# Statistical Machine Learning 2016

Exercises, 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

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

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

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

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

## Exercise 2

Properties of the univariate Gaussian distribution. The probability density of a univariate Gaussian $x$ with mean $\mu$ and variance $\sigma^2$ 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 $\sigma^2$).

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

# 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:

$$\begin{aligned} \mathbb{E}[x+z] &= \mathbb{E}[x] + \mathbb{E}[z] && (3) \\ \mathbb{E}[cx] &= c\mathbb{E}[x] && (4) \\ \mathrm{var}[f] &= \mathbb{E}[f^2] - (\mathbb{E}[f])^2 && (5) \end{aligned}$$

and for independent variables,

$$\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z] \quad (6)$$

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

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

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

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

1. Show, using result (3), that:
$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \quad (9)$$
2. To compute the expectation of $\sigma^2_{\mathrm{ML}}$, 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 4

More about multivariate Gaussians.

The general expression of a univariate Gaussian with mean $\mu$ and variance $\sigma^2$ is

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

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

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

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

## 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 $\alpha$, and the likelihood, with given $\beta$

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

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?

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

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

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 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.

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

$$V_S = \frac{a^D 2\pi^{D/2}}{D\Gamma(D/2)} \approx \frac{a^D 2\pi^{D/2}}{D\sqrt{2\pi}\mathrm{e}^{-(D/2-1)} \cdot (D/2-1)^{D/2-1}} \tag{16}$$

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