Statistical Machine Learning 2016

Exercises and answers, week 4

22 September 2016

Exercise 1

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

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

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

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

ANSWER: Expanding the determinant along the first row, using Laplace's formula (1) 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 (1) 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

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} \tag{2}$$

   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 (2) 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 2

Properties of the univariate Gaussian distribution. The probability density of a univariate Gaussian x with mean μ and variance σ^2 is given by:

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

1. Show, using the result on page 49 of the slides, which states

$$\int_{-\infty}^{\infty} \exp(-x^2)\, dx = \sqrt{\pi}$$

   that the univariate Gaussian density is properly normalized.

2. Calculate the expected value of x (Hint: use a change of variables).

3. Calculate the variance of x (Hint: differentiate both sides of the normalization condition for $p(x)$ with respect to σ^2).

4. Calculate the mode of x (i.e., the value of x that has maximum probability density).

ANSWER:

1. We perform a change of variables $z = \frac{x-\mu}{\sqrt{2}\sigma}$, noting that the integration region $(-\infty, \infty)$ transforms into itself:

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x)\, dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-z^2) \frac{dx}{dz} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-z^2) \sqrt{2}\sigma\, dz \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-z^2)\, dz = 1.
\end{aligned}$$

2. We use the same change of variables to calculate $\mathbb{E}(x)$:

$$\begin{aligned}
\mathbb{E}(x) &= \int_{-\infty}^{\infty} xp(x)\,dx \\
&= \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,dx \\
&= \int_{-\infty}^{\infty} (z\sqrt{2}\sigma+\mu)\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-z^2)\,\sqrt{2}\sigma dz \\
&= \sqrt{2}\sigma\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} z\exp(-z^2)dz + \mu\int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}\exp(-z^2)dz \\
&= \sqrt{2}\sigma\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} z\exp(-z^2)dz + \mu
\end{aligned}$$

The remaining integral vanishes, as the integrand is an odd function (a function $f(x)$ is odd if $f(-x) = -f(x)$) and the integration interval is invariant under sign reversal (to show this explicitly, write the integral as the sum of two integrals, one from $-\infty$ to 0 and the other from 0 to ∞ and then show that these two integrals cancel).

3. The normalization condition reads:

$$1 = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,dx$$

Differentiating this equation with respect to σ^2 (note that we are allowed interchange the integration over x and the differentiation with respect to σ^2 as the integration domain does not depend on σ^2) yields:

$$0 = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\left[-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4}\right]\,dx = \mathbb{E}\left(-\frac{1}{2\sigma^2}+\frac{(x-\mu)^2}{2\sigma^4}\right)$$

Rewriting this (using linearity of expectation):

$$0 = -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\mathbb{E}\left((x-\mu)^2\right) = \frac{1}{2\sigma^2}\left(-1+\frac{\mathbb{V}\mathrm{ar}(x)}{\sigma^2}\right)$$

and hence

$$\mathbb{V}\mathrm{ar}(x) = \sigma^2.$$

4. At the mode (maximum) $\hat{x}$ of $p(x)$, $p'(\hat{x}) = 0$. So

$$p'(\hat{x}) = -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\hat{x}-\mu)^2}{2\sigma^2}\right)\frac{(\hat{x}-\mu)}{\sigma^2} = 0$$

which only happens for $\hat{x} = \mu$. As we can check that $p''(x) < 0$, this is indeed a maximum.

Exercise 3

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:

$$\mathbb{E}[x+z] = \mathbb{E}[x]+\mathbb{E}[z] \quad (3)$$

$$\mathbb{E}[cx] = c\mathbb{E}[x] \quad (4)$$

$$\mathrm{var}[f] = \mathbb{E}[f^2]-(\mathbb{E}[f])^2 \quad (5)$$

and for independent variables,

$$\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z] \tag{6}$$

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

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{7}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{k=1}^{N} x_k)^2 \tag{8}$$

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 (3), that:

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{9}$$

ANSWER:

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

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:

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{k=1}^{N} x_k)^2\right] \\
&\overset{(3),(4)}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[(x_n - \frac{1}{N}\sum_{k=1}^{N} x_k)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) + \left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right] \\
&\overset{(3)}{=} \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbb{E}[x_n^2] - 2\cdot\mathbb{E}\left[x_n\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right]\right\}
\end{aligned}$$

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

(a)

$$\mathbb{E}[x_n^2] \overset{(5)}{=} \mathbb{E}[x_n]^2 + \mathrm{var}[x_n] = \mu^2 + \sigma^2$$

(b)

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

(c)

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] \\
&\overset{(3),(4)}{=} \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[x_n \left(\frac{1}{N}\sum_{k=1}^{N} x_k\right)\right] \\
&\overset{(b)}{=} \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \frac{1}{N}\sigma^2\right) \\
&= \mu^2 + \frac{1}{N}\sigma^2
\end{aligned}$$

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

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

Exercise 4

More about multivariate Gaussians.

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

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{10}$$

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

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{11}$$

where $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$.
Now consider a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$ in which the covariance matrix $\mathbf{\Sigma}$ 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 (10) and (11) that a multivariate Gaussian with diagonal covariance matrix, $\Sigma_{ij} = \sigma_i^2 I_{ij}$, factorizes into a product of univariate Gaussians

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

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

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

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

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

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

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

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

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

Exercise 5

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 $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) \tag{12}$$

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

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}), \tag{13}$$

then the posterior can be found by applying Bayes' rule

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \alpha, \beta) = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{x}, \alpha, \beta)p(\mathbf{w}|\mathbf{x}, \alpha, \beta)}{p(\mathbf{t}|\mathbf{x}, \alpha, \beta)} \tag{14}$$

1. Provide an interpretation (in your own words) of what the prior (12) 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

   $$\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{15}$$

   ANSWER: In (14) 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. (13) and (12) we see that the likelihood and prior are dependent only on resp. $\{\mathbf{w}, \mathbf{x}, \beta\}$ and $\{\alpha\}$, and therefore

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

   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 (12) and (13) (cf. eq.1.54 in Bishop) then results in

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

   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 (15).

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 (15). 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

$$p(t|x, \mathbf{t}, \mathbf{x}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{t}, \mathbf{x}) \mathrm{d}\mathbf{w} \tag{18}$$

   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 6

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

$$V_S = \frac{a^D 2\pi^{D/2}}{D\Gamma(D/2)} \approx \frac{a^D 2\pi^{D/2}}{D\sqrt{2\pi}\mathrm{e}^{-(D/2-1)} \cdot (D/2-1)^{D/2-1}} \tag{19}$$

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 (19) 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).