Lecture Outline

Reading: Section 2.3

- Gaussian Distribution
- ML Estimation for the Gaussian Distributional Parameters
- MAP Estimation for the Gaussian Distributional Parameters

Introduction

Density Estimation Problem: Given continuous-valued data, $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, find the distributional parameters θ of $p(\mathbf{x})$. With $p(\mathbf{x})$, we can make a prediction and to quantify the uncertainty of the prediction.

2.3 The Gaussian Distribution (Review)

Note: Sections 2.3.1-2.3.3 regarding manipulations of Gaussian distributions, conditional and marginal Gaussian distributions, and Bayes' theorem for Gaussian distributions will be left to the student to read and study.

In the case of a scalar RV x, the Gaussian distribution (pdf) is defined as

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1}$$

where μ is the mean and σ^2 is the variance. For a D-dimensional random vector $\mathbf{x}$ the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2}$$

where $\boldsymbol{\mu}$ is the D-dimensional mean vector, $\boldsymbol{\Sigma}$ is the $D \times D$ covariance matrix, and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

Other Details of the Gaussian Distribution

Geometry Students should review this section p. 79-85, as it develops some important relations for multivariate Gaussian distributions which will serve us in subsequent work.

Conditional Gaussian Distributions Students should review this section 2.3.1, as it develops some important relations for multivariate Gaussian distributions which will serve us in subsequent work.

Marginal Gaussian Distributions Students should review this section 2.3.2, as it develops some important relations for multivariate Gaussian distributions which will serve us in subsequent work.

Bayes' Theorem for Gaussian Variables Students should review this section 2.3.3, as it develops some important relations for multivariate Gaussian distributions which will serve us in subsequent work.

©2016 Phillip De Leon

2.3.4 ML Estimation of Distributional Parameters for Multivariate Gaussian

(This was already done for a Gaussian scalar, random variable. We now extend to Gaussian random vectors)

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be independently drawn from a multivariate Gaussian distribution, we can estimate the parameters, $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ of the distribution by maximizing the log-likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{3}$$

In order to maximize the log-likelihood with respect to $\boldsymbol{\mu}$, we set the derivative of the log likelihood function to zero. The derivative of the log likelihood function with respect to $\boldsymbol{\mu}$ is given by[1]

$$\frac{\partial}{\partial \boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \tag{4}$$

which, when set to zero, leads to the ML estimate of the mean

$$\begin{aligned} 0 &= \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \\ &= \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}) \\ &= \sum_{n=1}^{N}\mathbf{x}_n - N\boldsymbol{\mu} \end{aligned} \tag{5}$$

or

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{6}$$

which is the sample mean. Maximizing the log-likelihood with respect to $\boldsymbol{\Sigma}$ (more involved) leads to the ML estimate of the covariance matrix

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}. \tag{7}$$

Evaluating the expectation of the ML estimates we obtain

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu} \tag{8}$$

and

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}. \tag{9}$$

We see that the expectation of the ML estimate for the mean is equal to the true mean. However, the ML estimate for the covariance is biased. The bias can be corrected by defining a different estimator, i.e. unbiased estimator

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{\mathrm{ML}} &= \frac{N}{N-1}\boldsymbol{\Sigma}_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}. \end{aligned} \tag{10}$$

[1] $\frac{\partial}{\partial \mathbf{x}}\mathbf{x}^T\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}$

2.3.6 Bayesian Inference for the Gaussian (MAP Estimation of Distributional Parameters for the Univariate Gaussian)

We now develop a Bayesian treatment by introducing prior distributions over μ and Σ. To simplify, we assume a scalar random variable, x modeled as a Gaussian random variable, $\mathcal{N}(\mu, \sigma^2)$, and a data set $\mathbf{x}$. We want to estimate μ and σ^2 by maximizing the posterior

$$p(\mu, \sigma^2|\mathbf{x}) = \frac{p(\mathbf{x}|\mu, \sigma^2)p(\mu, \sigma^2)}{p(\mathbf{x})}. \tag{11}$$

To make the problem tractible, we will solve (11) assuming we know σ^2 and then solve it assuming we know μ. Then we will solve assuming neither.

Assume σ^2 is known, MAP estimate μ

The likelihood function viewed as a function of μ is given by

$$\begin{aligned} p(\mathbf{x}|\mu) &= \prod_{n=1}^{N} p(x_n|\mu) \\ &= \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}. \end{aligned} \tag{12}$$

We see that the likelihood function takes the form of the exponential of a quadratic in μ. Thus, if we choose a prior $p(\mu)$ given by a Gaussian, it will be a conjugate distribution for this likelihood function and hence will also be a Gaussian. We therefore take our prior distribution to be

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{13}$$

where μ_0, σ_0^2 are the mean, variance of the prior (the subscript, $_0$ denotes prior and is equivalent to no observations). With some manipulation, we can show that

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2) \tag{14}$$

(the subscript, $_N$ denotes posterior and is equivalent to N observations) where the MAP estimate for μ is

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}} \tag{15}$$

μ_{ML} is the ML solution for μ, i.e. sample mean in this case and

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}. \tag{16}$$

We note that the mean of the posterior distribution (15) is a compromise between the prior mean μ_0 ($N = 0$) and the maximum likelihood solution μ_{ML} ($N = \infty$).

The variance of the posterior distribution is expressed as a precision. Precisions are additive so that the precision of the posterior is given by the precision of the prior plus one contribution of the data precision from each of the observed data points.

As we increase the number of observed data points, the precision steadily increases, corresponding to a posterior distribution with a steadily decreasing variance. With no observed data points, we have the prior variance, whereas if $N \to \infty$, the variance σ_N^2 goes to zero and the posterior distribution becomes infinitely peaked around the ML solution.

Recap: For a dataset $\mathbf{x}$ modeled as $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, we want to MAP estimate μ. The posterior is $p(\mu, \sigma^2|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$. One estimate of μ would be $\mathbb{E}[\mu|\mathbf{x}] = \mu_N$ from (14).

Assume μ is known, MAP estimate σ^2

Let $\lambda = 1/\sigma^2$. The likelihood function viewed as a function of σ^2 is given by

$$\begin{aligned} p(\mathbf{x}|\lambda) &= \prod_{n=1}^{N} p(x_n|\lambda) \\ &= \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \\ &\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}. \end{aligned} \tag{17}$$

The corresponding conjugate prior should therefore be proportional to the product of a power of λ and the exponential of a linear function of λ. This corresponds to a *gamma* distribution

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{18}$$

where $a,\ b$ are the distribution parameters. The mean and variance of the gamma distribution are given by

$$\mathbb{E}[\lambda] = \frac{a}{b} \tag{19}$$

and

$$\mathrm{var}[\lambda] = \frac{a}{b^2}. \tag{20}$$

Therefore our prior is

$$p(\lambda) = \mathrm{Gam}(\lambda|a_0, b_0) \tag{21}$$

where a_0, b_0 are the distribution parameters.

We obtain the posterior distribution as

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \tag{22}$$

which we recognize as a gamma distribution of the form $\mathrm{Gam}(\lambda|a_N, b_N)$ where

$$a_N = a_0 + N/2 \tag{23}$$

and

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 - b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}} \tag{24}$$

We note that there is no need to keep track of normalization constants in the prior and likelihood function because, if required, the correct coefficient can be found at the end using (18) with (23) and (24).

Assume neither μ or σ^2 is known

See discussion starting with 3rd paragraph on p. 101.

MAP Estimation for the multivariate Gaussian

See discussion starting with 1st paragraph on p. 102.