Julien Fer
10649441
julien.fer88@gmail.com

Homework Assignment 1
Machine Learning 1, 19/20

2019-09-18

# 1 Basic Linear Algebra and Derivatives

This part was not required to hand in.

# 2 Multivariate Calculus

## 2.1 Question 2.1

**a)** The following gradient is asked: $\nabla_{\boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. In order to find the gradient the following variables are defined:

(1) $\mathbf{y}(\boldsymbol{\mu}) \coloneqq \mathbf{x}-\boldsymbol{\mu}$

(2) $\mathbf{f}(\mathbf{y}(\boldsymbol{\mu})) \coloneqq \mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$

and thus by combining (1) and (2) in the original equation we get:

(3) $\nabla_{\boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}}\mathbf{f} = \dfrac{\partial \mathbf{f}}{\partial \mathbf{y}}\dfrac{\partial \mathbf{y}}{\partial \boldsymbol{\mu}}$ with

(4) $\dfrac{\partial \mathbf{f}}{\partial \mathbf{y}} = \mathbf{y}^T(\boldsymbol{\Sigma}^{-1}+(\boldsymbol{\Sigma}^{-1})^T) = 2\mathbf{y}^T\boldsymbol{\Sigma}^{-1}$ (as $\boldsymbol{\Sigma}^{-1}$ is symmetric)

(5) $\dfrac{\partial \mathbf{y}}{\partial \boldsymbol{\mu}} = -\boldsymbol{I}$

and by finally combining (4) and (5) in (3) we get the gradient:

$$\nabla_{\boldsymbol{\mu}}\mathbf{f} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \boldsymbol{\mu}} = -2\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{I} = -2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}$$

**b)** The derivation of the gradient is as follows:

$$\begin{aligned}
\nabla_{\mathbf{q}} - \mathbf{p}^T\log(\mathbf{q}) &= -\mathbf{p}^T\nabla_{\mathbf{q}}\log(\mathbf{q}) \\
&= -\mathbf{p}^T\begin{bmatrix} \frac{\partial \log(q_1)}{\partial q_1} & \cdots & \frac{\partial \log(q_1)}{\partial q_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \log(q_n)}{q_1} & \ddots & \frac{\partial \log(q_n)}{\partial q_n} \end{bmatrix} \\
&= -\mathbf{p}^T\begin{bmatrix} \frac{1}{\partial q_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ddots & \frac{1}{\partial q_n} \end{bmatrix} \\
&= -\begin{bmatrix} \frac{p_1}{q_1} & \cdots & \frac{p_n}{q_n} \end{bmatrix}
\end{aligned}$$

**c)** We want $\nabla_{\mathbf{W}}\mathbf{f}$, where $\mathbf{f} = \mathbf{W}\mathbf{x}$ with $\mathbf{W} \in \mathbb{R}^{2\times 3}$ and $\mathbf{x} \in \mathbb{R}^3$. According to example 5.11 of the textbook MML the dimension of the gradient $\nabla_{\mathbf{W}}\mathbf{f}$ will be $\mathbb{R}^{2\times(2\times 3)}$. Thus we have

(6) $\nabla_{\mathbf{W}}\mathbf{f} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{W}} \\ \frac{\partial f_2}{\partial \mathbf{W}} \end{bmatrix} = \begin{pmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{W}_1} \\ \frac{\partial f_1}{\partial \mathbf{W}_2} \end{bmatrix} \\ \begin{bmatrix} \frac{\partial f_2}{\partial \mathbf{W}_1} \\ \frac{\partial f_2}{\partial \mathbf{W}_1} \end{bmatrix} \end{pmatrix}$ with $f_i = \sum_{j=1}^{3} W_{ij}x_j \quad i = 1, 2$

and for the partial derivatives we get:

$$\frac{\partial f_i}{\partial W_{iq}} = x_q \tag{7}$$

This results in:

$$\frac{\partial f_i}{\partial \mathbf{W}_i} = \mathbf{x}^T \in \mathbb{R}^{1\times 1\times 3} \text{ and } \frac{\partial f_i}{\partial \mathbf{W}_{j\neq i}} = \mathbf{0}^T \in \mathbb{R}^{1\times 1\times 3} \tag{8}$$

From the result in (8) together with (6) we can construct the gradient:

$$\nabla_{\mathbf{W}}\mathbf{f} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{W}} \\ \frac{\partial f_2}{\partial \mathbf{W}} \end{bmatrix} = \left( \begin{bmatrix} \mathbf{x}^T \\ 0^T \end{bmatrix} \begin{bmatrix} \mathbf{0}^T \\ \mathbf{x}^T \end{bmatrix} \right)$$

**d)** The derivation of the gradient is as follows:

$$\begin{aligned}
\nabla_{\mathbf{W}}\mathbf{f} &= \nabla_{\mathbf{W}}\big((\boldsymbol{\mu} - \mathbf{W}\mathbf{x})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{W}\mathbf{x})\big) \\
&= \nabla_{\mathbf{W}}\big((\boldsymbol{\mu}^T - \mathbf{x}^T\mathbf{W}^T)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{W}\mathbf{x})\big) \\
&= \nabla_{\mathbf{W}}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{x} - \mathbf{x}^T\mathbf{W}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{x}^T\mathbf{W}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{x}) \\
&= 0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{x}^T - \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{x}^T + \boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{x}\mathbf{x}^T \\
&= -2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{W}\mathbf{x})\mathbf{x}^T
\end{aligned}$$

# 3 Probability Theory

## 3.1 Question 3.1

**a)** Based on an individual's experience with criminal activities and the given circumstances (observation) it seems more likely (higher probability) that the man is a criminal.

**b)** Define the following variables:

$$C = \begin{cases} 1, & \text{if } \textit{criminal} \\ 0, & \textit{otherwise} \end{cases}$$

$$O = \begin{cases} 1, & \text{if } \textit{observation is made} \\ 0, & \textit{otherwise} \end{cases}$$

and thus the probability, based on our beliefs of making the observation, that the man is a criminal can be formulated as follows:

$$p(C = 1 \mid O = 0) = \frac{p(O = 0 \mid C = 1)p(C = 1)}{p(O = 0)} \tag{9}$$

$$p(C = 1 \mid O = 1) = \frac{p(O = 1 \mid C = 1)p(C = 1)}{p(O = 1)} \tag{10}$$

**c)** We have the following probabilities:

$$\begin{aligned}
p(C = 1) &= \frac{1}{10^5} \\
p(C = 0) &= 1 - \frac{1}{10^5} \\
p(O = 1 \mid C = 0) &= \frac{1}{10^6} \\
p(O = 1 \mid C = 1) &= 0.8
\end{aligned}$$

and we get the probability of the man being a criminal given the described observation using Bayes' rule (equation 10):

$$\begin{aligned} p(C=1 \mid O=1) &= \frac{p(O=1 \mid C=1)p(C=1)}{p(O=1)} \\ &= \frac{p(O=1 \mid C=1)p(C=1)}{p(O=1 \mid C=0)p(C=0) + p(O=1 \mid C=1)p(C=1)} \quad \text{(product rule)} \\ &= \frac{0.8 \cdot \frac{1}{10^5}}{\frac{1}{10^6} \cdot (1 - \frac{1}{10^5}) + 0.8 \cdot \frac{1}{10^5}} = \frac{8}{9} \approx 0.89 \end{aligned}$$

**d)** Given the new information it seems more reasonable to assume that, for example, the shopowner entered to protect its belongings. Hence, the observation will probably made more often than before, and thus increasing $p(O=1)$. This in turn decreases the belief the man is a criminal when the observation is made ($p(C=1 \mid O=1)$)

## 3.2 Question 3.2

**a)** From Bishop 2.29 we get the following likelihood:

$$p(D \mid \rho) = \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_k^{x_{nk}} = \prod_{k=1}^{K} \rho_k^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{4} \rho_k^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{4} \rho_k^{m_k}$$

with $m_k = \sum_n x_{nk}$

**b)** From Bishop 2.33 and with $m_k$ the same as in exercise a) we get the following maximum likelihood estimator:

$$\rho_k^{ML} = \frac{m_k}{N}$$

and thus we have the following maximum likelihood solution:
$\rho_{ML} = \left[\frac{4}{8}, \quad 0, \quad 0, \quad \frac{4}{8}\right] = \left[\frac{1}{2}, \quad 0, \quad 0, \quad \frac{1}{2}\right]$

**c)** From Bishop 2.5 we get the following likelihood:

$$p(D \mid \rho) = \prod_{n=1}^{N} \rho^{x_n} (1-\rho)^{1-x_n}$$

**d)** From Bishop 2.7 we get the following maximunm likelihood estimation:

$$\rho_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{6}{8} = 0.75$$

**e)** The probability of drawing a red card is equal to the sumation of the probability to draw a card with suit heart or diamonds: $\rho = \rho_1 + \rho_3$

**f)** The Bayes rule is given by:

$$\overbrace{p(\rho \mid D)}^{posterior} = \frac{\overbrace{p(D \mid \rho)}^{likelihood} \overbrace{p(\rho)}^{prior}}{\underbrace{p(D)}_{evidence}}$$

**g)** The derivation of the MAP estimator is as follows:

$$
\begin{aligned}
p(\rho \mid D) &= \frac{p(D \mid \rho)p(\rho)}{p(D)} \\
&\propto p(D \mid \rho)p(\rho) \quad \text{(as } p(D) \text{ does not depend on } \rho\text{)} \\
&= \prod_{n=1}^{N} \rho^{x_n}(1-\rho)^{1-x_n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \rho^{\alpha-1}(1-\rho)^{\beta-1} \\
&\propto \prod_{n=1}^{N} \rho^{x_n}(1-\rho)^{1-x_n} \rho^{\alpha-1}(1-\rho)^{\beta-1} \quad \text{(as the } \Gamma() \text{ functions do not depend on } \rho\text{)} \\
&= \rho^{\sum_{n=1}^{N} x_n}(1-\rho)^{N-\sum_{n=1}^{N} x_n} \rho^{\alpha-1}\rho^{\beta-1} \\
&= \rho^{\sum_{n=1}^{N} x_n+\alpha-1}(1-\rho)^{N-\sum_{n=1}^{N} x_n+\beta-1} \\
&= \rho^{m+\alpha-1}(1-\rho)^{N-m+\beta-1}
\end{aligned}
$$

with $m = \sum_{n=1}^{N} x_n$. Subsequently we determine the derivative of the logarithm of the likelihood and set this equal to 0 (to find the optimal solution):

$$
\begin{aligned}
\frac{\partial \log(p(\rho \mid D))}{\partial \rho} &= \frac{\partial}{\partial \rho}\big((m+\alpha-1)\log(\rho) + (N-m+\beta-1)\log(1-\rho)\big) \\
&= \frac{m+\alpha-1}{\rho} - \frac{N-m+\beta-1}{1-\rho} = 0 \\
&\Leftrightarrow \frac{m+\alpha-1}{\rho} = \frac{N-m+\beta-1}{1-\rho} \implies m+\alpha-1 = N\rho+\alpha\rho+\beta\rho-2\rho \\
&\implies \rho_{MAP} = \frac{m+\alpha-1}{N+\alpha+\beta-2}
\end{aligned}
$$

As $m$ and $N$ are equal to each other this does not influence the parameter $\rho$. Thus to ensure an equal probability for drawing a red or black card one needs to solve the following equation:

$$
\rho = \frac{\alpha-1}{\alpha+\beta-2} = \frac{1}{2} \implies 2\alpha-2 = \alpha+\beta-2 \implies \alpha = \beta
$$