Statistical Machine Learning 2018

Exercises, 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

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

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

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

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:

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

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 $x_1, \ldots, x_N$,

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

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:
$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{11}$$

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

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

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 $\mathbf{x} = (x_1, \dots, x_N)$ and N outputs $\mathbf{t} = (t_1, \dots, t_N)$ 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?

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

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

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?

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.

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

$$V_S = \frac{a^D 2\pi^{D/2}}{D\Gamma(D/2)} \approx \frac{a^D 2\pi^{D/2}}{D\sqrt{2\pi}\mathrm{e}^{-(D/2-1)} \cdot (D/2-1)^{D/2-1}} \tag{18}$$

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 $D \to \infty$.

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 $\mathbf{x} = (x_1, \dots, x_D)^{\mathrm{T}}$ drawn from a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with both N and D large.

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

$$w_{ij} = w_{ij}^S + w_{ij}^A \tag{19}$$

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 $w_{ij}^A = -w_{ji}^A$.

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

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

Show that

$$\sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j = \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij}^S x_i x_j \tag{20}$$

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. $W = W^S$.

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

$$\mathbf{x}^{\mathrm{T}} \mathbf{W} \mathbf{x} = \mathbf{x}^{\mathrm{T}} \mathbf{W}^S \mathbf{x} \tag{21}$$

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

Exercise 7

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

$$\det(\mathbf{A}) = \sum_{j=1}^{n} A_{ij} (-1)^{i+j} \det(\mathbf{M}_{ij}) \tag{22}$$

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

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

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

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

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} \tag{23}$$

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

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)).

- 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:

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

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.