COM6509/4509 — Tutorial Sheet: Bayesian and Maximum Likelihood Manipulation of Gaussian Models

Neil Lawrence

November 9, 2015

1. Univariate Gaussian model. A Gaussian density governs a vector of univariate observations, $\mathbf{y} = \{y_i\}_{i=1}^n$. The associated objective function has the following form.

$$E(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2$$

   (a) Introduce the variance parameter, σ^2 and convert the objective 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 $\mathbf{y}$ and the posterior density for μ.

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

$$p(\mathbf{Y}) = \prod_{i=1}^{n} p(\mathbf{y}_{i,:})$$

   where the likelihood of each data point is

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

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

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

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

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

(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 $\boldsymbol{\mu}$, $p(\boldsymbol{\mu}|\mathbf{Y})$. Write down the terms that remain that would be required for the marginal likelihood of $\mathbf{Y}$, $p(\mathbf{Y})$ (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 jth basis function. For each data point we take the m basis functions and write them in a vector of the following form

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

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

$$\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_{1,:}\boldsymbol{\phi}_{2,:}\ldots\boldsymbol{\phi}_{n,:}\right]^\top.$$

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

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

(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

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

(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

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

and mean

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

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

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

where

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

by using the matrix inversion formula:

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

4. Consider a p-dimensional data set $\mathbf{Y} = [\mathbf{y}_{:,1} \ldots \mathbf{y}_{:,p}] \in \Re^{n \times p}$ containing n data points. Let's assume the data is inherently low dimensional and each element of this matrix can be represented by a corresponding latent variable, $\mathbf{x}_{i,:}$ and an associated vector of parameters, $\mathbf{w}_{j,:}$ which maps from the latent space to the data space. If there is Gaussian noise then this implies a likelihood of the form,

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

Now consider an independent Gaussian prior over the latent variables,

$$p(\mathbf{x}_{i,:}) = \prod_{j=1}^{q} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_{i,j}^2\right),$$

where q is the number of latent variables.

(a) Make independent and identically distributed assumptions for the elements of the data matrix $\mathbf{Y}$ and formulate the likelihood for the ith data point.

(b) Combine the prior density for the corresponding ith latent variable to show that the marginal likelihood of the ith data point is

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

where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$ and $\mathbf{W}$ is the matrix containing all the vectors that map from the latent variables to the data space, $\mathbf{W} = [\mathbf{w}_{1,:}\ldots\mathbf{w}_{p,:}]^\top \in \Re^{p\times q}$.