Homework set 1

Machine Learning 1

Yke Rusticus
11306386

September 12, 2019

2 Multivariate Calculus

Question 2.1

2.1 a)

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= \nabla_{\boldsymbol{\mu}}(\boldsymbol{x}^\top-\boldsymbol{\mu}^\top)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) && (1)\\
&= \nabla_{\boldsymbol{\mu}}[\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x}-\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{x}+\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}] && (2)\\
&= \nabla_{\boldsymbol{\mu}}[-\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{x}+\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}] && (3)\\
&= \nabla_{\boldsymbol{\mu}}[-\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}-(\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{x})^\top+\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}] && (4)\\
&= \nabla_{\boldsymbol{\mu}}[-2\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}] && (5)\\
&= \nabla_{\boldsymbol{\mu}}(-2\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu})+\nabla_{\boldsymbol{\mu}}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu} && (6)\\
&= -2\boldsymbol{x}^\top\Sigma^{-1}+\boldsymbol{\mu}^\top(\Sigma^{-1}+(\Sigma^{-1})^\top) && (7)\\
&= -2\boldsymbol{x}^\top\Sigma^{-1}+2\boldsymbol{\mu}^\top\Sigma^{-1} && (8)\\
&= 2(\boldsymbol{\mu}-\boldsymbol{x})^\top\Sigma^{-1} && (9)
\end{aligned}$$

Since $\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x}$ does not depend on $\boldsymbol{\mu}$, it was left out of Eq. (3). In Eq. (5) and (8) we used $(\Sigma^{-1})^\top = \Sigma^{-1}$. The identities $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{a}^\top\boldsymbol{x} = \boldsymbol{a}^\top$ and $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{B}+\boldsymbol{B}^\top)$, as given in section 5.5 of "Mathematics for Machine Learning", are used in Eq. (7).

2.1 b) Assuming $\log(\cdot) = \ln(\cdot)$ (not clearly specified in the exercise):

$$\nabla_{\boldsymbol{q}} - \boldsymbol{p}^\top \log \boldsymbol{q} = -\boldsymbol{p}^\top \nabla_{\boldsymbol{q}} \log \boldsymbol{q} \tag{10}$$

$$\left[\frac{\partial \log \boldsymbol{q}}{\partial \boldsymbol{q}}\right]_{i,j} = \frac{\partial \log q_i}{\partial q_j} = \frac{\delta_{i,j}}{q_i}, \tag{11}$$

where δ is the Kronecker delta. We then get the answer:

$$\nabla_{\boldsymbol{q}} - \boldsymbol{p}^\top \log \boldsymbol{q} = -\boldsymbol{p}^\top \begin{bmatrix} 1/q_1 & 0 & \cdots & 0 \\ 0 & 1/q_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1/q_N \end{bmatrix}, \tag{12}$$

for $N \in \mathbb{N}$.

2.1 c) $\boldsymbol{W} \in \mathbb{R}^{2\times 3}$ and $\boldsymbol{x} \in \mathbb{R}^3$, so $\boldsymbol{f} \in \mathbb{R}^2$.
This means we should get a final answer $\frac{\partial}{\partial \boldsymbol{W}}\boldsymbol{f} \in \mathbb{R}^{2\times(3\times 2)}$.

$$\frac{\partial}{\partial \boldsymbol{W}}\boldsymbol{f} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{W}} f_1 \\ \frac{\partial}{\partial \boldsymbol{W}} f_2 \end{bmatrix} \tag{13}$$

$$\left[\frac{\partial f_i}{\partial \boldsymbol{W}}\right]_{1,j,k} = \frac{\partial f_i}{\partial W_{k,j}} \tag{14}$$

$$= \frac{\partial}{\partial W_{k,j}} W_{i,1}x_1 + W_{i,2}x_2 + W_{i,3}x_3. \tag{15}$$

Working this out gives a matrix of which the i'th column is equal to $\boldsymbol{x}$, and the rest of the elements are zero. For our final answer we thus get:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{W}} = \begin{bmatrix} \begin{bmatrix} x_1 & 0 \\ x_2 & 0 \\ x_3 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & x_1 \\ 0 & x_2 \\ 0 & x_3 \end{bmatrix} \end{bmatrix} \tag{16}$$

2.1 d)

$$f = (\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x})^\top \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x}) \tag{17}$$

$$= (\boldsymbol{\mu} - \boldsymbol{y})^\top \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{y}), \tag{18}$$

where $\boldsymbol{y} := \boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^M$. We can then apply the chain rule:

$$\frac{\partial f}{\partial \boldsymbol{W}} = \frac{\partial f}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}} \tag{19}$$

From similarity with exercise 2.1a we can see that

$$\frac{\partial f}{\partial \boldsymbol{y}} = 2(\boldsymbol{y} - \boldsymbol{\mu})^\top \Sigma^{-1}. \tag{20}$$

The last part is given by solving $\frac{\partial}{\partial \boldsymbol{W}}\boldsymbol{y}$, which is similar to exercise 2.1e. However now, instead of fixed matrix dimensions we have

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}} \in \mathbb{R}^{M\times(K\times M)}, \tag{21}$$

so we get

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}} = \begin{bmatrix} \begin{bmatrix} x_1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ x_K & 0 & \cdots & \cdots & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & x_K & 0 & \cdots & 0 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} 0 & \cdots & \cdots & 0 & x_1 \\ \vdots & & & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & x_K \end{bmatrix} \end{bmatrix}, \tag{22}$$

with M matrices on the vertical, each consisting of K rows and M columns. Combining Eqs. (20) and (22) gives us

$$\frac{\partial f}{\partial \boldsymbol{W}} = 2(\boldsymbol{W}\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} \begin{bmatrix} \begin{bmatrix} \boldsymbol{x} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \end{bmatrix} \\ \begin{bmatrix} \boldsymbol{0} & \boldsymbol{x} & \boldsymbol{0} & \cdots & \boldsymbol{0} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} & \boldsymbol{x} \end{bmatrix} \end{bmatrix} \in \mathbb{R}^{1\times(K\times M)}, \tag{23}$$

in which $\boldsymbol{x} = [x_1, x_2, ..., x_K]^\top$ and $\boldsymbol{0} = [0, ..., 0]^\top \in \mathbb{R}^M$. (Sorry, I was struggling to get the matrices to align properly.)

3 Probability Theory

Question 3.1

3.1 a) Most of the times when people experience a man climbing through a broken window a jewelry store with a bag over his shoulder, either on television or in stories, the man in question is a criminal. We learn from experience, so a man with this exact description would be classified by many people as a criminal.

3.1 b) Say, c stands for "the man being a criminal" and o stands for the described observation. The chance of the man being a criminal given our observation is given by Bayes' theorem:

$$p(c|o) = \frac{p(o|c)p(c)}{p(o)} \tag{24}$$

3.1 c)

$$\begin{align} p(c) &= 10^{-5} \tag{25}\\ p(o|c) &= 0.8 \tag{26}\\ p(o|\neg c) &= 10^{-6} \tag{27}\\ p(o) &= p(o|c)p(c) + p(o|\neg c)p(\neg c) \tag{28}\\ p(\neg c) &= 1 - p(c) \tag{29} \end{align}$$

So, by combining Eqs. (24), (28) and (29) and using the given values gives:

$$\begin{align} p(c|o) &= \frac{p(o|c)p(c)}{p(o|c)p(c) + p(o|\neg c)(1 - p(c))} \tag{30}\\ &= \left(1 + \frac{p(o|\neg c)(1 - p(c))}{p(o|c)p(c)}\right)^{-1} \tag{31}\\ &= 0.889 \tag{32}\\ &\approx \frac{8}{9} \tag{33} \end{align}$$

3.1 d) If some kids smashed multiple storefronts in the neighbourhood, then the man in question might be the jewelry store owner bringing his belongings to safety. In other words,

$p(o|\neg c)$ increases, although it is difficult to say how much. Say, it increases by a factor 10, i.e. $p(o|\neg c) = 10^{-5}$. Then Eq. (30) evaluates to $p(c|o) = 0.444 \approx \frac{4}{9}$. So in this case we would believe the man is (most probably) not a criminal, given the observation.

Question 3.2

3.2 a) The n'th observation is $\boldsymbol{x}_n \in \mathbb{R}^4$. $\mathcal{D}$ is the data set consisting of N independent observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, i.e. $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$. The likelihood function for the observed data set is (Bishop Eq. 2.29):

$$p(\mathcal{D}|\boldsymbol{\rho}) = \prod_{k=1}^{4} \rho_k^{m_k}, \tag{34}$$

where m_k is given by

$$m_k = \sum_{n=1}^{N} x_{n,k} \tag{35}$$

3.2 b) Maximum likelihood solution for $\boldsymbol{\rho}$:

$$\boldsymbol{\rho} = [1/2, 0, 0, 1/2] \tag{36}$$

This is based on the fact that half of the draws are hearts (i.e. $\rho_1 = 1$) and the other half are spades (i.e. $\rho_4 = 1$).

3.2 c) Define $r_n := x_{n,1} + x_{n,3}$, then

$$\text{if} \quad r_n = 1 \rightarrow \text{the } n\text{'th observation is a red card} \tag{37}$$
$$\text{if} \quad r_n = 0 \rightarrow \text{the } n\text{'th observation is a black card} \tag{38}$$

To avoid confusion, a capital P will now be used in $P(X)$ to denote the probability of an event X, instead of $p(X)$. The likelihood function for the card colors is (Bishop Eq. 2.5):

$$P(\mathcal{D}|p) = \prod_{n=1}^{N} P(r_n|p) = \prod_{n=1}^{N} p^{r_n}(1-p)^{1-r_n} \tag{39}$$

3.2 d) Using Bishop Eq. 2.7:

$$p_{ML} = \frac{1}{N}\sum_{n=1}^{N} r_n = \frac{6}{8} = 0.75 \tag{40}$$

3.2 e)

$$p = \rho_1 + \rho_3 \tag{41}$$

3.2 f)

$$\underbrace{P(p|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{P(\mathcal{D}|p)}^{\text{likelihood}}\,\overbrace{P(p)}^{\text{prior}}}{\underbrace{P(\mathcal{D})}_{\text{evidence}}} \tag{42}$$

3.2 g) We have

$$P(\mathcal{D}|p) = \prod_{n=1}^{N} p^{r_n}(1-p)^{1-r_n} \tag{43}$$

$$P(p) = \text{Beta}(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{44}$$

The log likelihood is

$$\mathcal{L} = \log(P(p|\mathcal{D})) \tag{45}$$

The MAP estimate p_{MAP} is given by

$$
\begin{aligned}
p_{\text{MAP}} &= \arg\max_{p} \mathcal{L} &(46)\\
&= \arg\max_{p} \log(P(\mathcal{D}|p)(p)) &(47)\\
&= \arg\max_{p} \underbrace{\sum_{n=1}^{N} \log(p^{r_n}(1-p)^{1-r_n})}_{:=A} + \underbrace{\log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1})}_{:=B} &(48)
\end{aligned}
$$

To find p_{MAP}, we need to set $\frac{\partial}{\partial p}(A+B) = \frac{\partial A}{\partial p} + \frac{\partial B}{\partial p} = 0$. For term A we have:

$$
\begin{aligned}
\frac{\partial A}{\partial p} &= \sum_{n=1}^{N} \frac{\partial}{\partial p} \log(p^{r_n}(1-p)^{1-r_n}) &(49)\\
y &:= p^{r_n}(1-p)^{1-r_n} &(50)\\
\frac{\partial \log y}{\partial p} &= \frac{\partial \log y}{\partial y}\frac{\partial y}{\partial p} &(51)\\
&= \frac{1}{y}[r_n p^{r_n-1}(1-p)^{1-r_n} + p^{r_n}(r-1)(1-p)^{-r_n}] &(52)\\
\frac{\partial A}{\partial p} &= \sum_{n=1}^{N} \frac{1}{p^{r_n}(1-p)^{1-r_n}}[r_n p^{r_n} p^{-1}(1-p)^{1-r_n} + p^{r_n}(r-1)\frac{(1-p)^{1-r_n}}{1-p}] &(53)\\
&= \sum_{n=1}^{N}\left[\frac{r_n}{p} + \frac{r_n-1}{1-p}\right] &(54)\\
&= \frac{1}{p}\sum_{n=1}^{N} r_n + \frac{1}{1-p}\sum_{n=1}^{N}(r_n-1) &(55)
\end{aligned}
$$

For term B we have:

$$
\begin{aligned}
\frac{\partial B}{\partial p} &= \frac{\partial}{\partial p}\log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}) + \frac{\partial}{\partial p}\log(p^{\alpha-1}(1-p)^{\beta-1}) &(56)\\
&= \frac{\partial}{\partial p}\log(p^{\alpha-1}(1-p)^{\beta-1}) &(57)\\
&= \frac{1}{p^{\alpha-1}(1-p)^{\beta-1}}[(\alpha-1)p^{\alpha-2}(1-p)^{\beta-1} + p^{\alpha-1}(\beta-1)(1-p)^{\beta-2}] &(58)\\
&= \frac{(\alpha-1)p^{\alpha-1}p^{-1}(1-p)^{\beta-1}}{p^{\alpha-1}(1-p)^{\beta-1}} + \frac{p^{\alpha-1}(1-\beta)(1-p)^{\beta-1}(1-p)^{-1}}{p^{\alpha-1}(1-p)^{\beta-1}} &(59)\\
&= \frac{\alpha-1}{p} + \frac{1-\beta}{1-p} &(60)
\end{aligned}
$$

Setting $\frac{\partial A}{\partial p} + \frac{\partial B}{\partial p} = 0$ gives

$$\begin{aligned} 0 &= \frac{1}{p_{\mathrm{MAP}}} \sum_{n=1}^{N} r_n + \frac{1}{1 - p_{\mathrm{MAP}}} \sum_{n=1}^{N} (r_n - 1) + \frac{\alpha - 1}{p_{\mathrm{MAP}}} + \frac{1 - \beta}{1 - p_{\mathrm{MAP}}} && (61) \\ &= \sum_{n=1}^{N} r_n + \frac{p_{\mathrm{MAP}}}{1 - p_{\mathrm{MAP}}} \sum_{n=1}^{N} (r_n - 1) + \alpha - 1 + \frac{p_{\mathrm{MAP}}(1 - \beta)}{1 - p_{\mathrm{MAP}}} && (62) \end{aligned}$$

$$\begin{aligned} p_{\mathrm{MAP}} \Big[\sum_{n=1}^{N} (r_n - 1) + 1 - \beta \Big] &= (1 - p_{\mathrm{MAP}}) \Big[1 - \alpha - \sum_{n=1}^{N} r_n \Big] && (63) \\ &= \Big[1 - \alpha - \sum_{n=1}^{N} r_n \Big] - p_{\mathrm{MAP}} \Big[1 - \alpha - \sum_{n=1}^{N} r_n \Big] && (64) \end{aligned}$$

$$\begin{aligned} p_{\mathrm{MAP}} \Big[\sum_{n=1}^{N} (r_n - 1) - \sum_{n=1}^{N} r_n + 2 - \beta - \alpha \Big] &= 1 - \alpha - \sum_{n=1}^{N} r_n && (65) \\ p_{\mathrm{MAP}} = \frac{1 - \alpha - \sum_{n=1}^{N} r_n}{2 - N - \beta - \alpha} &= \frac{\sum_{n=1}^{N} r_n + \alpha - 1}{N + \alpha + \beta - 2} && (66) \end{aligned}$$

To encode the belief that there is an equal probability of drawing a red or black card, we would choose $\alpha = \beta$. The greater their values, the stronger the belief that $p \approx 0.5$.