THE UNIVERSITY OF TENNESSEE KNOXVILLE

AICIP RESEARCH

COSC 522 – Machine Learning

Lecture 4 – Parametric Estimation (MLE)

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu

Course Website: http://web.eecs.utk.edu/~hqi/cosc522/

Questions

- How to estimate the parameters of a pdf? Take the example of a multivariate Gaussian.
- What is maximum likelihood estimation (MLE)?
- What is the derivative of the quadratic form?
- What's the cost function when estimating the parameters of the pdf?

Multivariate Normal Density

AICIP RESEARCH

$$p(\vec{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right]$$

$\vec{x}$: d - component column vector

$\vec{\mu}$: d - component mean vector

Σ : d - by - d covariance matrix

$|\Sigma|$: determinant

Σ^{-1} : inverse

Estimating Normal Densities

AICIP RESEARCH

- Calculate μ, Σ

$$\vec{\mu}_i = \begin{bmatrix} \mu_{i1} = \dfrac{1}{n_i}\sum_{k=1}^{n_i} x_{k1} \\ \vdots \\ \mu_{id} = \dfrac{1}{n_i}\sum_{k=1}^{n_i} x_{kd} \end{bmatrix}$$

$$\Sigma_i = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix} = \frac{1}{n_i - 1}\sum_{k=1}^{n_i} \left(\vec{x}_k - \vec{\mu}_i\right)\left(\vec{x}_k - \vec{\mu}_i\right)^T$$

Method 1 – Maximum Likelihood Estimation

$D = \{x_1, x_2, \cdots, x_k, \cdots, x_n\}$ is a data set of n training samples

Compare "likelihood"	
$p(x \mid \omega_i)$	$p(D \mid \vec{\theta})$

$$\vec{\theta} = \begin{bmatrix} \vec{\mu} \\ \Sigma \end{bmatrix}$$

$$p(D \mid \vec{\theta}) \xrightarrow{\text{assume samples are drawn independently}} \prod_{k=1}^{n} p(x_k \mid \vec{\theta})$$

$$l(\vec{\theta}) = \ln p(D \mid \vec{\theta}) = \sum_{k=1}^{n} \ln p(x_k \mid \vec{\theta})$$

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

Derivation

$$\theta = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} = \begin{bmatrix}\vec{\mu} \\ \Sigma\end{bmatrix}$$

$$p(\vec{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right]$$

$$p(\vec{x}_k \mid \vec{\theta}) = \frac{1}{(2\pi)^{d/2}|\theta_2|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}_k-\theta_1)^T \frac{1}{\theta_2} (\vec{x}_k-\theta_1)\right]$$

$$l(\vec{\theta}) = \sum_{k=1}^{n} \left(-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\theta_2| - \frac{1}{2}(\vec{x}_k-\theta_1)^T \frac{1}{\theta_2}(\vec{x}_k-\theta_1)\right)$$

$$\frac{\partial l}{\partial \theta_1} = 0 \qquad \frac{\partial l}{\partial \theta_2} = 0$$

Derivative of a Quadratic Form

A matrix A is "positive definite" if $x^T Ax > 0 \quad \forall x \in R^d, x \neq 0$

$x^T Ax$ is also called a "quadratic form".

The derivative of a quadratic form is particularly useful:

$$\frac{d}{dx}\left(x^T Ax\right) = \left(A + A^T\right)x$$

AICIP RESEARCH

Method 2 – Baysian Estimation

- Maximum likelihood estimation
  - The parameters are fixed
  - Find value for θ that best agrees with or supports the actually observed training samples – likelihood of θ w.r.t. the set of samples

$$p\left(D \mid \vec{\theta}\right)$$

- Baysian estimation
  - Treat parameters as random variable themselves

AICIP RESEARCH

* The pdf of the Parameter (μ) is Gaussian

$$p(\mu \mid D) = \frac{p(D \mid \mu)p(\mu)}{C} = \frac{1}{C}\prod_{k=1}^{n} p(x_k \mid \mu)p(\mu)$$

$$p(x_k \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right]$$

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]$$

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right]$$

AICIP RESEARCH

* Derivation

$$p(\mu \mid D) = \frac{1}{C} \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]$$

$$= \alpha \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{x_k - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha \exp\left[-\frac{1}{2}\left(\frac{\sum_{k=1}^{n} x_k^2 - 2\mu \sum_{k=1}^{n} x_k + n\mu^2}{\sigma^2} + \frac{\mu^2 - 2\mu_0\mu + \mu_0^2}{\sigma_0^2}\right)\right]$$

$$= \beta \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right]$$

AICIP RESEARCH

* μ_n and σ_n

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) + \left(\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\right)\mu_0$$

Our best guess for μ after observing *n* samples

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

Measures our uncertainty about this guess

◆ Behavior of Bayesian learning

- The larger the *n*, the smaller the σ_n – each additional observation decreases our uncertainty about the true value of μ
- As *n* approaches infinity, $p(\mu|D)$ becomes more and more sharply peaked, approaching a Dirac delta function.
- μ_n is a linear combination between the sample mean and μ_0