Andrii Skliar
11636785
andrii.skliar@student.uva.nl

Homework Assignment 1
Machine Learning 1, 17/18

2019-01-08

1 Basic Linear Algebra and Derivatives

Question 1.1

Let A = $\begin{bmatrix} -7 & 8 & 1 \\ -4 & 3 & 5 \\ 7 & 7 & -8 \end{bmatrix}$ and b = $\begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$

1. Compute $\mathbf{A}\mathbf{b}$

Ab = $\begin{bmatrix} -7 & 8 & 1 \\ -4 & 3 & 5 \\ 7 & 7 & -8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} -7+16+5 \\ -4+6+25 \\ 7+14-40 \end{bmatrix} = \begin{bmatrix} 14 \\ 27 \\ -19 \end{bmatrix}$

2. Compute $\mathbf{b}^{\mathbf{T}}\mathbf{A}$

$\mathbf{b}^{\mathbf{T}}$ = $\begin{bmatrix} 1 & 2 & 5 \end{bmatrix}$

$$\mathbf{b}^{\mathbf{T}}A = \begin{bmatrix} 1 & 2 & 5 \end{bmatrix} \begin{bmatrix} -7 & 8 & 1 \\ -4 & 3 & 5 \\ 7 & 7 & -8 \end{bmatrix} = \begin{bmatrix} -7-8+35 & 8+6+35 & 1+10-40 \end{bmatrix} = \begin{bmatrix} 20 & 49 & -29 \end{bmatrix}$$

3. Compute the vector $\mathbf{c}$ for which $\mathbf{A}\mathbf{c} = \mathbf{b}$ through elimination.

To calculate vector c we are going to use gaussian elimination on augmented matrix.
Note: L_i means i-th row.

$$\left[\begin{array}{ccc|c} -7 & 8 & 1 & 1 \\ -4 & 3 & 5 & 2 \\ 7 & 7 & -8 & 5 \end{array}\right] \rightarrow \left(L_1 \cdot \tfrac{1}{-7} \rightarrow L_1 \right) \rightarrow \left[\begin{array}{ccc|c} 1 & -\frac{8}{7} & -\frac{1}{7} & -\frac{1}{7} \\ -4 & 3 & 5 & 2 \\ 7 & 7 & -8 & 5 \end{array}\right] \rightarrow \begin{pmatrix} L_2 + 4 \cdot L_1 \rightarrow L_2 \\ L_3 - 7 \cdot L_1 \rightarrow L_3 \end{pmatrix} \rightarrow$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & -\frac{8}{7} & -\frac{1}{7} & -\frac{1}{7} \\ 0 & -\frac{11}{7} & \frac{31}{7} & \frac{10}{7} \\ 0 & 15 & -7 & 6 \end{array}\right] \rightarrow \left(L_2 \cdot -\tfrac{11}{7} \rightarrow L_2 \right) \rightarrow \left[\begin{array}{ccc|c} 1 & -\frac{8}{7} & -\frac{1}{7} & -\frac{1}{7} \\ 0 & 1 & -\frac{31}{11} & -\frac{10}{11} \\ 0 & 15 & -7 & 6 \end{array}\right] \rightarrow \begin{pmatrix} L_1 + \frac{8}{7} \cdot L_2 \rightarrow L_1 \\ L_3 - 15 \cdot L_2 \rightarrow L_3 \end{pmatrix} \rightarrow$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -\frac{37}{11} & -\frac{13}{11} \\ 0 & 1 & -\frac{31}{11} & -\frac{10}{11} \\ 0 & 0 & \frac{388}{11} & \frac{216}{11} \end{array}\right] \rightarrow \left(L_3 \cdot \tfrac{388}{11} \rightarrow L_3 \right) \rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -\frac{37}{11} & -\frac{13}{11} \\ 0 & 1 & -\frac{31}{11} & -\frac{10}{11} \\ 0 & 0 & 1 & \frac{54}{97} \end{array}\right] \rightarrow \begin{pmatrix} L_1 + \frac{37}{11} \cdot L_3 \rightarrow L_1 \\ L_2 + \frac{31}{11} \cdot L_3 \rightarrow L_2 \end{pmatrix} \rightarrow$$

$$\rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & \frac{67}{97} \\ 0 & 1 & 0 & \frac{64}{97} \\ 0 & 0 & 1 & \frac{54}{97} \end{array}\right]$$

$$c = \begin{bmatrix} \frac{67}{97} \\ \frac{64}{97} \\ \frac{54}{97} \end{bmatrix}$$

4. Compute the inverse of A denoted by $\mathbf{A^{-1}}$.

$$\det(A) = \begin{vmatrix} -7 & 8 & 1 \\ -4 & 3 & 5 \\ 7 & 7 & -8 \end{vmatrix} = -7 \cdot (-24 - 35) - 8 \cdot (32 - 35) + 1 \cdot (-28 - 21) = 7 \cdot 59 + 24 - 49 = 388$$

$$A^{-1} = \frac{1}{\det(A)} * \begin{bmatrix} C_{11} & -C_{12} & C_{13} \\ -C_{21} & C_{22} & -C_{23} \\ C_{31} & -C_{32} & C_{33} \end{bmatrix}^T = \frac{1}{388} \cdot \begin{bmatrix} -24-35 & -(32-35) & -28-21 \\ -(-64-7) & 56-7 & -(-49-56) \\ 40-3 & -(-35+4) & -21+32 \end{bmatrix}^T =$$

$$= \frac{1}{388} \cdot \begin{bmatrix} -59 & 3 & -49 \\ 71 & 49 & 105 \\ 37 & 31 & 11 \end{bmatrix}^T = \frac{1}{388} \cdot \begin{bmatrix} -59 & 71 & 37 \\ 3 & 49 & 31 \\ -49 & 105 & 11 \end{bmatrix} = \begin{bmatrix} -\frac{59}{388} & \frac{71}{388} & \frac{37}{388} \\ \frac{3}{388} & \frac{49}{388} & \frac{31}{388} \\ -\frac{49}{388} & \frac{105}{388} & \frac{11}{388} \end{bmatrix}$$

5. Verify that in this case $\mathbf{A^{-1}b = c}$. Show that this must be the case for general $\mathbf{A}, \mathbf{b}$ and $\mathbf{c}$.

$$A^{-1}b = \begin{bmatrix} -\frac{59}{388} & \frac{71}{388} & \frac{37}{388} \\ \frac{3}{388} & \frac{39}{388} & \frac{31}{388} \\ -\frac{49}{105} & \frac{105}{388} & \frac{11}{388} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} = \frac{1}{388} \begin{bmatrix} -59 + 142 + 37 \cdot 5 \\ 3 + 98 + 155 \\ -49 + 210 + 55 \end{bmatrix} = \begin{bmatrix} \frac{268}{388} \\ \frac{256}{388} \\ \frac{216}{388} \end{bmatrix} = \begin{bmatrix} \frac{67}{97} \\ \frac{64}{97} \\ \frac{54}{97} \end{bmatrix}$$

General proof:
Note: here $b = Ac$ as implied from previous question.
$A^{-1} \cdot b = A^{-1} \cdot A \cdot c = I \cdot c = c$, where I is Identity matrix

Question 1.2

Find the derivative of the following functions with respect to x.

1. $f(x) = (\frac{2}{x^2} + x^{-7} + x^3)^2$

$$f'(x) = 2 \cdot (\frac{2}{x^2} + x^{-7} + x^3)(-\frac{4}{x^3} - \frac{7}{x^8} + 3x^2) = 2(-8x^{-5} - 14x^{-10} + 6 - 4x^{-10} - 7x^{-15} + 3x^{-5} - 4 - 7x^{-5} + 3x^5)$$
$$= 2 \cdot (2 - 12x^{-5} - 18x^{-10} - 7x^{-15} + 3x^5) = 6x^5 - 24x^{-5} - 36x^{-10} - 14x^{-15} + 4$$

2. $f(x) = x^2 \sqrt{e^{-\sqrt[3]{x}}}$

$$f'(x) = 2xe^{-\frac{1}{2}\sqrt[3]{x}} + x^2 e^{-\frac{1}{2}\sqrt[3]{x}}(-\frac{1}{2} \cdot \frac{1}{3}x^{-\frac{2}{3}}) = 2xe^{-\frac{1}{2}x^{\frac{1}{3}}} - \frac{1}{6}x^{\frac{4}{3}}e^{-\frac{1}{2}x^{\frac{1}{3}}} = (2x - \frac{1}{6}x^{\frac{4}{3}})e^{-\frac{1}{2}x^{\frac{1}{3}}}$$

3. $f(x) = x + \ln x$

$$f'(x) = 1 + \frac{1}{x} = \frac{x+1}{x}$$

4. $f(x) = x \cdot \ln \sqrt{x}$

$$f'(x) = \ln \sqrt{x} + \frac{1}{2}x \cdot x^{-\frac{1}{2}} \cdot \frac{1}{\sqrt{x}} = \ln \sqrt{x} + \frac{1}{2}\sqrt{x} \cdot \frac{1}{\sqrt{x}} = \ln \sqrt{x} + \frac{1}{2} = \frac{1}{2}\ln x + \frac{1}{2} = \frac{1}{2}(\ln x + 1)$$

5. $f(x) = 6(x^2 - 1)\sin x$

$$f'(x) = 12x \cdot sinx + 6(x^2 - 1)\cos x$$

6. $f(x) = \ln\left(\sqrt[3]{\frac{e^{3x}}{1+e^{3x}}}\right)$

$$
\begin{aligned}
1.f_1(x) &= \frac{e^{3x}}{1+e^{3x}} \\
f_1'(x) &= \frac{3e^{3x}(1+e^{3x}) - 3e^{3x}\cdot e^{3x}}{(1+e^{3x})^2} = \frac{3e^{3x}+3e^{6x}-3e^{6x}}{(1+e^{3x})^2} = \frac{3e^{3x}}{(1+e^{3x})^2} \\
2.f_2(x) &= \sqrt[3]{f_1(x)} \\
f_2'(x) &= \frac{3e^{3x}}{(1+e^{3x})^2}\cdot\frac{1}{3}\left(\frac{e^{3x}}{1+e^{3x}}\right)^{-\frac{2}{3}} = \frac{e^{3x}}{(1+e^{3x})^2}\cdot\left(\frac{e^{3x}}{1+e^{3x}}\right)^{-\frac{2}{3}} \\
3.f(x) &= \ln f_2(x) \\
f'(x) &= \frac{e^{3x}}{(1+e^{3x})^2}\cdot\left(\frac{e^{3x}}{1+e^{3x}}\right)^{-\frac{2}{3}}\cdot\left(\frac{e^{3x}}{1+e^{3x}}\right)^{-\frac{1}{3}} = \frac{e^{3x}}{(1+e^{3x})^2}\cdot\frac{(1+e^{3x})}{e^{3x}} = \frac{1}{1+e^{3x}} = \frac{1}{e^{3x}+1}
\end{aligned}
$$

$f'(x) = \frac{1}{e^{3x}+1}$

7. What does it mean to build the partial derivative of a multivariate function?

Building partial derivative of a multivariate function $f(x_1, x_2, \dots x_n)$ with respect to one of the variables means to find derivative ($\frac{\partial f}{\partial x_i}$) by one of the parameters while keeping all the other parameters constant. The reason to use it is to see how the function will change if we keep all the variables fixed except for desired one. In practice, that has a lot of applications, i.e. finding gradient (direction of the fastest change of a function) allows to either minimize or maximize function and thus is being used, among other applications, in gradient descent algorithm.

8.$f(x,y,z) = 2\ln\left(y - \exp\left(x^{-1}\right) - \sin\left(z\cdot x^2\right)\right)$

$$
\begin{aligned}
\frac{\partial f}{\partial x} &= 2\cdot\frac{1}{y-\exp(x^{-1})-\sin(z\cdot x^2)}\cdot\frac{\partial(y-\exp\left(x^{-1}\right)-\sin\left(z\cdot x^2\right))}{\partial x} = 2\cdot\frac{x^{-2}\cdot exp(x^{-1}) - 2z\cdot x\cdot cos(z\cdot x^2)}{y-\exp(x^{-1})-\sin(z\cdot x^2)} \\
\frac{\partial f}{\partial y} &= 2\cdot\frac{1}{y-\exp(x^{-1})-\sin(z\cdot x^2)}\cdot\frac{\partial(y-\exp\left(x^{-1}\right)-\sin\left(z\cdot x^2\right))}{\partial y} = \frac{2}{y-\exp(x^{-1})-\sin(z\cdot x^2)} \\
\frac{\partial f}{\partial z} &= 2\cdot\frac{1}{y-\exp(x^{-1})-\sin(z\cdot^2)}\cdot\frac{\partial(y-\exp\left(x^{-1}\right)-\sin\left(z\cdot x^2\right))}{\partial z} = \frac{-2x^2\cos(z\cdot x^2)}{y-\exp(x^{-1})-\sin(z\cdot^2)}
\end{aligned}
$$

9. $f(x,y,z) = \ln(\sqrt[\gamma]{z^\alpha y^\beta x^\gamma})$

First, we should simplify a bit:

$f(x,y,z) = \ln(\sqrt[\gamma]{z^\alpha y^\beta x^\gamma}) = \frac{1}{\gamma}\cdot\ln(z^\alpha y^\beta x^\gamma)$

Now, let's take derivative step by step:

$$
\begin{aligned}
1.f_1(x,y,z) &= z^\alpha y^\beta x^\gamma \\
\frac{\partial f_1}{\partial x} &= \gamma z^\alpha y^\beta x^{\gamma-1} \\
\frac{\partial f_1}{\partial y} &= \beta z^\alpha y^{\beta-1} x^\gamma \\
\frac{\partial f_1}{\partial z} &= \alpha z^{\alpha-1} y^\beta x^\gamma \\
2.f(x,y,z) &= \frac{1}{\gamma}\ln(f_1(x,y,z)) \\
\frac{\partial f}{\partial x} &= \frac{1}{\gamma}\frac{\gamma z^\alpha y^\beta x^{\gamma-1}}{z^\alpha y^\beta x^\gamma} = x^{-1} = \frac{1}{x} \\
\frac{\partial f}{\partial y} &= \frac{1}{\gamma}\frac{\beta z^\alpha y^{\beta-1} x^\gamma}{z^\alpha y^\beta x^\gamma} = \frac{\beta}{\gamma y} \\
\frac{\partial f}{\partial z} &= \frac{1}{\gamma}\frac{\alpha z^{\alpha-1} y^\beta x^\gamma}{z^\alpha y^\beta x^\gamma} = \frac{\alpha}{\gamma z}
\end{aligned}
$$

Question 1.3

The following questions are good practice in manipulating vectors and matrices and they are very important for solving for posterior distributions.
Given the following expression:

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathbf{T}}\mathbf{S}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$$

where $\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\mu}_0$ are vectors and $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}^{-1}$ are symmetric, invertible matrices.

1. Expand the expression and gather terms.

$$\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathbf{T}}\mathbf{S}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\\
&=(\mathbf{x}^{\mathbf{T}}-\boldsymbol{\mu}^{\mathbf{T}})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+(\boldsymbol{\mu}^{\mathbf{T}}-\boldsymbol{\mu}_0^{\mathbf{T}})\mathbf{S}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\\
&=\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0
\end{aligned}$$

2. Collect all the terms that depend on μ and those that do not.

$$\begin{aligned}
&\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0\\
&=(\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0)-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0\\
&=(\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0)+(\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu})
\end{aligned}$$

3. Take the derivative with respect to μ, set to 0, and solve for μ.

Note: The last step of calculation implies, that $(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})$ is invertible.

$$\begin{aligned}
&\mathbf{f}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\mu}_0,\boldsymbol{\Sigma},\mathbf{S})=(\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0)+(\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})\boldsymbol{\mu}-\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\boldsymbol{\mu}^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu}_0-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\boldsymbol{\mu})\\
&\frac{\partial \mathbf{f}}{\partial \boldsymbol{\mu}}=\mathbf{2}\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}-\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}-\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}=\mathbf{2}\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})-\mathbf{2}\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}-\mathbf{2}\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\\
&\frac{\partial \mathbf{f}}{\partial \boldsymbol{\mu}}=\mathbf{2}\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})-\mathbf{2}\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}-\mathbf{2}\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}=\mathbf{0}\\
&\mathbf{2}\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})=\mathbf{2}\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}+\mathbf{2}\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\\
&\boldsymbol{\mu}^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})=\frac{\mathbf{2}\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}+\mathbf{2}\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}}{\mathbf{2}}=\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1}\\
&\boldsymbol{\mu}^{\mathbf{T}}=(\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1})(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})^{-1}\\
&\boldsymbol{\mu}=((\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1})(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})^{-1})^{\mathbf{T}}=((\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})^{\mathbf{T}})^{-1}(\mathbf{x}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}+\boldsymbol{\mu}_0^{\mathbf{T}}\mathbf{S}^{-1})^{\mathbf{T}}=(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0)
\end{aligned}$$

It is also possible to check if our calculations were correct using matrix derivative applied to the initial expression $\mathbf{f}(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\mu}_0,\boldsymbol{\Sigma},\mathbf{S})=(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathbf{T}}\mathbf{S}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$:

$$\begin{aligned}
&\frac{\partial \mathbf{f}}{\partial \boldsymbol{\mu}}=-\mathbf{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+\mathbf{2}\mathbf{S}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)=\mathbf{2}(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})\boldsymbol{\mu}-\mathbf{2}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{2}\mathbf{S}^{-1}\boldsymbol{\mu}_0=\mathbf{0}\\
&\boldsymbol{\mu}=(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\boldsymbol{\mu}_0)
\end{aligned}$$

2 Probability Theory

Question 2.1

Being a student in the Netherlands, you spend all your time in the cities of Amsterdam and Rotterdam. Based on your experience, the weather in Amsterdam is much nicer: the probability that it rains when you are in Amsterdam is 0.5, while the probability that it rains when in Rotterdam is 0.75. Amsterdam is where you spend most of your time: at any given moment, the probability that you are in Amsterdam is 0.8 and the probability that you are in Rotterdam is 0.2.
Based on the above:

1. Define the random variables and the values they can take on, both with symbols and numerically.

Note: if x indicates an event, then $\overline{x}$ indicates the negation of an event.

- X is a random variable, taking on values from $\{rain, \overline{rain}\}$
- Y is a random variable, taking on values from $\{Amsterdam, Rotterdam\}$
- $\begin{cases} P(Y = Amsterdam) = 0.8 \\ P(Y = Rotterdam) = 0.2 \end{cases}$
- $\begin{cases} P(X = rain|Y = Amsterdam) = 0.5 \\ P(X = rain|Y = Rotterdam) = 0.75 \end{cases}$

2. What is the probability that it does not rain when you are in Rotterdam?

$$P(X = \overline{rain}|Y = Rotterdam) = 1 - P(X = rain|Y = Rotterdam) = 1 - 0.75 = 0.25$$

3. What is the probability that it rains at your current location?

$$\begin{aligned}
&P(X = rain) = \sum_{y \in \{Amsterdam, Rotterdam\}} P(X = rain, Y = y) \\
&= P(X = rain, Y = Amsterdam) + P(X = rain, Y = Rotterdam) \\
&= P(X = rain|Y = Amsterdam)P(Y = Amsterdam) + P(X = rain|Y = Rotterdam)P(Y = Rotterdam) \\
&= 0.5 \cdot 0.8 + 0.75 \cdot 0.2 = 0.4 + 0.15 = 0.55
\end{aligned}$$

4. You wake up on the sidewalk, after a night out which you can't remember anything about but which clearly was not such a great idea. You can't recognize your surroundings, but you must be either in Amsterdam or Rotterdam. It is raining. What is the probability that you are in Amsterdam?

To solve this problem, we have to apply Bayes rule:

$$P(Y = Amsterdam|X = rain) = \frac{P(X = rain|Y = Amsterdam)P(Y = Amsterdam)}{P(X = rain)} = \frac{0.5 \cdot 0.8}{0.55} = 0.727$$

Question 2.2

According to a 1988 study, home pregnancy tests conducted by experienced technicians have a false negative rate of only 2.6%, while when carried out by consumers the rate goes up to 25%, due to improper use or misunderstanding of the instructions. Generally, false positive rates are rare events (0.01%) independent of the conductor. With this information, given a group of 10 000 women that could be evenly likely pregnant or not, 2000 are being assisted by a professional (Group A), and 8000 conduct the test on their own (Group B).

1. Define the random variables and the values they can take on.

- X is a random variable, meaning if woman is **actually pregnant or not** and taking on values from $\{pregnant, \overline{pregnant}\}$.
- Y is a random variable, meaning if woman was **diagnosed pregnant or not** and taking on values from $\{diagnosed\ pregnant, \overline{diagnosed\ pregnant}\}$.
- Z is a random variable, meaning who was carrying out a test, taking on values from $\{technician, consumer\}$.
- False negative rate for technicians: $P(Y = \overline{diagnosed\ pregnant}|X = pregnant, Y = technician) = 0.026$
- False negative rate for consumers: $P(Y = \overline{diagnosed\ pregnant}|X = pregnant, Y = consumer) = 0.25$
- False positive rate for technicians: $P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Y = technician) = 0.0001$
- False positive rate for consumers: $P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Y = consumer) = 0.0001$
- $P(X = pregnant) = 0.5$
- $P(X = \overline{pregnant}) = 0.5$

2. How many women in each group do you expect being diagnosed as pregnant?

$$P(Y = diagnosed\ pregnant|Z = technician) = \sum_{x \in \{pregnant, \overline{pregnant}\}} P(X = x, Y = diagnosed\ pregnant|Z = technician)$$
$$= P(Y = diagnosed\ pregnant|X = pregnant, Z = technician) \cdot P(X = pregnant)$$
$$+ P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Z = technician) \cdot P(X = \overline{pregnant})$$
$$= \big(1 - P(Y = \overline{diagnosed\ pregnant}|X = pregnant, Z = technician)\big) \cdot P(X = pregnant)$$
$$+ P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Z = technician) \cdot P(X = \overline{pregnant})$$
$$= (1 - 0.026) \cdot 0.5 + 0.0001 \cdot 0.5 = 0.48705$$

$$P(Y = diagnosed\ pregnant|Z = customer) = \sum_{x \in \{pregnant, \overline{pregnant}\}} P(X = x, Y = diagnosed\ pregnant|Z = customer)$$
$$= P(Y = diagnosed\ pregnant|X = pregnant, Z = customer) \cdot P(X = pregnant)$$
$$+ P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Z = customer) \cdot P(X = \overline{pregnant})$$
$$= \big(1 - P(Y = \overline{diagnosed\ pregnant}|X = pregnant, Z = customer)\big) \cdot P(X = pregnant)$$
$$+ P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Z = customer) \cdot P(X = \overline{pregnant})$$
$$= (1 - 0.25) \cdot 0.5 + 0.0001 \cdot 0.5 = 0.37505$$

- For group A, expected number of patients being diagnosed as pregnant is:

$$\#A \cdot (Y = diagnosed\ pregnant|Z = technician) = 2000 \cdot 0.48705 = 974.1$$

- For group B, expected number of patients being diagnosed as pregnant is:

$$\#B \cdot (Y = diagnosed\ pregnant|Z = customer) = 2000 \cdot 0.37505 = 3000.4$$

3. In each group, for how many do you expect a wrong result?

$$P(Y = diagnosed\ pregnant, X = \overline{pregnant}|Z = technician)$$
$$+ P(Y = \overline{diagnosed\ pregnant}, X = pregnant|Z = technician)$$
$$= P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Z = technician) \cdot P(X = \overline{pregnant})$$
$$+ P(Y = \overline{diagnosed\ pregnant}|X = pregnant, Z = technician) \cdot P(X = pregnant)$$
$$= 0.026 \cdot 0.5 + 0.0001 \cdot 0.5 = 0.01305$$

$$P(Y = diagnosed\ pregnant, X = \overline{pregnant}|Z = customer)$$
$$+ P(Y = \overline{diagnosed\ pregnant}, X = pregnant|Z = customer)$$
$$= P(Y = diagnosed\ pregnant|X = \overline{pregnant}, Z = customer) \cdot P(X = \overline{pregnant})$$
$$+ P(Y = \overline{diagnosed\ pregnant}|X = pregnant, Z = customer) \cdot P(X = pregnant)$$
$$= 0.25 \cdot 0.5 + 0.0001 \cdot 0.5 = 0.12505$$

- For group A, expected number of wrong results is:

$$\#A \cdot \big(P(Y = diagnosed\ pregnant, X = \overline{pregnant}|Z = technician)$$
$$+ P(Y = \overline{diagnosed\ pregnant}, X = pregnant|Z = technician) = 2000 \cdot 0.01305 = 26.1$$

- For group B, expected number of wrong results is:

$$\#B \cdot \big(P(Y = diagnosed\ pregnant, X = \overline{pregnant}|Z = technician)$$
$$+ P(Y = \overline{diagnosed\ pregnant}, X = pregnant|Z = technician) = 8000 \cdot 0.12505 = 1000.4$$

Question 2.3

For this question you will compute the expression for the posterior parameter distribution for a simple data problem. Assume we observe N univariate data points $\{x_1, x_2, \dots, x_N\}$. Further, we assume that they are generated by a Gaussian distribution with known variance σ^2, but unknown mean μ.
Assume a prior Gaussian distribution over the unknown mean, i.e. $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. When answering these questions, use $\mathcal{N}(a|b, c^2)$ to indicate a Gaussian (normal) distribution over a with mean b and variance c^2.

1. Write down the general expression for a posterior distribution, using θ for the parameter, $\mathcal{D}$ for the data. Indicate the *prior*, *likelihood*, *evidence*, and *posterior*

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})},$$

where $p(\theta) - prior;\ p(\mathcal{D}|\theta) - likelihood;\ p(\mathcal{D}) - evidence;\ p(\theta|\mathcal{D}) - posterior$

2. Write the posterior for this particular example. You do not need an analytic solution.

$$p(\mu|x_1, x_2, \dots x_N) = \frac{p(x_1, x_2, \dots x_N|\mu)p(\mu)}{p(x_1, x_2, \dots x_N)} = \frac{p(x_1, x_2, \dots x_N|\mu)p(\mu)}{\int_{\mu'} p(x_1, x_2, \dots x_N|\mu')p(\mu')d\mu'}$$

$$= \frac{\mathcal{N}(x_1, x_2, \dots x_N|\mu, \sigma^2)\mathcal{N}(\mu|\mu_0, \sigma_0^2)}{\int_{\mu'} \left(\mathcal{N}(x_1, x_2, \dots x_N|\mu', \sigma^2)\mathcal{N}(\mu'|\mu_0, \sigma_0^2)\right) d\mu'} = \frac{(\prod_{i=1}^N \mathcal{N}(x_i|\mu, \sigma^2))\mathcal{N}(\mu|\mu_0, \sigma_0^2)}{\int_{-\infty}^{+\infty} \left((\prod_{i=1}^N \mathcal{N}(x_i|\mu', \sigma^2))\mathcal{N}(\mu'|\mu_0, \sigma_0^2)\right) d\mu'},$$

where

$\mathcal{N}(\mu|\mu_0, \sigma_0^2)$- prior,

$\prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2)$- likelihood,

$\int_{-\infty}^{+\infty} \left((\prod_{i=1}^{N} \mathcal{N}(x_i|\mu', \sigma^2))\mathcal{N}(\mu'|\mu_0, \sigma_0^2)\right) d\mu'$- evidence,

$p(\mu|x_1, x_2, \dots x_N)$- posterior