Statistical Machine Learning 2018

Exercises and answers, week 3

21 September 2018

TUTORIAL

Exercise 1

We consider the Gaussian distribution in one dimension (see Bishop, p. 27-28)

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

with parameters μ and $\sigma^2 > 0$. Now suppose we have a data set of observations χ

$$\chi = \{x_1, \ldots, x_N\}$$

The observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown. The probability of the data set χ, given these unknown parameters is

$$p(\chi|\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

1. Show that the log likelihood function can be written in the form

$$\ln p(\chi|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) \tag{2}$$

ANSWER: Make use of the fact that $\ln\prod_i \ldots = \sum_i \ln\ldots$, and in particular $\ln(ab) = \ln(a) + \ln(b)$, furthermore, $\ln(1/a) = -\ln(a)$, and $\ln(\sqrt{a}) = 1/2\ln(a)$. In addition $\ln\exp(a) = a$, and finally $\sum_{i=1}^{N} c = Nc$ if c does not depend on i. Then it should more or less easily follow:

$$\begin{aligned}
\ln p(\chi|\mu,\sigma^2) &= \ln\prod_{i=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2) \\
&= \sum_{i=1}^{N}\ln\mathcal{N}(x_n|\mu,\sigma^2) \\
&= \sum_{i=1}^{N}\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right) \\
&= N\ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi)
\end{aligned}$$

2. By maximizing (2) with respect to μ (i.e., take the partial derivative with respect to μ and set to zero), we obtain the maximum likelihood solution μ_{ML}. Verify that it is given by

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \equiv \bar{x} \tag{3}$$

3. In the previous item, you may have noticed that the maximum likelihood solution μ_{ML} does not depend on σ^2. We can now substitute the solution $\mu = \mu_{\mathrm{ML}} = \bar{x}$ in (2) and maximize the result with respect to σ^2_{ML} (i.e., take the partial derivative with respect to σ^2 and set to zero), we then obtain the maximum likelihood solution σ^2_{ML}. Verify that it is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \tag{4}$$

ANSWER: Taking the derivative of (2) with respect to μ gives $\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$. Set equal to zero and rearranging terms, this gives the solution

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Note that this solution does not depend on σ^2.

Maximizing with respect to σ^2: substitute $\mu(= \mu_{\mathrm{ML}}) = \bar{x}$ and $v = \sigma^2$ in (2). Take the derivative w.r.t. v and set equal to zero.

$$\frac{1}{2v^2}\sum_{n=1}^{N}(x_n - \bar{x})^2 - \frac{N}{2v} = 0$$

So

$$\frac{N}{2v^2}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 - v\right) = 0$$

This means that either $v = \sigma^2 = \infty$ (but that is actually a minimum since $\ln(\sigma^2) \to \infty$), or $v = \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$ which is indeed a maximum and therefore the ML solution.

Exercise 2

Maximum likelihood estimate of variance underestimates true variance (Bishop p 27).

In this exercise, we will make use of definitions and results we have seen in previous exercises:

$$\begin{aligned} \mathbb{E}[x+z] &= \mathbb{E}[x] + \mathbb{E}[z] \qquad (5)\\ \mathbb{E}[cx] &= c\mathbb{E}[x] \qquad (6)\\ \mathrm{var}[f] &= \mathbb{E}[f^2] - (\mathbb{E}[f])^2 \qquad (7) \end{aligned}$$

and for independent variables,

$$\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z] \tag{8}$$

The maximum likelihood solutions for the univariate Gaussian, μ_{ML} and σ_{ML}, are functions of the data set values $x_1, \ldots, x_N$,

$$\begin{aligned} \mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n \quad (9) \\ \sigma^2_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{k=1}^{N} x_k)^2 \quad (10) \end{aligned}$$

Now assume that data is generated i.i.d from a univariate Gaussian with parameters μ and σ^2, (so $p(x_n) = \mathcal{N}(x_n|\mu, \sigma^2)$ for all n).

1. Show, using result (5), that:

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \quad (11)$$

ANSWER:

$$\begin{aligned} \mathbb{E}[\mu_{\mathrm{ML}}] &= \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} x_n] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mu \\ &= \mu \end{aligned}$$

2. To compute the expectation of σ^2_{ML}, one has to be a bit careful with the bookkeeping. (Hint: Expand the square and use the fact that $\mathbb{E}[x_i^2] = \mu^2 + \sigma^2$ and $\mathbb{E}[x_i x_j] = \mu^2$ for $i \neq j$, since the draws are independent.) Show that:

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \frac{N-1}{N}\sigma^2$$

ANSWER:

$$\begin{aligned} \mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{k=1}^{N} x_k)^2\right] \\ &\overset{(5),(6)}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[(x_n - \frac{1}{N}\sum_{k=1}^{N} x_k)^2\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) + \left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right] \\ &\overset{(5)}{=} \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}[x_n^2] - 2\cdot\mathbb{E}\left[x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right]\right\} \end{aligned}$$

We now compute the three expected value terms inside the curly brackets separately:

(a)

$$\mathbb{E}[x_n^2] \overset{(7)}{=} \mathbb{E}[x_n]^2 + \text{var}[x_n] = \mu^2 + \sigma^2$$

(b)

$$\begin{aligned}
\mathbb{E}\left[x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] &= \frac{1}{N}\left\{\mathbb{E}[x_n^2] + \sum_{k=1,k\neq n}^{N} \mathbb{E}[x_n x_k]\right\} \\
&\overset{(a),(8)}{=} \frac{1}{N}\left\{\mu^2 + \sigma^2 + (N-1)\mu^2\right\} \\
&= \mu^2 + \frac{1}{N}\sigma^2
\end{aligned}$$

(c)

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] \\
&\overset{(5),(6)}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] \\
&\overset{(b)}{=} \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \frac{1}{N}\sigma^2\right) \\
&= \mu^2 + \frac{1}{N}\sigma^2
\end{aligned}$$

We now combine the values we obtained for the three terms to get our result:

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\text{ML}}] &= \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}[x_n^2] - 2\cdot\mathbb{E}\left[x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right]\right\} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{(\mu^2+\sigma^2) - 2(\mu^2 + \frac{1}{N}\sigma^2) + (\mu^2 + \frac{1}{N}\sigma^2)\right\} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{N-1}{N}\sigma^2\right\} \\
&= \frac{N-1}{N}\sigma^2
\end{aligned}$$

Exercise 3

The general expression of a univariate Gaussian with mean μ and variance σ^2 is

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{12}$$

The general expression of a multivariate Gaussian over a D dimensional vector $\mathbf{x}$ with D dimensional mean vector $\boldsymbol{\mu}$ and $D \times D$ covariance matrix $\boldsymbol{\Sigma}$ is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{13}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

Now consider a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in which the covariance matrix $\boldsymbol{\Sigma}$ is a diagonal matrix, i.e., its elements can be written as $\Sigma_{ij} = \sigma_i^2 I_{ij}$, where I_{ij} are the matrix elements of the identity matrix (so $I_{ij} = 0$ if $i \neq j$ and $I_{ii} = 1$).

- Show, using (12) and (13) that a multivariate Gaussian with diagonal covariance matrix, $\Sigma_{ij} = \sigma_i^2 I_{ij}$, factorizes into a product of univariate Gaussians

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{D} \mathcal{N}(x_i|\mu_i, \sigma_i^2)$$

ANSWER: First we note that the inverse covariance matrix has elements

$$(\boldsymbol{\Sigma}^{-1})_{ij} = \frac{1}{\sigma_i^2} I_{ij}$$

The next step is to evaluate the exponent in the multivariate Gaussian.

$$\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) &= -\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(\mathbf{x}-\boldsymbol{\mu})_i(\boldsymbol{\Sigma}^{-1})_{ij}(\mathbf{x}-\boldsymbol{\mu})_j \\
&= \sum_{i=1}^{D}\sum_{j=1}^{D}\left\{-\frac{1}{2\sigma_i^2}(x_i-\mu_i)I_{ij}(x_j-\mu_j)\right\} \\
&= \sum_{i=1}^{D}\left\{-\frac{1}{2\sigma_i^2}(x_i-\mu_i)^2\right\}
\end{aligned}$$

Then we have a look at the determinant $|\Sigma|$ which appear in the denominator. Since Σ is diagonal, the determinant is just the product of its diagonal terms, so

$$|\boldsymbol{\Sigma}| = \prod_{i=1}^{D} \sigma_i^2$$

Now, we combine results. The multivariate Gaussian can be written as

$$\begin{aligned}
\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{D/2}}\frac{1}{(\prod_{i=1}^{D}\sigma_i^2)^{1/2}}\exp\left\{\sum_{i=1}^{D}-\frac{1}{2\sigma_i^2}(x_i-\mu_i)^2\right\} \\
&= \prod_{i=1}^{D}\left(\frac{1}{(2\pi\sigma_i^2)^{1/2}}\right)\prod_{i=1}^{D}\exp\left\{-\frac{1}{2\sigma_i^2}(x_i-\mu_i)^2\right\} \\
&= \prod_{i=1}^{D}\frac{1}{(2\pi\sigma_i^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma_i^2}(x_i-\mu_i)^2\right\} \\
&= \prod_{i=1}^{D}\mathcal{N}(x_i|\mu_i, \sigma_i^2)
\end{aligned}$$

Exercise 4

Curve fitting of a polynomial of the familiar form $y(x;\mathbf{w}) = \sum_{j=0}^{M} w_j x^j$ based on training data of N inputs $\mathbf{x} = (x_1, \ldots, x_N)$ and N outputs $\mathbf{t} = (t_1, \ldots, t_N)$ by the MAP solution.

Given the prior of the M-dimensional parameter vector $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) \tag{14}$$

with given hyperparameter α, and the likelihood, with given β

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1}), \tag{15}$$

then the posterior can be found by applying Bayes' rule

$$p(\mathbf{w}|\mathbf{t},\mathbf{x},\alpha,\beta) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{x},\alpha,\beta)p(\mathbf{w}|\mathbf{x},\alpha,\beta)}{p(\mathbf{t}|\mathbf{x},\alpha,\beta)} \tag{16}$$

1. Provide an interpretation (in your own words) of what the prior (14) represents. Do you think this is a reasonable prior or could you come up with a better one?

   ANSWER: The prior $p(\mathbf{w}|\alpha)$ is such that the weights are considered most likely to be small, either positive or negative; the higher precision parameter α, the more closely centered around zero. Given that there is just a single hyperparameter, all weights are treated the same.
   From what we saw before (introduction of regularizer to keep weights small in order to avoid overfitting), this seems a reasonable assumption, although one could argue that it is desirable to be able to penalize higher order weight terms more than lower order ones.

2. Show that for the given prior and likelihood the posterior is proportional to $p(\mathbf{t}|\mathbf{w},\mathbf{x},\beta)p(\mathbf{w}|\alpha)$, and that the MAP solution $\mathbf{w}_{MAP}$ that maximizes this posterior distribution is equal to the parameter vector that minimizes

$$\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{17}$$

   ANSWER: In (16) the normalization factor $p(\mathbf{t}|\mathbf{x},\alpha,\beta)$ is not dependent on $\mathbf{w}$, and is therefore a constant *given* $\{\mathbf{t},\mathbf{x},\alpha,\beta\ \}$. From eqs. (15) and (14) we see that the likelihood and prior are dependent only on resp. $\{\mathbf{w},\mathbf{x},\beta\}$ and $\{\alpha\}$, and therefore

$$p(\mathbf{w}|\mathbf{t},\mathbf{x},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{w},\mathbf{x},\beta)p(\mathbf{w}|\alpha) \tag{18}$$

   where '$\propto$' means 'is proportional to'. The location of the maximum of the posterior w.r.t. $\mathbf{w}$ does not change when it is multiplied by a constant $c > 0$ or when it is transformed by a monotonically increasing function, e.g. a logarithm. Since maximizing is equivalent to minimizing the negative, the problem is equivalent to minimizing $-\ln p(\mathbf{t}|\mathbf{w},\mathbf{x},\beta) - \ln p(\mathbf{w}|\alpha)$. Plugging in the definitions from (14) and (15) (cf. eq.1.54 in Bishop) then results in

$$\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \frac{N}{2}\ln(\beta^{-1}) + \frac{(M+1)}{2}\ln(\alpha) + \frac{(N+M+1)}{2}\ln(2\pi) \tag{19}$$

   in which the last 3 terms are constants and can be dropped. Therefore the MAP solution is given by the weights vector $\mathbf{w}_{MAP}$ that minimizes (17).

3. Why is this not yet a fully 'Bayesian' approach? What would be required to make it so, and what would be the (qualitative) impact on the result?

   ANSWER: In this example, the aim of Bayesian fitting is to find a prediction t for a new value x given the available data $(\mathbf{x}, \mathbf{t})$ and prior information α and β, that also expresses our uncertainty about the answer, from a consistent application of the rules of probability. In other words, in a fully Bayesian approach we try to find: $p(t|x, \mathbf{t}, \mathbf{x}, \alpha, \beta)$ using (only) the sum and product rules.
   In the MAP solution, we find $p(t|x, \mathbf{w}_{MAP})$, using a single weight vector $\mathbf{w}_{MAP}$, found by minimizing (17). In a fully Bayesian approach *all* possible weight vectors $\mathbf{w}$ should be taken into account, meaning that we are looking for a solution in the form

$$p(t|x, \mathbf{t}, \mathbf{x}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{t}, \mathbf{x}) \mathrm{d}\mathbf{w} \tag{20}$$

   As a result, not only the mean but also the variance in the prediction t becomes dependent on x. Exactly how to do all this will be tackled later on in the course.

Exercise 5

What does a high dimensional cube look like? Consider a hypercube with sides $2a$ in D-dimensions.

1. Calculate the ratio of the distance from the center of the hypercube to one if its corners, divided by the perpendicular distance to one of its sides.

   ANSWER: The perpendicular distance from the center to one of the sides is simply $d_s = a$. For the distance to the corner use Pythagoras:

$$d_c^2 = \left(a_1^2 + \ldots + a_D^2\right) = Da^2$$

   and so for the ratio we have

$$\frac{d_c}{d_s} = \frac{\sqrt{Da^2}}{a} = \sqrt{D}$$

Now consider a hypersphere of radius a in D-dimensions that just touches the hypercube at the centers of its sides. In Bishop, ex.1.19, the following approximation for the volume of a sphere with radius a in high dimensions $D \gg 1$ is derived

$$V_S = \frac{a^D 2\pi^{D/2}}{D\Gamma(D/2)} \approx \frac{a^D 2\pi^{D/2}}{D\sqrt{2\pi}\mathrm{e}^{-(D/2-1)} \cdot (D/2-1)^{D/2-1}} \tag{21}$$

2. Calculate the ratio of the volume of the hypersphere divided by the volume of the cube as $D \to \infty$. What do these answers tell you about the shape of a cube in high dimensions? Hint: no exact calculation, only the behaviour in the limit $D \to \infty$.

   ANSWER: Dividing V_S by the volume of the cube $V_C = (2a)^D$ and ignoring irrelevant factors in expression (21) in the limit $D \to \infty$ gives

$$\begin{aligned}
\frac{\text{volume of sphere}}{\text{volume of cube}} &= \frac{V_S}{V_D} \\
&\approx \frac{a^D 2\pi^{D/2}}{(2a)^D \cdot D\sqrt{2\pi}\mathrm{e}^{-(D/2)} \cdot (D/2)^{D/2}} \\
&\approx \left(\frac{\pi \mathrm{e}}{2D}\right)^{D/2} \cdot \frac{1}{D} \\
&\approx \left(\frac{1}{\infty}\right)^{\infty} \cdot \frac{1}{\infty} \\
&= 0
\end{aligned}$$

It means that in high dimensions almost the entire volume of a cube is contained in its corners which themselves become very elongated ‘spikes’.

3. Try to interpret this result in terms of what it means for a dataset $\mathbf{X}$ consisting of N i.i.d. observations of a vector valued variable $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ drawn from a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with both N and D large.

   ANSWER: It means that if the number of elements D in $\mathbf{x}$ is very large then it is extremely unlikely to find a completely ‘average’ observation, i.e. an observation with $\mathbf{x} \approx \boldsymbol{\mu}$. Furthermore, with large N it is almost guaranteed that many extreme values for the individual elements x_i will be observed in the data set. (Variant of the law of large numbers).

BONUS PRACTICE

Exercise 6

(Exercise 1.14 from Bishop.)

1. Show that a matrix $\mathbf{W}$ with elements w_{ij} can be written as the sum of a symmetric matrix $\mathbf{W}^S$ and an anti-symmetric matrix $\mathbf{W}^A$. In other words, show that

$$w_{ij} = w_{ij}^S + w_{ij}^A \tag{22}$$

with symmetric matrix elements $w_{ij}^S = (w_{ij} + w_{ji})/2$ and anti-symmetric matrix elements $w_{ij}^A = (w_{ij} - w_{ji})/2$. Verify that $w_{ij}^S = w_{ji}^S$ and $w_{ij}^A = -w_{ji}^A$.

ANSWER: For a matrix with elements w_{ij} we have

$$w_{ij} = \frac{1}{2}[w_{ij} + w_{ji}] + \frac{1}{2}[w_{ij} - w_{ji}] = w_{ij}^S + w_{ij}^A$$

where $\mathbf{W}^S$ is symmetric

$$w_{ij}^S = \frac{1}{2}[w_{ij} + w_{ji}] = \frac{1}{2}[w_{ji} + w_{ij}] = w_{ji}^S$$

and $\mathbf{W}^A$ is anti-symmetric

$$w_{ij}^A = \frac{1}{2}[w_{ij} - w_{ji}] = -\frac{1}{2}[w_{ji} - w_{ij}] = -w_{ji}^A$$

2. Consider the 2^{nd} order terms in a 2^{nd} order polynomial in d dimensions, i.e. $\mathbf{x} = (x_1, \ldots, x_d)^T$.

$$\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}x_ix_j$$

Show that

$$\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}x_ix_j = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^S x_ix_j \tag{23}$$

i.e. there is no contribution from anti-symmetric matrix elements. This demonstrates that, without loss of generality, in problems involving (only) quadratic terms a matrix W can be taken to be *symmetric*, i.e. $W = W^S$.

ANSWER: We need to show that only the symmetric part contributes to the overall sum. By using (22) we can split the sum in two parts: one for the symmetric matrix W^S and one for the anti-symmetric matrix W^A:

$$\begin{aligned}\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}x_ix_j &= \sum_{i=1}^{d}\sum_{j=1}^{d}[w_{ij}^S + w_{ij}^A]x_ix_j \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^S x_ix_j + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^A x_ix_j\end{aligned}$$

Using the anti-symmetric property of w^A in combination with the symmetry in $x_i x_j = x_j x_i$, the interchangeability of the summation sequence $\sum_{i=1}^d \sum_{j=1}^d \ldots = \sum_{j=1}^d \sum_{i=1}^d \ldots$, and the fact the that i and j are 'just' dummy-indices we find for the second sum S^A:

$$\begin{aligned} S^A &= \sum_{i=1}^d \sum_{j=1}^d w_{ij}^A x_i x_j \\ &= \sum_{j=1}^d \sum_{i=1}^d w_{ji}^A x_j x_i \\ &= -\sum_{j=1}^d \sum_{i=1}^d w_{ij}^A x_j x_i \\ &= -\sum_{i=1}^d \sum_{j=1}^d w_{ij}^A x_i x_j \\ &= -S^A \end{aligned}$$

and $S^A = -S^A \Leftrightarrow S^A = 0$. Therefore the net contribution from the anti-symmetric sum is zero, and so

$$\sum_{i=1}^d \sum_{j=1}^d w_{ij} x_i x_j = \sum_{i=1}^d \sum_{j=1}^d w_{ij}^S x_i x_j$$

Note: This property will feature prominently when the multivariate Gaussian distribution is discussed.

3. Show that the previous statement can also be stated in matrix notation as

$$\mathbf{x}^{\mathrm{T}} \mathbf{W} \mathbf{x} = \mathbf{x}^{\mathrm{T}} \mathbf{W}^S \mathbf{x} \tag{24}$$

with $\mathbf{W}^S = \frac{1}{2}\left(\mathbf{W} + \mathbf{W}^{\mathrm{T}}\right)$, the symmetric part of matrix $\mathbf{W}$.

ANSWER: From the definition of matrix multiplication, and the fact that a vector is just a single column matrix, the left hand side of (24) is compactly written as

$$\begin{aligned} \sum_{i=1}^d \sum_{j=1}^d w_{ij} x_i x_j &= \sum_{i=1}^d \sum_{j=1}^d [w_{ij}^S + w_{ij}^A] x_i x_j \\ &= \sum_{i=1}^d \sum_{j=1}^d w_{ij}^S x_i x_j + \underbrace{\sum_{i=1}^d \sum_{j=1}^d w_{ij}^A x_i x_j}_{=0} \\ &= \sum_{i=1}^d x_i \sum_{j=1}^d w_{ij}^S x_{j1} \\ &= \sum_{i=1}^d x_{1i} \left(\mathbf{W}^S \mathbf{x}\right)_{i1} \\ &= \left(\mathbf{x}^{\mathrm{T}} \mathbf{W}^S \mathbf{x}\right)_{11} \\ &= \mathbf{x}^{\mathrm{T}} \mathbf{W}^S \mathbf{x} \end{aligned}$$

where the last step follows from the fact that the overall result is a scalar, i.e. a single cell matrix. The right hand side goes similar, after substituting the definition for the symmetric part of a matrix.

Exercise 7

The determinant of an $N \times N$ matrix $\mathbf{A}$ can be calculated using Laplace's formula as

$$\det(\mathbf{A}) = \sum_{j=1}^{n} A_{ij}(-1)^{i+j}\det(\mathbf{M}_{ij}) \tag{25}$$

where A_{ij} is the element in $\mathbf{A}$ at row i, column j, and $\mathbf{M}_{ij}$ is the smaller matrix obtained by removing the i-th row and j-th column from $\mathbf{A}$. (The determinant of submatrix $\mathbf{M}_{ij}$ is also known as the *minor* M_{ij}.)

1. Calculate $|\mathbf{A}|$, the determinant of the matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 0 \\ \text{-1} & 1 & 3 \\ 2 & 0 & \text{-1} \end{pmatrix}$$

ANSWER: Expanding the determinant along the first row, using Laplace's formula (25) with $i = 1$ we find

$$\begin{aligned} \det(\mathbf{A}) &= 2 \cdot \det\left(\begin{bmatrix} 1 & 3 \\ 0 & -1 \end{bmatrix}\right) - 2 \cdot \det\left(\begin{bmatrix} -1 & 3 \\ 2 & -1 \end{bmatrix}\right) + 0 \cdot \det\left(\begin{bmatrix} -1 & 1 \\ 2 & 0 \end{bmatrix}\right) \\ &= 2 \cdot (-1 - 0) - 2 \cdot (1 - 6) + 0 \cdot (0 - 2) \\ &= -2 + 10 + 0 \\ &= 8 \end{aligned}$$

2. Verify that the determinant of a diagonal matrix $\mathbf{\Lambda}$ is just the product of its elements.

ANSWER: A diagonal matrix $\mathbf{\Lambda}$ only has nonzero elements Λ_{ii}, and so $\Lambda_{i \neq j} = 0$. Substituting in (25) and expanding along the first row of $\mathbf{\Lambda}$ gives

$$\begin{aligned} |\mathbf{\Lambda}| &= \sum_{j=1}^{n} \Lambda_{1j}(-1)^{1+j}\det(\mathbf{M}_{1j}) \\ &= \Lambda_{11} \cdot \det(\mathbf{M}_{11}) - 0 \cdot \det(\mathbf{M}_{12}) + \cdots + (-1)^{1+n} \cdot 0 \cdot \det(\mathbf{M}_{1n}) \\ &= \Lambda_{11} \cdot \det(\mathbf{\Lambda}_{[2\ldots n, 2\ldots n]}) \end{aligned}$$

in which $\mathbf{\Lambda}_{[2\ldots n,2\ldots n]}$ is a diagonal matrix of size $n - 1$. Repeatedly expanding along the first row of each subsequent submatrix then results in

$$\begin{aligned} |\mathbf{\Lambda}| &= \Lambda_{11} \cdot \Lambda_{22} \cdot \ldots \cdot \Lambda_{nn} \\ &= \prod_{i=1}^{n} \Lambda_{ii} \end{aligned}$$

3. The determinant of the product of two matrices is given by $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.
Use this to show that for the determinant of an inverse matrix

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} \tag{26}$$

What does this tell you about the existence of the inverse of a matrix $\mathbf{A}$?

ANSWER: The inverse $\mathbf{A}^{-1}$ of an $N \times N$ matrix $\mathbf{A}$ is defined as

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_N$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Since $\mathbf{I}_N$ is a diagonal matrix with only 1s on the diagonal, from the previous question we have $|\mathbf{I}_N| = 1$.
With the product rule for matrix determinants this gives

$$|\mathbf{A}^{-1}||\mathbf{A}| = |\mathbf{A}\mathbf{A}^{-1}| = |\mathbf{I}_N| = 1$$

which reduces to (26) after dividing both sides by $|\mathbf{A}|$. It suggests (and is in fact true) that the inverse of a matrix with $|\mathbf{A}| = 0$ is not defined.

Exercise 8

Matrix identities (Exercises 2.24 and 2.26 in Bishop).

- Prove the *partitioned matrix inversion formula*:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix},$$

where $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$. This identity is used, for example, to simplify the expression of the inverse of the precision in a linear Gaussian model (see Bishop (2.104) and (2.105)).

ANSWER: If we multiply the right-hand side of the equation by $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$, we get:

$$\begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{M}\mathbf{A} - \mathbf{M}\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{M}\mathbf{B} - \mathbf{M}\mathbf{B}\mathbf{D}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{A} + \mathbf{D}^{-1}\mathbf{C} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & -\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B} + \mathbf{D}^{-1}\mathbf{D} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\mathbf{D} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{M}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) & \mathbf{M}\mathbf{B} - \mathbf{M}\mathbf{B} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) + \mathbf{D}^{-1}\mathbf{C} & -\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B} + \mathbf{I} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} + \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

$$= \mathbf{I}$$

We can now right multiply by $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1}$ to obtain the required identity. □

- The *Woodbury matrix inversion formula* (see below) is useful when we have a large diagonal matrix $\mathbf{A}$, which is easy to invert, while $\mathbf{B}$ has many rows, but few columns (and conversely for $\mathbf{D}$), so that the right-hand side is much cheaper to evaluate than the left-hand side. A common application is finding the inverse of a low-rank update $\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}$ of $\mathbf{A}$, for example in the Kalman filter algorithm. Prove the correctness of the identity, which is given by:

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

ANSWER: If we multiply the right-hand side of the equation by $(\mathbf{A}+\mathbf{BCD})$, we get:

$$
\begin{aligned}
&(\mathbf{A}+\mathbf{BCD})(\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}) = \\
&= \mathbf{I}+\mathbf{BCDA}^{-1}-\mathbf{AA}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}-\mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&= \mathbf{I}+\mathbf{BCDA}^{-1}-\mathbf{BCC}^{-1}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}-\mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&= \mathbf{I}+\mathbf{BCDA}^{-1}-\mathbf{BC}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&= \mathbf{I}+\mathbf{BCDA}^{-1}-\mathbf{BCDA}^{-1} \\
&= \mathbf{I}
\end{aligned}
$$

We can now left multiply by $(\mathbf{A}+\mathbf{BCD})^{-1}$ to obtain the required identity. □

Exercise 9

(Exercise 2.34 in Bishop) Find the maximum likelihood solution for the covariance matrix of a multivariate Gaussian by maximizing the log likelihood function

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

with respect to $\boldsymbol{\Sigma}$. In order to perform a straightforward maximization, ignore the constraints of symmetry and positive definitiness on $\boldsymbol{\Sigma}$, i.e. treat $\boldsymbol{\Sigma}$ as if it contained D^2 free parameters instead of just $\frac{D(D+1)}{2}$.

Hint: Use the results from Appendix C in Bishop to compute the matrix derivatives.

ANSWER:

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}} &= -\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) \\
&= -\frac{N}{2}(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}-\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\sum_{n=1}^{N}\mathsf{Tr}\,[\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}] \\
&= -\frac{N}{2}\boldsymbol{\Sigma}^{-1}-\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathsf{Tr}\,[\boldsymbol{\Sigma}^{-1}\mathbf{S}], \text{ where } \mathbf{S}=\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}.
\end{aligned}
$$

For each element (i,j) in $\boldsymbol{\Sigma}$, which we denote Σ_{ij}, we get:

$$
\begin{aligned}
\frac{\partial}{\partial\Sigma_{ij}}\mathsf{Tr}\,[\boldsymbol{\Sigma}^{-1}\mathbf{S}] &= \mathsf{Tr}\,[\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\Sigma_{ij}}\mathbf{S}] \\
&= -\mathsf{Tr}\,[\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}] \\
&= -\mathsf{Tr}\,[\frac{\partial\boldsymbol{\Sigma}}{\partial\Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}] \\
&= -(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})_{ij}
\end{aligned}
$$

Note that in the last step we have ignored the fact that $\Sigma_{ij}=\Sigma_{ji}$, so that $\frac{\partial\boldsymbol{\Sigma}}{\partial\Sigma_{ij}}$ has a one only in position (i,j) and zero everywhere else. Nevertheless, treating the last result as valid, we have:

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}} &= -\frac{N}{2}\boldsymbol{\Sigma}^{-1}-\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathsf{Tr}\,[\boldsymbol{\Sigma}^{-1}\mathbf{S}] \\
&= -\frac{N}{2}\boldsymbol{\Sigma}^{-1}+\frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}
\end{aligned}
$$

Setting the last expression to zero, we obtain $\frac{N}{2}\boldsymbol{\Sigma}^{-1}=\frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1} \iff \boldsymbol{\Sigma}=\mathbf{S}$.