, Barcelona, especially the ones by Stefan Kersten and Srikanth Cherla, closely following

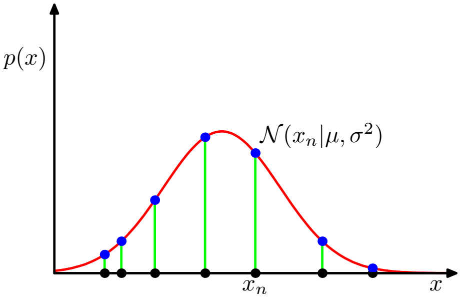
- *Andrew Moore's machine learning tutorial lectures: Gaussians, Gaussian Mixture Models,* *http://www.autonlab.org/tutorials/*
- *Christopher Bishop: Pattern Recognition and Machine Learning: Chapter 1 (Introduction) 1.2.3 (The Gaussian Distribution) p. 24 - 27, 2.3 (The Gaussian Distribution) p. 78, 84 bottom - 85 top.*

Outline

1 Gaussian Parameter Estimation
- ML

- Let's consider a vector $\mathbf{x}$ of N samples from a random distribution
- Elements of the vector are drawn independently and are identically distributed (i.i.d)
- Then the probability of $\mathbf{x}$ being produced by a Gaussian with parameters μ and σ^2 is

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$



Maximum Likelihood Estimation

- Given data $\mathbf{x}$, we want to find the most probable μ_{MAP} and σ^2_{MAP} that have generated $\mathbf{x}$: (*maximum a posteriori (MAP) estimation*)

$$(\hat{\mu}_{MAP}, \hat{\sigma}^2_{MAP}) = \arg\max_{\mu,\sigma^2} p(\mu, \sigma^2 \mid \mathbf{x})$$

- Bayes' Theorem: $p(\mu, \sigma^2 \mid \mathbf{x}) = \frac{p(\mathbf{x}|\mu,\sigma^2)}{p(\mathbf{x})} \cdot p(\mu, \sigma^2)$
- Marginalization $p(\mathbf{x}) = \sum_{\theta_1,\theta_2} p(\mathbf{x}|\theta_1, \theta_2) \cdot p(\theta_1, \theta_2)$ across all (θ_1, θ_2) does not dep. on $(\mu, \sigma^2) \Rightarrow p(\mathbf{x})$ const.
- Assume no prior knowledge about $(\mu, \sigma^2) \Rightarrow p(\mu, \sigma^2)$ const:

$$\arg\max_{\mu,\sigma^2} p(\mathbf{x} \mid \mu, \sigma^2) = \arg\max_{\mu,\sigma^2} p(\mu, \sigma^2 \mid \mathbf{x})$$

- *Maximum Likelihood Estimation*:

$$\boxed{(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML}) = \max_{\mu,\sigma^2}\{p(\mathbf{x} \mid \mu, \sigma^2)\}}$$

- The joint probability of (two) independent variables factorizes into the product of each marginal probability:

$$p(X, Y) = p(X)p(Y)$$

- Assume data points are *independent and identically distributed (i.id.)*
- $\Rightarrow$ Likelihood function of the Gaussian distribution:

$$\begin{aligned} p(\mathbf{x} \mid \mu, \sigma^2) &= \prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2) \\ &= N(x_1 \mid \mu, \sigma^2) \cdot N(x_2 \mid \mu, \sigma^2) \ldots \cdot N(x_N \mid \mu, \sigma^2) \end{aligned}$$

- ML adjusts mean μ and variance σ^2 to Gaussian distribution
- It is preferred to maximise the *log-likelihood* $\ln p(\mathbf{x}|\mu, \sigma^2)$, because
  - The natural logarithm is a monotonically increasing function. So maximizing log of a function is equivalent to maximizing function itself.
  - $\prod$ turns into $\sum$ and makes math simpler (Remember: $\ln(x \cdot y) = \ln x + \ln y$)
  - It gets rid of the exponentials.
  - A sum of logarithms is less likely to underflow a machine representation than a product of small probabilities
- Find parameter values maximizing the likelihood function, in two stages:
  1. Maximize with respect to mean
  2. Maximize with respect to variance

- Logarithm of the likelihood function:

$$\ln(p(\mathbf{x} \mid \mu, \sigma^2)) = \ln(\prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2))$$

$$= \ln(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{-\frac{(x_n-\mu)^2}{2\sigma^2}})$$

$$\ln\,(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi \cdot \sigma^2}}) + \ln\,(\prod_{n=1}^{N} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}})$$

  - First term:

$$\ln\,((\frac{1}{\sqrt{2\pi \cdot \sigma^2}})^N) = \ln(1^N) - ln(2\pi \cdot \sigma^2)^{\frac{N}{2}} = -\frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2)$$

  - Second term:

$$\sum_{n=1}^{N} -\frac{(x_n-\mu)^2}{2\sigma^2} = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n-\mu)^2$$

- Log likelihood function:

$$\ln(p(\mathbf{x} \mid \mu, \sigma^2)) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2) \qquad (1)$$

Estimation of the Mean

- Partial derivative with respect to μ:

$$\frac{\partial \ln(p(\mathbf{x} \mid \mu, \sigma^2))}{\partial \mu} = \frac{\partial}{\partial \mu}\{(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2)\})$$

$$= -\frac{1}{2\sigma^2} \cdot \frac{\partial}{\partial \mu}(\sum_{n=1}^{N}(x_n - \mu)^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- Find maximum by setting derivative to 0:

$$0 = \frac{\partial \ln(p(\mathbf{x} \mid \mu, \sigma^2))}{\partial \mu} = \sum_{n=1}^{N}(x_n - \mu) = \sum_{n=1}^{N} x_n - N \cdot \mu$$

- Solution for μ:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Estimation of the Variance

- Maximize for variance (for which $\theta = (\mu, \sigma^2)$ is most likely):
- Derive for σ^2 (mean is known):

$$\frac{\partial \ln(p(\mathbf{x} \mid \mu, \sigma^2))}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\{(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2)\})$$
$$= \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^2}$$

- First term (derivation rule: a $\frac{d}{dy}\frac{1}{y} = -a\frac{1}{y^2}$):

$$-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\frac{\partial}{\partial\sigma^2}\frac{1}{\sigma^2} = \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\frac{1}{\sigma^4}$$

- Second term (derivation rule: $\frac{d}{dy}\ln y = \frac{1}{x}$):

$$-\frac{N}{2}\frac{\partial}{\partial\sigma^2}\ln(\sigma^2) - = -\frac{N}{2}\cdot\frac{1}{\sigma^2} = -\frac{N}{2\sigma^2} \tag{2}$$

Estimation of the Variance

$$\frac{\partial \ln(p(\mathbf{x} \mid \mu, \sigma^2))}{\partial \sigma^2} = 0$$

$$\Leftrightarrow \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2\sigma^2} = 0 \qquad \backslash \cdot 2\sigma^2$$

$$\Leftrightarrow \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - N = 0 \qquad \backslash + N$$

$$\Leftrightarrow \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 = N \qquad \Leftrightarrow \sigma^2_{ML} = \tfrac{1}{N} \textstyle\sum_{n=1}^{N} (x_n - \mu)^2$$

- Solving $\ln p(\mathbf{x}|\mu, \sigma^2)$ yields maximum likelihood solutions for mean and variance

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$$

- Same as our simple estimators for mean and variance above!