COMP3206: Exercises on Matrix calculus for optimisation

Srinandan Dasmahapatra

2017/8

1 Learning the parameters of a Gaussian by Maximum Likelihood Estimation (MLE)

Estimation: Data $\mathcal{X} = \{x^1, \ldots, x^n, \ldots x^N\}$ Find mean vector μ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Assuming data are drawn i.i.d. from Gaussian $\mathcal{N}(x|\mu, \boldsymbol{\Lambda})$, $\boldsymbol{\Lambda} \triangleq \boldsymbol{\Sigma}^{-1}$, the log likelihood $\mathcal{L}(\mu, \boldsymbol{\Lambda})$ is

$$\mathcal{L}(\mu, \boldsymbol{\Lambda}) = -\sum_{n=1}^{N} \frac{1}{2}(x^n - \mu)^T \boldsymbol{\Lambda}(x^n - \mu) + \frac{N}{2} \log \det(\boldsymbol{\Lambda}) + \text{constant}.$$

The exercises below will enable you to find the optimal mean and covariance matrix using maximum likelihood estimation.

- Optimal μ: $\frac{\partial}{\partial \mu}\mathcal{L}(\mu, \boldsymbol{\Lambda}) = 0 = \sum_n \boldsymbol{\Lambda}\,(x^n - \mu)$

$$\sum_n \boldsymbol{\Lambda} x^n = \sum_n \boldsymbol{\Lambda}\mu = N\boldsymbol{\Lambda}\mu \Rightarrow \mu = \frac{1}{N}\sum_n x^n.$$

- Optimal $\boldsymbol{\Lambda}$: $\frac{\partial}{\partial \boldsymbol{\Lambda}}\mathcal{L}(\mu, \boldsymbol{\Lambda}) = 0$.

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1} = \frac{1}{N}\sum_{n=1}^{N}(x^n - \mu)(x^n - \mu)^T.$$

1.1 Exercises

1. Verify that the log likelihood function for

$$p(x^n|\mu, \boldsymbol{\Sigma}) = (\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|})^{-1} \exp\left(-\frac{1}{2}(x^n - \mu)^T \boldsymbol{\Sigma}^{-1}(x^n - \mu)\right)$$

is as shown above.

2. For $p \times p$ matrices $\mathbf{A}, \mathbf{B}$ with matrix elements $(\mathbf{A})_{ij} = a_{ij}$ and $(\mathbf{B})_{ij} = b_{ij}$, show that tr(**AB**)=tr(**BA**) by writing $\text{tr}(\mathbf{A}) = \sum_i a_{ii})$ and the product of matrices as
$$(\mathbf{AB})_{ij} = \sum_k a_{ik} b_{kj}.$$

3. For $n \times p$ matrix $\mathbf{A}$ with matrix elements $(\mathbf{A})_{ij} = a_{ij}$, show that the sum of the squares of the matrix elements
$$\sum_{ij} a_{ij}^2 = \text{tr}(\mathbf{A}\mathbf{A}^T), \text{ where tr is the matrix trace.}$$

4. For the Kronecker delta δ_{ij} defined as
$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise.} \end{cases}$$
show that
   (i) $\sum_j a_{ij} \delta_{kj} = a_{ik}$,
   (ii) The diagonal elements of the product of matrices $\mathbf{A}, \mathbf{B}$ is $\sum_{jk} a_{ij} b_{jk} \delta_{ki}$,
   (iii) Trace tr(**A**) = $\sum_{ij} a_{ij} \delta_{ij}$,
   (iv) Make sure you grok $\dfrac{\partial}{\partial x_i} x_j = \delta_{ij}$.

5. For $p \times p$ matrix $\mathbf{A}$ with matrix elements $(\mathbf{A})_{ij} = a_{ij}$ $1 \leqslant i, j \leqslant p$ and vector $\mathbf{x} = (x_1, \dots, x_p)^T$ the i-th element of vector $(\mathbf{Ax})$ is $(\mathbf{Ax})_i = \sum_{j=1}^p a_{ij} x_j$. Show that $\frac{\partial}{\partial \mathbf{x}}(\mathbf{Ax}) = \mathbf{A}^T$ by writing out the indices explicitly:
$$\left(\frac{\partial}{\partial \mathbf{x}} (\mathbf{Ax}) \right)_{ij} = \frac{\partial}{\partial x_i} (\mathbf{Ax})_j = \frac{\partial}{\partial x_i} \sum_{k=1}^p a_{jk} x_k.$$

6. Show, by writing out the matrix elements as above, that the gradient of the scalar quadratic form $\mathbf{xAx}$ is $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{Ax} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$. *Hint:* the i-th matrix element of the gradient is
$$\frac{\partial}{\partial x_i} \left(\sum_{r,s=1}^p x_r a_{rs} x_s \right).$$
This should lead to the expression for the MLE of the mean.

7. The partial derivative of the quadratic form $\mathbf{x}\mathbf{A}\mathbf{x}$ with respect to $\mathbf{A}$ can be evaluated for each matrix element a_{ij}, $1 \leqslant i, j \leqslant p$:

$$\frac{\partial}{\partial a_{ij}}\left(\sum_{r,s=1}^{p} x_r a_{rs} x_s\right).$$

Show that the result is $\mathbf{x}\mathbf{x}^\top$ (a $p \times p$ matrix).

8. Remember that the determinant of a matrix can be written as a sum

$$\det(\mathbf{A}) = \sum_{j=1}^{p} a_{ij}\operatorname{cof}(a_{ij})$$

where cof(a_{ij}) is $(-1)^{i+j}$ times the determinant of the submatrix of $\mathbf{A}$ obtained by deleting the i-th row and j-th column. In particular, cof(a_{ij}) does not contain a_{ij}. Show that

$$\left(\frac{\partial}{\partial \mathbf{A}}\ln\det\mathbf{A}\right)_{ij} = \frac{\partial}{\partial a_{ij}}\ln\left(\sum_{s=1}^{p} a_{rs}\operatorname{cof}(a_{rs})\right) = \left(\mathbf{A}^{-1}\right)_{ij}.$$

This and the previous problem should help you derive the MLE of the covariance matrix.

2 Regularised linear regression

In regression problems, we have been minimising the residual sum of errors (RSS) with respect to the parameters θ that are the weight vectors $\boldsymbol{w}$ in

$$f(\mathbf{x}; \boldsymbol{w}) = w_0 + \sum_{i=1}^{p} w_i \phi_i(\mathbf{x}).$$

If we introduce $x_0 = 1$, we can write, for each data point $(\mathbf{x}^n, y^n) = (1, x_1^n, \ldots, x_p^n, y^n)$, and the RSS is

$$\text{RSS} = \sum_{n=1}^{N} (r^n)^2 = \sum_{n=1}^{N} \left(y^n - f(\mathbf{x}^n; \boldsymbol{w})\right)^2.$$

When we introduce a L_2 regularisation term $\|\boldsymbol{w}\|_2 = \boldsymbol{w}^\top\boldsymbol{w}$ for the weights $\boldsymbol{w}$, the minimisation is then over a loss function $\ell(\boldsymbol{w})$:

$$\ell(\boldsymbol{w}) = -\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} \left(y^n - f(\mathbf{x}^n; \boldsymbol{w})\right)^2 + \lambda\boldsymbol{w}^\top\boldsymbol{w},$$

where λ controls the trade-off between where the data wants the learnt functions to go and how small the modeller wants to keep $\|w\|_2$. Minimising $\ell(w)$ is equivalent to maximising $\mathcal{L}(w)$. This L_2 regularised version of linear regression is called **ridge regression**.

2.1 Exercises

1. Take the gradient of the loss function $\ell(w)$ with respect to the weight vector w and set it equal to zero. For each vector component w_i compute

$$\frac{\partial}{\partial w_i} \sum_{n=1}^{N} \left(y^n - \sum_{j=0}^{p} w_j \phi_j(x^n) \right) \left(y^n - \sum_{k=0}^{p} w_k \phi_k(x^n) \right) + \lambda \sum_{j,k=0}^{p} w_j w_k \delta_{jk}.$$

   Show that the derivative reduces to -2 multiplied by

$$\sum_{n=1}^{N} \left\{ \phi_i(x^n) y^n - \sum_{j=0}^{p} (w_j + \lambda \delta_{ij}) \phi_j(x^n) \phi_i(x^n) \right\},$$

   a quantity that we will set to zero for max/minimisation.

2. Keep in mind that the data index n in the superscript is a row index while the i, j, k indices for weights stand for columns. Introduce the matrix $\mathbf{\Phi}$ with matrix elements $(\mathbf{\Phi})_{nj} = \phi_j(x^n)$. Also, the column vector of y values is $\mathbf{y}$. Use this to rewrite the above as a matrix equation

$$\mathbf{\Phi}^{\mathsf{T}} \mathbf{y} = (\mathbf{\Phi}^{\mathsf{T}} \mathbf{\Phi} + \lambda \mathbb{I}) w \Rightarrow w = (\mathbf{\Phi}^{\mathsf{T}} \mathbf{\Phi} + \lambda \mathbb{I})^{-1} \mathbf{\Phi}^{\mathsf{T}} \mathbf{y}.$$

3. Identify the negative of the loss function $\ell(w)$ with the quantity you worked with for the maximum likelihood estimation problem for the Gaussian. Think about the correspondence. The $\lambda\|w\|_2$ term becomes a prior distribution on weights in a Bayesian interpretation.