

COM6509/4509 — Tutorial Sheet 1

Bayesian and Maximum Likelihood Manipulation of Gaussian Models

Neil Lawrence

October 5th, 2012

1. Univariate Gaussian model. A Gaussian density governs a vector of univariate observations, $\mathbf{t} = \{t_i\}_{i=1}^N$. The associated error function has the following form.

$$E(\mu) = \sum_{i=1}^N (t_i - \mu)^2$$

- (a) Introduce the variance parameter, σ^2 and convert the error function to the Gaussian density. Find the maximum likelihood solutions for both μ and σ^2 .
- (b) Place the following Gaussian prior over the mean,

$$p(\mu) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{1}{2\alpha}\mu^2\right)$$

and compute the marginal likelihood for \mathbf{t} and the posterior density for μ .

2. Maximum likelihood in a multivariate Gaussian. A data set consists of p dimensional vectors, $\mathbf{t}_{i,:}$, from a matrix $\mathbf{T} = \{\mathbf{t}_{i,:}\}_{i=1}^N$ (i.e. $\mathbf{T} \in \mathbb{R}^{N \times p}$). The likelihood is given by

$$p(\mathbf{T}) = \prod_{i=1}^N p(\mathbf{t}_{i,:})$$

where the likelihood of each data point is

$$p(\mathbf{t}_{i,:}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{C}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{t}_{i,:} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{t}_{i,:} - \boldsymbol{\mu})\right).$$

- (a) Write down the log likelihood and use the following matrix and vector derivatives

$$\begin{aligned} \frac{d\mathbf{x}^\top \mathbf{A} \mathbf{x}}{d\mathbf{x}} &= \mathbf{A} \mathbf{x} + \mathbf{A}^\top \mathbf{x} \\ \frac{d \log |\mathbf{C}|}{d\mathbf{C}} &= \mathbf{C}^{-1} \\ \frac{d\mathbf{a}^\top \mathbf{C}^{-1} \mathbf{a}}{d\mathbf{C}} &= -\mathbf{C}^{-1} \mathbf{a} \mathbf{a}^\top \mathbf{C}^{-1} \end{aligned}$$

to show that the maximum likelihood solutions for the mean, $\hat{\boldsymbol{\mu}}$ and covariance matrix, $\hat{\mathbf{C}}$, are

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^N \mathbf{t}_{i,:},$$

$$\hat{\mathbf{C}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{t}_{i,:} - \hat{\boldsymbol{\mu}})(\mathbf{t}_{i,:} - \hat{\boldsymbol{\mu}})^\top.$$

(b) Now consider an independent Gaussian prior over the elements of the mean vector,

$$p(\boldsymbol{\mu}) = \prod_{i=1}^p \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{1}{2\alpha}\mu_i^2\right)$$

i. Show that this can be written in vector form as follows:

$$p(\boldsymbol{\mu}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\boldsymbol{\mu}^\top \boldsymbol{\mu}\right).$$

ii. Now compute the posterior density for $\boldsymbol{\mu}$, $p(\boldsymbol{\mu}|\mathbf{T})$. Write down the terms that remain that would be required for the marginal likelihood of \mathbf{T} , $p(\mathbf{T})$ (note given the matrix algebra we've covered you won't be able to write down the full form of the marginal likelihood).

3. **Regression with a basis function model.** Assume that we wish to perform a nonlinear regression by computing a set of basis functions, for example,

$$\phi_j(\mathbf{x}_{i,:}) = \exp\left(-\frac{1}{2\ell_j^2}(x_i - \mu_j)^2\right),$$

where μ is a location parameter and ℓ is a width parameter for the j th basis function. For each data point we take the m basis functions and write them in a vector of the following form

$$\boldsymbol{\phi}_{i,:} = [\phi_1(\mathbf{x}_{i,:}) \dots \phi_m(\mathbf{x}_{i,:})]^\top$$

and the complete set of basis functions is written in a matrix, $\boldsymbol{\Phi} \in \mathbb{R}^{N \times m}$ of the following form,

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_{1,:} \boldsymbol{\phi}_{2,:} \dots \boldsymbol{\phi}_{N,:}]^\top.$$

If we assume Gaussian noise we can write down the Gaussian likelihood of a single data point, i ,

$$p(t_i | \boldsymbol{\phi}_{i,:}, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t_i - \mathbf{w}^\top \boldsymbol{\phi}_{i,:})^2\right).$$

(a) Assume the noise is independent and identically distributed and write down the corresponding likelihood and log likelihood of the entire data set.

(b) Show that the maximum likelihood solution for \mathbf{w} is given by

$$\hat{\mathbf{w}} = \left(\Phi^\top \Phi \right)^{-1} \Phi^\top \mathbf{t}.$$

(c) Consider a Gaussian prior over the parameters, \mathbf{w} ,

$$p(\mathbf{w}) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{1}{2\alpha} w_i^2\right).$$

Show that the posterior for \mathbf{w} is given by a Gaussian with covariance

$$\mathbf{C}_w = \left(\frac{1}{\sigma^2} \Phi^\top \Phi + \alpha^{-1} \mathbf{I} \right)^{-1}$$

and mean

$$\boldsymbol{\mu}_w = \frac{1}{\sigma^2} \mathbf{C}_w \Phi^\top \mathbf{t}$$

- i. Compare the solution for the maximum likelihood and the posterior mean over \mathbf{w} . When do they become the same?
- ii. What problems occur for the maximum likelihood solution if $m > N$?

(d) Show that the marginal likelihood of the data set is given by

$$p(\mathbf{t}|\mathbf{X}, \alpha, \sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \mathbf{t}^\top \mathbf{K}^{-1} \mathbf{t}\right)$$

where

$$\mathbf{K} = \alpha \Phi \Phi^\top + \sigma^2 \mathbf{I}$$

by using the matrix inversion formula:

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{DA}^{-1} \mathbf{B})^{-1} \mathbf{DA}^{-1}.$$