Statistical Machine Learning – Homework

Prof. Dr. Jan Peters
Daniel Palenicek, An Thai Le, Theo Gruner, Maximilian Tölle & Théo Vincent

TECHNISCHE UNIVERSITÄT DARMSTADT

Sommersemester 2023 – Due date: 15.05.2023, 23:59
Sheet 1

Each (sub)task in this homework is worth 1 point. For example, you can get up to 3 points in task 1.1. The points you achieve in this homework will count towards the exam bonus points.
The solutions need to be uploaded to moodle before the deadline. You can provide us with scans of your handwritten solutions or directly write into our .tex file and submit as .pdf. Please make sure that we can exactly determine which solution belongs to which (sub)task. If you decide for handwritten solutions, please understand that we can only give points for answers which we can also read!

Linear Algebra

Task 1.1: Pseudo inverse

1.1a)

Can the right-pseudo inverse matrix of $A = \begin{bmatrix} 1 & 0 & 2 \\ 5 & 1 & 3 \end{bmatrix}$ be computed? If yes, compute it. If no, explain why.

Solution:

Since the matrix has full row rank, the right-pseudo inverse A^+ can be computed as follow:

$$A^+ = A^\intercal \cdot (A \cdot A^\intercal)^{-1} \tag{1.1.a}$$

$$A^\intercal = \begin{bmatrix} 1 & 5 \\ 0 & 1 \\ 2 & 3 \end{bmatrix}; A \cdot A^\intercal = \begin{bmatrix} 1 & 0 & 2 \\ 5 & 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 5 \\ 0 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 5 & 11 \\ 11 & 35 \end{bmatrix}$$

$$(A \cdot A^\intercal)^{-1} = \frac{1}{\det(A \cdot A^\intercal)} \cdot \begin{bmatrix} 35 & -11 \\ -11 & 5 \end{bmatrix} = \frac{1}{54} \begin{bmatrix} 35 & -11 \\ -11 & 5 \end{bmatrix} = \begin{bmatrix} 35/54 & -11/54 \\ -11/54 & 5/54 \end{bmatrix}$$

Apply the calculation of $A^\intercal$ and $(A, A^\intercal)^{-1}$ to formular (1.1.a) above we have:

$$A^+ = \begin{bmatrix} 1 & 5 \\ 0 & 1 \\ 2 & 3 \end{bmatrix} \left(\frac{1}{54} \begin{bmatrix} 35 & -11 \\ -11 & 5 \end{bmatrix} \right) = \frac{1}{54} \begin{bmatrix} -20 & 14 \\ -11 & 5 \\ 37 & -7 \end{bmatrix}$$

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

1.1b)

Can the left-pseudo inverse matrix of $B = \begin{bmatrix} 1 & 7 & 2 & 3 \\ 0 & 4 & 1 & 5 \\ 0 & 1 & 0 & 4 \end{bmatrix}$ be computed? If yes, compute it. If no, explain why.

Solution:

Since B does not have full column rank, it does not have a left-pseudo inverse matrix.

1.1c)

Can the left-pseudo inverse matrix of $C = \begin{bmatrix} 6 & 3 & 2 \\ 0 & 2 & 4 \\ 9 & 3 & 0 \\ 3 & 5 & 8 \end{bmatrix}$ be computed? If yes, compute it. If no, explain why.

Solution:

Since C does not have full row rank, it does not have a left pseudo-inverse matrix

Task 1.2: Determinant

1.2a)

Compute the determinant of $D = \begin{bmatrix} 3 & 9 & 17 & 4 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 5 & 1 \\ 0 & 5 & 3 & 8 \end{bmatrix}$ using Leibniz formular.

Solution:

$\det(D) = \sum_{\sigma \in s_n} \text{sign}(\sigma) \cdot d_{\sigma(1)1} \dots d_{\sigma(n)n}$

$$= d_{11}d_{22}d_{33}d_{44} + d_{11}d_{32}d_{43}d_{24} + d_{11}d_{42}d_{23}d_{34}$$

$$+d_{21}d_{12}d_{43}d_{34} + d_{21}d_{32}d_{13}d_{44} + d_{21}d_{42}d_{33}d_{14}$$

$$+d_{31}d_{12}d_{23}d_{44} + d_{31}d_{22}d_{43}d_{14} + d_{31}d_{42}d_{13}d_{24}$$

$$+d_{41}d_{32}d_{23}d_{14} + d_{41}d_{22}d_{13}d_{34} + d_{41}d_{12}d_{33}d_{24}$$

$$-d_{11}d_{22}d_{43}d_{34} - d_{11}d_{42}d_{33}d_{24} - d_{11}d_{32}d_{23}d_{44}$$

$$-d_{21}d_{12}d_{33}d_{44} - d_{21}d_{32}d_{43}d_{14} - d_{21}d_{42}d_{13}d_{34}$$

$$-d_{31}d_{12}d_{43}d_{24} - d_{31}d_{22}d_{13}d_{44} - d_{31}d_{42}d_{23}d_{14}$$

$$-d_{41}d_{12}d_{23}d_{34} - d_{41}d_{22}d_{33}d_{14} - d_{41}d_{32}d_{13}d_{24}$$

$$= 0 + 0 + 3.5.2.1 - 0 - 0 - 3.5.5.3 = -195$$

(Since $d_{21} = d_{22} = d_{31} = d_{32} = d_{41} = 0$, the second, third, fourth, sixth, seventh, eighth rows from the calculation above is zero. For this reason we only have to calculate the first and fifth row.)

1.2b)

Compute the determinant of $E = \begin{bmatrix} 0 & 4 & 7 & 0 \\ 10 & 3 & 8 & 2 \\ 3 & 1 & 4 & 0 \\ 0 & 5 & 1 & 0 \end{bmatrix}$ using Laplace formular.

Solution:

$$\det(E) = 2 \cdot \det\left(\begin{bmatrix} 0 & 4 & 7 \\ 3 & 1 & 4 \\ 0 & 5 & 1 \end{bmatrix}\right) = 2 \cdot (-3) \cdot \det\left(\begin{bmatrix} 4 & 7 \\ 5 & 1 \end{bmatrix}\right) = (-6)(-31) = 186$$

1.2c)

Compute the determinant of $E^\intercal D^{-1}$.

Solution:

$$\det\left(E^\top \cdot D^{-1}\right) = \det\left(E^\top\right) \cdot \det\left(D^{-1}\right) = \det(E) \cdot \frac{1}{\det(D)} = \frac{186}{-195} = -\frac{62}{65}.$$

.

Task 1.3: Matrix decomposition

1.3a)

Compute the eigenvalues and the eigenvectors of $F = \begin{bmatrix} -3 & 0 & 0 \\ -42 & 3 & -6 \\ 14 & -2 & -1 \end{bmatrix}$.

Hint: The characteristic polynomial of F has a double root.

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Solution:

$$
\begin{aligned}
P(\lambda) &= \det(F - \lambda I) \\
\det(F - \lambda I) &= \det\left(\begin{bmatrix} -3-\lambda & 0 & 0 \\ -42 & 3-\lambda & -6 \\ 14 & -2 & -1-\lambda \end{bmatrix}\right) \\
&= (-3-\lambda)\det\left(\begin{bmatrix} 3-\lambda & -6 \\ -2 & -1-\lambda \end{bmatrix}\right) \\
&= (-3-\lambda)((3-\lambda)(-1-\lambda) - 12) \\
&= (-3-\lambda)((\lambda-3)(\lambda+1) - 12) \\
&= (-3-\lambda)\left(\lambda\lambda^2 - 2\lambda - 15\right) \\
&= -(\lambda+3)(\lambda-5)(\lambda+3) \\
P(\lambda) = \det(F - \lambda I) = 0 \quad &\text{iff} \begin{bmatrix} \lambda_1 = 5 \\ \lambda_{23} = -3 \end{bmatrix}
\end{aligned}
$$

For $\lambda_1 = 5$:

$$
\begin{aligned}
&(F - 5I)x_1 = 0 \\
&\text{iff} \begin{bmatrix} -8 & 0 & 0 \\ -42 & -2 & -6 \\ 14 & -2 & -6 \end{bmatrix} x_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
&\text{iff} \begin{bmatrix} -8 & 0 & 0 \\ -56 & 0 & 0 \\ 14 & -2 & -6 \end{bmatrix} x_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
&\text{iff } x_1 \text{ takes the form } \begin{pmatrix} 0 \\ 3a \\ -a \end{pmatrix} \text{ with a } \in R \\
&\text{We can choose } x_1 = \begin{pmatrix} 0 \\ 3 \\ -1 \end{pmatrix}
\end{aligned}
$$

For $\lambda_{2,3} = -3$:

$$
\begin{aligned}
&(F + 3I)x_{2,3} = 0 \\
&\text{iff} \begin{bmatrix} 0 & 0 & 0 \\ -42 & 6 & -6 \\ 14 & -2 & 2 \end{bmatrix} x_{2,3} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
&\text{iff} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 14 & -2 & 2 \end{bmatrix} x_{2,3} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\end{aligned}
$$

iff $x_{2,3}$ takes the form $\begin{pmatrix} 0 \\ a \\ a \end{pmatrix}$ for $a \in R$ or $\begin{pmatrix} -b \\ 0 \\ +7b \end{pmatrix}$ for $b \in R$ or $\begin{pmatrix} c \\ 7c \\ 0 \end{pmatrix}$ for $c \in R$

We can choose $x_2 = \begin{pmatrix} -1 \\ 0 \\ 7 \end{pmatrix}$ ard $x_3 = \begin{pmatrix} 1 \\ 7 \\ 0 \end{pmatrix}$

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

1.3b)

Let $G \in \mathbb{R}^{n \times n}, n \in \mathbb{N}$ be a positive definite symetric matrix. Show that there exists $H \in \mathbb{R}^{n \times n}$ such that $G = H^2$.

Solution:

Since the matrix G is real and symmetric, it can be decomposed into

$$G = Q \cdot \Lambda \cdot Q^\intercal$$

where the column of Q are the eigenvectors of G, and Λ is a diagonal matrix where the entries are the eigenvalues respectively.

Because G is also positive definite, all of its eigenvalues are also positive. Therefore, there exists matrix $\Lambda^{1/2}$ whose diagonal entries are square roots of the diagonal entries of Λ. We can choose matrix H such that:

$$H = Q \cdot \Lambda^{1/2} \cdot Q^\intercal$$

where H^2 satisfies

$$H^2 = H \cdot H = (Q \cdot \Lambda^{1/2} \cdot Q^\intercal)(Q \cdot \Lambda^{1/2} \cdot Q^\intercal) = Q \cdot \Lambda^{1/2} \cdot (Q^\intercal \cdot Q) \cdot \Lambda^{1/2} \cdot Q^\intercal = Q \cdot (\Lambda^{1/2} \cdot \Lambda^{1/2}) \cdot Q^\intercal = Q \cdot \Lambda \cdot Q^\intercal = G$$

.

For this reason for a real square symmetric positive definite matrix G, there exists a matrix H with $G = H^2$.

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Statistics

In this section, you are not allowed to use any kind of formula without adding its proof.

Task 1.4: Variances & Covariances

Hint: $\mathbb{E}\left[X^2\right] = \mathbb{E}[X(X-1)] + \mathbb{E}[X]$.

1.4a)

Compute the variance of X a random variable following a geometric distribution in function of its parameter p.

Solution:

$$\Pr(X = k) = p(k) = (1-p)^k p$$

$$\begin{aligned}\text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= E[X(X-1)] + E[X] - (E[X])^2\end{aligned}$$

$$\begin{aligned}E[X] &= \sum_{k=0}^{\infty} p(k)k \\ &= \sum_{k=0}^{\infty}(1-p)^k pk \\ &= p\sum_{k=0}^{\infty}(1-p)^k k \\ &= p(1-p)\sum_{k=0}^{\infty}(1-p)^{k-1}k \\ &= p(1-p)\sum_{k=0}^{\infty}\frac{d}{dp}(-(1-p)^k) \\ &= p(1-p)\frac{d}{dp}(-\sum_{k=0}^{\infty}(1-p)^k)\end{aligned}$$

$\sum_{k=0}^{\infty}(1-p)^k$ is a geometric series with $|1-p| < 1$, since $0 \leq p < 1$. Therefore

$$\sum_{k=0}^{\infty}(1-p)^k = \frac{1}{1-(1-p)} = \frac{1}{p}$$

Substitute into $E[X]$

$$\begin{aligned}
E[X] &= p(1-p)\frac{d}{dp}(-\sum_{k=0}^{\infty}(1-p)^k) \\
&= p(1-p)\frac{d}{dp}(-\frac{1}{p}) \\
&= p(1-p)\frac{1}{p^2} \\
&= \frac{1-p}{p}
\end{aligned}$$

$$\begin{aligned}
E[X(X-1)] &= \sum_{k=0}^{\infty}(1-p)^k pk(k-1) \\
&= p\sum_{k=0}^{\infty}(1-p)^k k(k-1) \\
&= p\sum_{k=0}^{\infty}(1-p)^k k(k-1) \\
&= p(1-p)\sum_{k=0}^{\infty}(1-p)^{k-1} k(k-1) \\
&= p(1-p)\sum_{k=0}^{\infty}(k-1)\frac{d}{dp}(-(1-p)^k) \\
&= p(1-p)\frac{d}{dp}(-\sum_{k=0}^{\infty}(k-1)(1-p)^k) \\
&= p(1-p)\frac{d}{dp}(-\sum_{k=1}^{\infty}(k-1)(1-p)^k) \\
&= p(1-p)\frac{d}{dp}(-(1-p)^2\sum_{k=1}^{\infty}(k-1)(1-p)^{k-2}) \\
&= p(1-p)\frac{d}{dp}((1-p)^2\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^{k-1}) \\
&= p(1-p)\frac{d}{dp}((1-p)^2\frac{d}{dp}\sum_{k=0}^{\infty}(1-p)^k) \\
&= p(1-p)\frac{d}{dp}((1-p)^2\frac{d}{dp}\frac{1}{p}) \\
&= p(1-p)\frac{d}{dp}(-\frac{(1-p)^2}{p^2})
\end{aligned}$$

$$\begin{aligned}\frac{d}{dp}(-\frac{(1-p)^2}{p^2}) &= -\frac{-2(1-p)p^2 - 2p(1-p)^2}{p^4} \\ &= \frac{2(1-p)p^2 + 2p(1-p)^2}{p^4} \\ &= \frac{2(1-p)(p^2 + p(1-p))}{p^4} \\ &= \frac{2(1-p)(p^2 + p - p^2))}{p^4} \\ &= \frac{2p(1-p)}{p^4} \\ &= \frac{2(1-p)}{p^3} \\ &= \frac{2-2p}{p^3}\end{aligned}$$

Plugging into E[X[X-1], then we have

$$\begin{aligned}E[X(X-1)] &= p(1-p)\frac{d}{dp}(-\frac{(1-p)^2}{p^2}) \\ &= \frac{p(1-p)(2-2p)}{p^3} \\ &= \frac{(1-p)(2-2p)}{p^2}\end{aligned}$$

$$\begin{aligned}E[X(X-1)] &= p(1-p)\frac{d}{dp}(-\frac{(1-p)^2}{p^2}) \\ &= \frac{p(1-p)(2-2p)}{p^3} \\ &= \frac{(1-p)(2-2p)}{p^2}\end{aligned}$$

$$\begin{aligned}\text{Var}(X) &= E[X(X-1)] + E[X] - (E[X])^2 \\ &= \frac{(1-p)(2-2p)}{p^2} + \frac{1-p}{p} - \frac{(1-p)^2}{p^2} \\ &= \frac{(1-p)(2-2p) + p(1-p) - (1-p)^2}{p^2} \\ &= \frac{(1-p)(2-2p+p-1+p)}{p^2} \\ &= \frac{(1-p)}{p^2}\end{aligned}$$

1.4b)

Compute the covariance between X a random variable following a poisson distribution and Y a random variable following a binomial distribution in function of the parameters λ, n, p. If the question seems unsolvable at first sight, you are allowed to add one more assumption on the relationship between X and Y.

Solution:

Assume X and Y are independent. This means Pr($X = x, Y = y$) $= p(x, y) = p(x)p(y)$.

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} xyp(x,y)dxdy \\ &= \int_{-\infty}^{\infty} xyp(x)p(y)dxdy \\ &= \int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} yp(y)dy \\ &= E[X]E[Y] \end{aligned}$$

$$\begin{aligned} \text{Cov}(X,Y) &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \\ &= 0 \end{aligned}$$

1.4c)

Compute the covariance between X a random variable following a poisson distribution and $Y = X^2 - 3X + 2$ in function of the parameter λ.

Solution:

$$\begin{aligned} \text{Cov}(X,Y) &= \text{Cov}(X, X^2 - 3X + 2) \\ &= E[X(X^2 - 3X + 2)] - E[X]E[X^2 - 3X + 2] \\ &= E[X^3 - 3X^2 + 2X] - E[X]E[X^2 - 3X + 2] \end{aligned}$$

Making use of the linearity of expectation, then we have

$$\begin{aligned} E[X^3 - 3X^2 + 2X] &= E[X^3] - 3E[X^2] + 2E[X] \\ E[X^2 - 3X + 2] &= E[X^2] - 3E[X] + 2 \end{aligned}$$

We need to calculate $E[X]$, $E[X^2]$ and $E[X^3]$.

$$\begin{aligned}
E[X] &= \sum_{k=0}^{\infty} k\frac{\lambda^k k^{-\lambda}}{k!} \\
&= e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\
&= e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\
&= e^{-\lambda}\lambda\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}
\end{aligned}$$

$\frac{\lambda^k}{k!}$ can be written as the exponential function e^{λ}, therefore

$$\begin{aligned}
E[X] &= e^{-\lambda}\lambda\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\lambda e^{\lambda} \\
&= \lambda \quad (1)
\end{aligned}$$

$$\begin{aligned}
E[X^2] &= \sum_{k=0}^{\infty} k^2\frac{\lambda^k k^{-\lambda}}{k!} \\
&= e^{-\lambda}\sum_{k=0}^{\infty}\frac{k^2\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=1}^{\infty}\frac{k(k-1+1)\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^k}{k!} + e^{-\lambda}\sum_{k=1}^{\infty}\frac{k(k-1)\lambda^k}{k!} \\
&= E[X] + e^{-\lambda}\sum_{k=1}^{\infty}\frac{k(k-1)\lambda^k}{k!}
\end{aligned}$$

$$\begin{aligned}
e^{-\lambda}\sum_{k=1}^{\infty}\frac{k(k-1)\lambda^k}{k!} &= e^{-\lambda}\sum_{k=2}^{\infty}\frac{k(k-1)\lambda^k}{k!} \\
&= e^{-\lambda}\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} \\
&= e^{-\lambda}\lambda^2\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\lambda^2 e^{\lambda} \\
&= \lambda^2 \quad (2)
\end{aligned}$$

$$\rightarrow E[X^2] = E[X] + \lambda^2 = \lambda + \lambda^2$$

$$\begin{aligned}
E[X^3] &= \sum_{k=0}^{\infty} k^3 \frac{\lambda^k k^{-\lambda}}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k^3 \lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k^2 - 1 + 1)\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k^2 - 1) + k)\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)(k+1) + k)\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)(k-2+3) + k)\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)(k-2) + 3k(k-1) + k)\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)(k-2)\lambda^k}{k!} + e^{-\lambda} \sum_{k=0}^{\infty} \frac{3k(k-1)\lambda^k}{k!} + e^{-\lambda} \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} \\
&=^{(1),(2)} e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)(k-2)\lambda^k}{k!} + 3\lambda^2 + \lambda
\end{aligned}$$

$$\begin{aligned}
e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)(k-2)\lambda^k}{k!} &= e^{-\lambda} \sum_{k=3}^{\infty} \frac{k(k-1)(k-2)\lambda^k}{k!} \\
&= e^{-\lambda} \lambda^3 \sum_{k=3}^{\infty} \frac{\lambda^{k-3}}{(k-3)!} \\
&= e^{-\lambda} \lambda^3 \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \lambda^3 e^{\lambda} \\
&= \lambda^3
\end{aligned}$$

$$\rightarrow E[X^3] = \lambda^3 + 3\lambda^2 + \lambda$$

$$\begin{aligned}
E[X^3 - 3X^2 + 2X] &= E[X^3] - 3E[X^2] + 2E[X] \\
&= \lambda^3 + 3\lambda^2 + \lambda - 3\lambda^2 - 3\lambda + 2\lambda \\
&= \lambda^3
\end{aligned}$$

$$\begin{aligned}
E[X^2 - 3X + 2] &= E[X^2] - 3E[X] + 2 \\
&= \lambda^2 + \lambda - 3\lambda + 2 \\
&- \lambda^2 - 2\lambda + 2
\end{aligned}$$

$$\begin{aligned}\text{Cov}(X,Y) &= E[X^3 - 3X^2 + 2X] - E[X]E[X^2 - 3X + 2] \\ &= \lambda^3 - \lambda(\lambda^2 - 2\lambda + 2) \\ &= \lambda^3 - \lambda^3 + 2\lambda^2 - 2\lambda \\ &= 2\lambda^2 - 2\lambda\end{aligned}$$

Task 1.5: Exponential family

1.5a)

Is the uniform distribution part of the exponential family? Justify the answer.

Solution:

The continuous uniform distribution has the PDF

$$\begin{aligned}p(x|a,b) &= \begin{cases}\frac{1}{b-a} & a \le x \le b \\ 0 & x < a \vee x > b\end{cases} \\ &= \frac{1}{b-a} I_{[a,b]}(x)\end{aligned}$$

with

$$I_{[a,b]}(x) = \begin{cases}1 & a \le x \le b \\ 0 & x < a \vee x > b\end{cases}$$

Assume $p(x|a,b)$ is of the exponential family with natural parameter $\boldsymbol{\eta} = (a,b)$

$$p(x|\boldsymbol{\eta}) = h(x)g(\boldsymbol{\eta})e^{\boldsymbol{\eta}^T u(x)}$$

Since both $\frac{1}{b-a}$ and $I_{[a,b]}$ depend on $\boldsymbol{\eta}$, they cannot be represented by $h(x)$. Therefore $h(x) = 1 \quad \forall x$.
$g(\boldsymbol{\eta}) = \frac{1}{b-a}$, since the function depends only on $\boldsymbol{\eta}$ and the other factor, $I_{[a,b]}(x)$, depends additionally on x.
So what is left, $e^{\boldsymbol{\eta}^T u(x)}$ has to represent $I_{[a,b]}(x)$. However from the definition of $I_{[a,b]}$, there is no way to plug the function in the exponential.
Therefore the continuous uniform distribution cannot be of the exponential family.

Task 1.6: Central Limit Theorem [coding question]

1.6a)

For this question, add the code to the answer. It should not be more than 30 lines of code. Using Python, show the effect of the Central Limit Theorem by reproducing the 3 plots of slide 22/47 for an exponential distribution with $\lambda = 2$. In addition, you are only allowed to sample from a uniform distribution. Explain what is happening in the 3 plots.

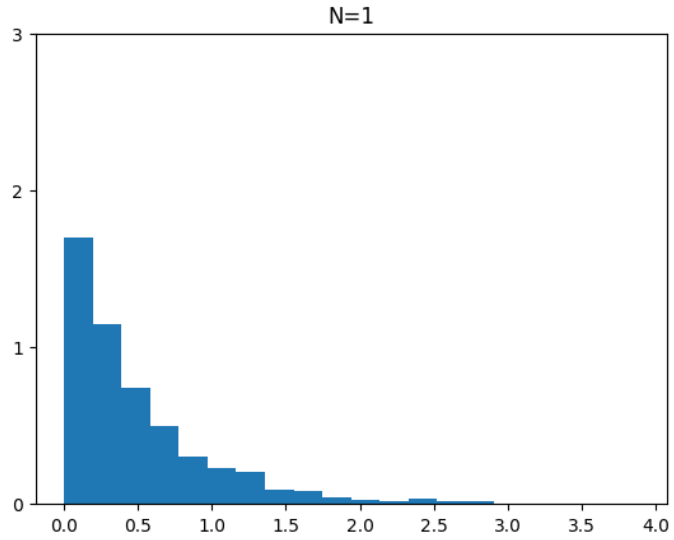
Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

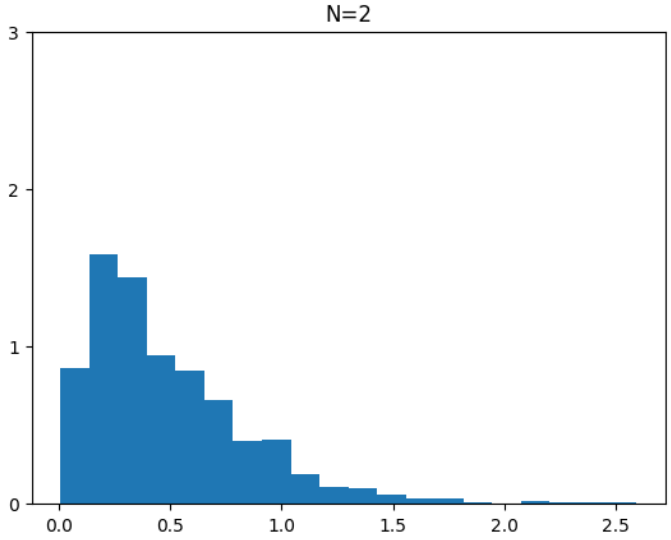
Solution:

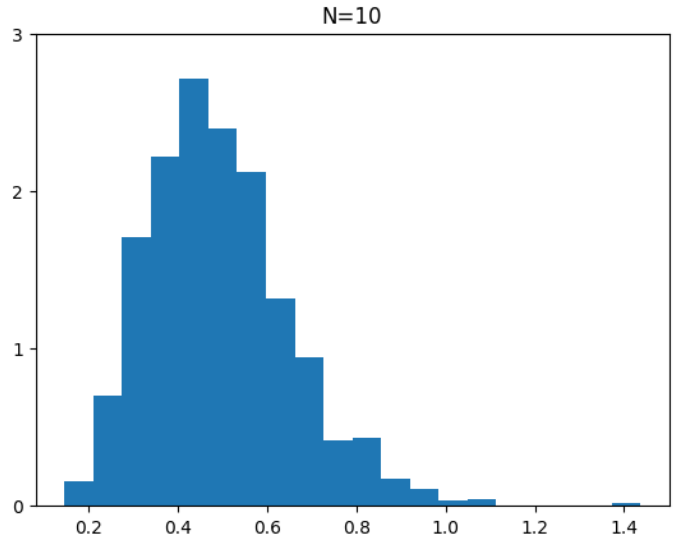
```
import numpy as np
import matplotlib.pyplot as plt

n_samples = [1,2,10]
for i in n_samples:
    np.random.seed(40)
    x = [np.mean(-np.log(np.random.random_sample(i))/2) for j in range(1000)]
    plt.hist(x,20,density=True)
    plt.yticks([0,1,2,3])
    plt.show()
```

When we have only 1 sample or 2 samples, the sampling distribution of the mean still looks like the exponential distribution. However when 10 samples are taken, it is already looking close to a Gaussian distribution.







Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Task 1.7: KL Divergence

1.7a)

Let p be a PDF. We set $p(x) = \frac{1}{2}\mathcal{N}(x|\mu, \sigma^2) + \frac{1}{2}\mathcal{N}(x|\lambda, \gamma^2)$. Let q be the PDF of $\mathcal{N}(\alpha, \beta^2)$. Find α^*, β^* such that $(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta} D_{\mathrm{KL}}(p, q)$.

Solution:

$$\begin{aligned} KL(p\|q) &= -\int p(x) \ln \frac{q(x)}{p(x)} dx \\ &= -\int p(x) \ln q(x) dx + \int p(x) \ln p(x) dx \\ &= -\int \frac{1}{2} p_1(x) \ln q(x) dx - \int \frac{1}{2} p_2(x) \ln q(x) dx \end{aligned}$$

(Explanation for the transition from second line to third line: $\int p(x) \ln p(x) dx$ is known and fixed, therefore we only have to consider the first term $-\int p(x) \ln q(x) dx$)

$$\begin{aligned} -\int \frac{1}{2} p_1(x) \ln q(x) dx &= -\int \frac{1}{2} \left(-\ln \left(\left(2\pi\beta^2\right)^{1/2} \right) p_1(x) dx - \int \frac{1}{2} \left(-\frac{1}{2\beta^2} (x-\alpha)^2 \right) p_1(x) dx \right. \\ &= \frac{1}{4} \ln \left(2\pi\beta^2\right) \cdot E_{x\sim p_1} \left[x^0\right] + \frac{1}{4\beta^2} E_{x\sim p_1} \left[(x-\alpha)^2\right] \\ &= \frac{1}{4} \ln \left(2\pi\beta^2\right) + \frac{1}{4\beta^2} E_{x\sim p_1} \left[x^2 - 2\alpha x + \alpha^2\right] \\ &= \frac{1}{4} \ln \left(2\pi\beta^2\right) + \frac{1}{4\beta^2} \left(\sigma^2 + \mu^2 - 2\alpha\mu + \alpha^2\right) \end{aligned}$$

Similarly: we have

$$-\int \frac{1}{2} p_2(x) \ln q(x) dx = \frac{1}{4} \ln \left(2\pi\beta^2\right) + \frac{1}{4\beta^2} \left(\gamma^2 + \lambda^2 - 2\alpha\lambda + \alpha^2\right)$$

Therefore:

$$\begin{aligned} -\int p(x) \ln q(x) dx &= \frac{1}{2} \ln \left(2\pi\beta^2\right) + \frac{1}{4\beta^2} \left(\sigma^2 + \mu^2 + \gamma^2 + \lambda^2 - 2\alpha(\mu+\lambda) + 2\alpha^2\right) \\ \frac{\partial KL(p\|q)}{\partial \alpha} &= \frac{1}{4\beta^2}(-2\alpha(\mu+\lambda) + 4\alpha) = 0 \quad \text{iff } \alpha^* = \frac{\mu+\lambda}{2}. \\ \frac{\partial KL\left(p\|q\right)}{\partial \beta} &= \frac{1}{\beta} - \frac{1}{2\beta^3} \left(\sigma^2 + \mu^2 + \gamma^2 + \lambda^2 - 2\alpha(\mu+\lambda) + 2\alpha^2\right) = 0 \end{aligned}$$

iff $2\beta^2 - \left(\sigma^2 + \mu^2 + \gamma^2 + \lambda^2 - 2\alpha(\mu+\lambda) + 2\alpha^2\right) = 0.$

iff $2\beta^2 - \left(\sigma^2 + \mu^2 + \gamma^2 + \lambda^2 - \frac{1}{2}(\mu+\lambda)^2\right) = 0.$

iff $\beta^2 = \frac{1}{2} \left(\sigma^2 + \mu^2 + \gamma^2 + \lambda^2 - \frac{1}{2}(\mu+\lambda)^2\right)$

iff $\beta^* = \sqrt{\frac{1}{2} \left(\sigma^2 + \mu^2 + \gamma^2 + \lambda^2 - \frac{1}{2}(\mu+\lambda)^2\right)}$

1.7b)

When $\mu = 1, \lambda = 10, \sigma = 1$ and $\gamma = 1.5$, the solution of $\arg\min_{\alpha,\beta} D_{\mathrm{KL}}(q, p)$ is $(\alpha^{**}, \beta^{**}) = (10, 1.5)$. Using python, draw the obtained solution of the previous question along with the solution given in this question and the distribution p. Comment on the different behaviours by giving the intuition behind the minimizers.

The code is not expected for this question.

Solution:

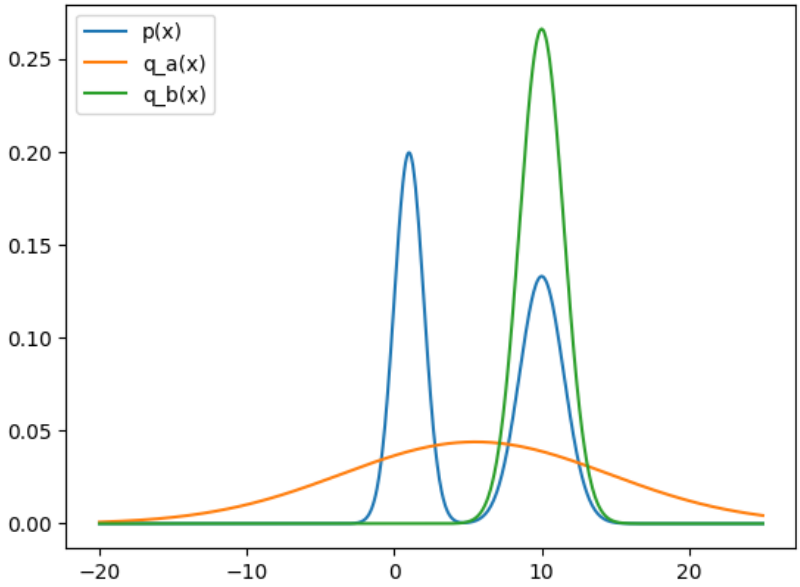


Figure 1: Solution to 1.7.b

Task 1.8: Manipulating Gaussians

1.8a)

Let $A \in \mathbb{R}^{n\times n}, n \in \mathbb{N}^*, b \in \mathbb{R}^n, X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$. Let $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$. Which distribution is $Y = AX + b + \epsilon$ following? Compute its parameters in function of $A, b, \mu, \Sigma, \Sigma_\epsilon$.

Hint: You can use the fact without proving it that the characterictic function of X is, for $t \in \mathbb{R}^n$: $\phi_X(t) = \mathbb{E}(e^{it^\intercal \mu - \frac{1}{2}t^\intercal \Sigma t})$.

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Solution:

$$\begin{aligned}
\varphi_y(t) &= E(e^{it^T y}) \\
&= E(e^{it^T Ax + it^T B + it^T \epsilon}) \\
&= e^{it^T B} E(e^{it^T Ax}) E(e^{it^T \epsilon}) \\
&= e^{it^T B} E(e^{it^T Ax}) E(e^{it^T \epsilon}) \\
&= e^{it^T B} E(e^{i(A^T t)^T x}) \phi_\epsilon(t) \\
&= e^{it^T B} \varphi_x(A^T t) E(e^{-\frac{1}{2} t^T \Sigma_\epsilon t}) \\
&= e^{it^T B} E(e^{it^T A\mu - \frac{1}{2} t^T A\Sigma A^T t}) E(e^{-\frac{1}{2} t^T \Sigma_\epsilon t}) \\
&= E(e^{it^T B + it^T A\mu - \frac{1}{2} t^T A\Sigma A^T t - \frac{1}{2} t^T \Sigma_\epsilon t}) \\
&= E(e^{it^T (B + A\mu) - \frac{1}{2} t^T (A\Sigma A^T + \Sigma_\epsilon) t})
\end{aligned}$$

$\rightarrow Y$ is a normal distributed random variable with

$$\begin{aligned}
\text{mean} &= B + A\mu \\
\text{var} &= A\Sigma A^T + \Sigma_\epsilon
\end{aligned}$$

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Optimization

Task 1.9: Convexity

For $n = 1, ..., N$ input samples $(\mathbf{x}_n, y_n)$, we want to find the weights $\mathbf{w}$ that minimize the differentiable cost function

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2.$$

Proof under which conditions for $\mathbf{X}$ and $\mathbf{y}$ a unique global minimum exists. Provide the unique solution $\mathbf{w}^*$.
Hint: Use the Hessian.

Solution:

Consider each sample x_n has a dimension of m: $x_n = (x_n^1, x_n^2, ...x_n^m)$
Rewrite the cost function in summation form: $\mathcal{J}(w) = \frac{1}{2}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 = \frac{1}{2}\Sigma_{i=1}^N(x_i\mathbf{w} - y_i)$
The gradient of $\mathcal{J}(\mathbf{w})$, $\nabla\mathcal{J}(\mathbf{w})$ is a row vector. We take the derivative with respect to each w_j of $\mathbf{w}$:
$\frac{\partial\mathcal{J}}{\partial w_j} = \frac{1}{2}\Sigma_{i=1}^N 2x_i^j(x_i\mathbf{w} - y_i) = \Sigma_{i=1}^N x_i^j(x_i\mathbf{w} - y_i)$ with $j = 1, ...m$
Rewrite the gradient in matrix form:

$$\begin{aligned}\nabla\mathcal{J}(\mathbf{w}) = (\frac{\partial\mathcal{J}}{\partial w_1}, ...\frac{\partial\mathcal{J}}{\partial w_m}) &= (\Sigma_{i=1}^N x_i^1(x_i\mathbf{w} - y_i), ..., \Sigma_{i=1}^N x_i^2(x_i\mathbf{w} - y_i)) \\ &= (\mathbf{X}\mathbf{w} - \mathbf{y})^T\mathbf{X}\end{aligned}$$

With the matrix notation of the gradient, we compute the Hessian using matrix calculus:

$$\begin{aligned}H &= \frac{\partial\nabla\mathcal{J}}{\partial\mathbf{w}} \\ &= \frac{\partial}{\partial\mathbf{w}}(\mathbf{X}\mathbf{w} - \mathbf{y})^T\mathbf{X} \\ &= \frac{\partial}{\partial\mathbf{w}}\mathbf{X}^\mathbf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= \mathbf{X}^\mathbf{T}\mathbf{X}\end{aligned}$$

$$v\mathbf{X}^\mathbf{t}\mathbf{X}v^T = (\mathbf{X}v^T)^T(\mathbf{X}v^T) = ||\mathbf{X}v^T||_2^2 \geq 0 \quad \forall v \rightarrow H\text{is always positive semidefinite}$$

The solution $\mathbf{w}^*$ first have to be a stationary point, which means

$$\nabla\mathcal{J}(\mathbf{w}^*) = (\mathbf{X}\mathbf{w}^* - \mathbf{y})^T\mathbf{X} = 0$$

Case 1: $\mathbf{X} = \mathbf{0}$, then $H = \mathbf{X}^\mathbf{T}\mathbf{X} = \mathbf{0}$, which is only positive semidefinite. We need H to be positive definite in order for w^* to be stationary.
Case 2: $\mathbf{X}\mathbf{w}^* - \mathbf{y} = \mathbf{0}$, then $\mathbf{X}$ must be non-singular for there to be a solution $w^* = \mathbf{X}^{-1}\mathbf{y}$
Therefore the first condition is that $\mathbf{X}$ is non-singular.
Since $H = \mathbf{X}^\mathbf{T}\mathbf{X}$ is positive semidefinite for any $\mathbf{X}$, this means $\mathbf{w}^*$ is always a local miminum.
For $\mathbf{w}^*$ to be a global minimum, $\mathcal{J}$ has to be strictly convex $\leftrightarrow$ H has to be positive definite.

$$\begin{aligned}&H = \mathbf{X}^\mathbf{T}\mathbf{X}\text{ is positive definite} \\ &\leftrightarrow v^T\mathbf{X}^\mathbf{T}\mathbf{X}v > 0 \quad \forall v \neq \mathbf{0} \\ &\leftrightarrow ||\mathbf{X}v||_2^2 > 0 \quad \forall v \neq \mathbf{0} \\ &\leftrightarrow \mathbf{X}v \quad \text{is a nontrivial linear combination} \\ &\leftrightarrow \text{Each column of X is linear independent} \\ &\leftrightarrow \text{X is non-singular}\end{aligned}$$

When $\mathbf{X}$ is non-singular, $\mathbf{X}\mathbf{w}^* - \mathbf{y} = \mathbf{0}$ has a unique solution $\mathbf{w}^* = \mathbf{X}^{-1}\mathbf{y}$
Therefore the condition for there to exist a unique minimum is that $\mathbf{X}$ must be non-singular.

Task 1.10: Numerical Optimization

Given is the following cost function

$$\mathcal{J}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log\left(\sigma(\mathbf{w}^\intercal \mathbf{x}_n)\right) + (1-y_n)\log\left(1-\sigma(\mathbf{w}^\intercal \mathbf{x}_n)\right)]$$

with

$$\sigma(\mathbf{w}^\intercal \mathbf{x}_n) = \frac{1}{1+e^{-\mathbf{w}^\intercal \mathbf{x}_n}}$$

being the sigmoid function.

1.10a)

As we are not able to find a closed-form solution, we fall back to a numerical optimization method. Here, we would like to use **stochastic gradient descent** to minimize the cost function. Please derive the corresponding update rule. Then express your derived update rule in matrix form.

Solution:

We first simplify the cost function:

$$\begin{aligned}
\ln(\sigma(\mathbf{w^T x_n})) &= \ln(\frac{1}{1+e^{-\mathbf{w^T x_n}}}) = -\ln(1+e^{-\mathbf{w^T x_n}}) \\
\ln(1-\sigma(\mathbf{w^T x_n})) &= \ln(1-\frac{1}{1+e^{-\mathbf{w^T x_n}}}) \\
&= \ln(\frac{e^{-\mathbf{w^T x_n}}}{1+e^{-\mathbf{w^T x_n}}}) \\
&= \ln(e^{-\mathbf{w^T x_n}}) - \ln(1+e^{-\mathbf{w^T x_n}}) \\
&= e^{-\mathbf{w^T x_n}} - \ln(1+e^{-\mathbf{w^T x_n}}) \\
&= -\mathbf{w^T x_n} - \ln(1+e^{-\mathbf{w^T x_n}})
\end{aligned}$$

$$\begin{aligned}
\mathcal{J}(\mathbf{w}) &= -\frac{1}{N}\sum_{n=1}^{N}[y_n \ln\left(\sigma(\mathbf{w}^\intercal \mathbf{x}_n)\right) + (1-y_n)\ln\left(1-\sigma(\mathbf{w}^\intercal \mathbf{x}_n)\right)] \\
&= -\frac{1}{N}\sum_{n=1}^{N}[-y_n \ln(1+e^{-\mathbf{w^T x_n}}) + (1-y_n)(-\mathbf{w^T x_n} - \ln(1+e^{-\mathbf{w^T x_n}}))] \\
&= -\frac{1}{N}\sum_{n=1}^{N}(y_n \mathbf{w^T x_n} - \mathbf{w^T x_n} - \ln(1+e^{-\mathbf{w^T x_n}}) \\
&= -\frac{1}{N}\sum_{n=1}^{N}(y_n \mathbf{w^T x_n} - (\ln(e^{\mathbf{w^T x_n}}) + \ln(1+e^{-\mathbf{w^T x_n}})) \\
&= -\frac{1}{N}\sum_{n=1}^{N}(y_n \mathbf{w^T x_n} - \ln(e^{\mathbf{w^T x_n}} * (1+e^{-\mathbf{w^T x_n}})) \\
&= -\frac{1}{N}\sum_{n=1}^{N}(y_n \mathbf{w^T x_n} - \ln(e^{\mathbf{w^T x_n}} + 1))
\end{aligned}$$

Now we compute the partial derivatives of $\mathcal{J}$ w.r.t to w_j, with $j = 1, ..., m$.

$$\frac{\partial}{\partial w_j}(y_n \mathbf{w^T x_n}) = y_n x_n^j$$

$$\begin{aligned}\frac{\partial}{\partial w_j}(\ln(e^{\mathbf{w^T x_n}} + 1)) &= \frac{x_n^j e^{\mathbf{w^T x_n}}}{1 + e^{\mathbf{w^T x_n}}} \\ &= \frac{x_n^j}{\frac{1+e^{\mathbf{w^T x_n}}}{e^{\mathbf{w^T x_n}}}} \\ &= \frac{x_n^j}{1 + \frac{1}{e^{\mathbf{w^T x_n}}}} \\ &= \frac{x_n^j}{1 + e^{-\mathbf{w^T x_n}}} \\ &= x_n^j \sigma(\mathbf{w^T x_n})\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial w_j} &= -\frac{1}{N}\sum_{n=1}^{N} x_n^j (y_n - \sigma(\mathbf{w^T x_n})) \\ &= -\frac{1}{N}\sum_{n=1}^{N} x_n^j (\sigma(\mathbf{w^T x_n}) - y_n)\end{aligned}$$

Convert summation form into matrix form:

$$\begin{aligned}\nabla \mathcal{J} &= (-\frac{1}{N}\sum_{n=1}^{N} x_n^1 (\sigma(\mathbf{w^T x_n}) - y_n), ..., -\frac{1}{N}\sum_{n=1}^{N} x_n^m (\sigma(\mathbf{w^T x_n}) - y_n))^T \\ &= -\frac{1}{N}\mathbf{X}^T(\varsigma - \mathbf{y})\end{aligned}$$

where $\varsigma = (\sigma(\mathbf{w^T x_1}), ..., \mathbf{w^T x_n})^T$
Then the update rule for SGD is as follows

$$\mathbf{w_{k+1}} = \mathbf{w_k} - \alpha \nabla \mathcal{J}(\mathbf{w_k}) = \mathbf{w_k} + \alpha \frac{1}{N}\mathbf{X}^T(\varsigma(\mathbf{w_k}) - \mathbf{y})$$

1.10b)

Does the cost function have a local or global minimum? Please proof your answer.
Hint: Use the Hessian.

Solution:

We compute the second order partial derivative of $\mathcal{J}$ w.r.t to $\mathbf{w}$.

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

$$\frac{\partial}{\partial w_k}\sigma(\mathbf{w^T x_n}) = x_n^k \sigma(\mathbf{w^T x_n})(1 - \sigma(\mathbf{w^T x_n}))$$

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial w_k w_j} &= \frac{\partial}{\partial w_k}(-\frac{1}{N}\sum_{n=1}^{N} x_n^j (\sigma(\mathbf{w^T x_n}) - y_n)) \\ &= -\frac{1}{N}\sum_{n=1}^{N} x_n^j \frac{\partial}{\partial w_k}(\sigma(\mathbf{w^T x_n})) \\ &= -\frac{1}{N}\sum_{n=1}^{N} x_n^j x_n^k \sigma(\mathbf{w^T x_n})(1 - \sigma(\mathbf{w^T x_n}))\end{aligned}$$

Convert summation form to matrix form:

$$\nabla^2 \mathcal{J} = -\frac{1}{N}\mathbf{X^T D X}$$

where $\mathbf{D}$ is a NxN diagonal matrix with $D_{ii} = \sigma(\mathbf{w^T x_n})(1 - \sigma(\mathbf{w^T x_n}) > 0$. Therefore $D > \mathbf{0}$.
For every $v \in R^m$, we have

$$\begin{aligned} v\mathbf{X^T D X}v^T &= (\mathbf{X}v^T)^T\mathbf{D}(\mathbf{X}v^T) \\ &= ||\mathbf{X D}v^T||_2^2 \geq 0 \end{aligned}$$

Therefore $H = \nabla^2\mathcal{J}$ is always positive semidefinite $\leftrightarrow$ $\mathcal{J}$ is convex $\leftrightarrow$ Local minimum is also global minimum. So we are sure that we have reached global minimum.

1.10c)

Stochastic gradient descent is a prominent but slow optimization algorithm. Under the assumption of convergence, can Newton's method increase convergence speed? Consider the following two criteria in your answer: number of optimization steps until convergence and computing time per optimization step.

Solution:

Newton's method usually has less optimization steps until convergence than SGD. However the computing time per optimization step of Newton's is higher than that of SGD, because at every step we need to compute the second order derivative, while for SGD only the first order derivative is needed. So Newton's can potentially increase the convergence speed if the computation time for the second order derivative is not too high. The computation time is partly dependent on the number of samples, since more samples = more effort to calculate the Hessian.

Task 1.11: Lagrange multiplier

Given is the constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_n(\mathbf{w}^\intercal \mathbf{x}_n + b) \geq 1 \quad \text{for} \quad n = 1, ..., N \end{aligned}$$

Hint: For the following tasks, the KKT-conditions should be considered (λ_n: Lagrange multiplier, $f_n(\mathbf{w}, b)$: constraint)

$$\begin{aligned} \lambda_n &\geq 0 \quad \text{for} \quad n = 1, ..., N \\ f_n(\mathbf{w}, b) &\geq 0 \quad \text{for} \quad n = 1, ..., N \\ \lambda_n f_n(\mathbf{w}, b) &= 0 \quad \text{for} \quad n = 1, ..., N. \end{aligned}$$

1.11a)

Formulate the Lagrangian and solve for **w** and b.
Hint: $y_n \in \{-1, 1\}$, so $y_n^2 = 1$

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Solution:

We consider that $\mathbf{w}$ and each $\mathbf{x_n}$ have dimension m.

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\lambda}) &= \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} \lambda_n (y_n(\mathbf{w^T} x_n + b) - 1) \\ &= \frac{1}{2} \sum_{i=1}^{m} (w_1^2 + w_2^2 + ... + w_m^2) - \sum_{n=1}^{N} \lambda_n (y_n(\mathbf{w^T x_n} + b) - 1) \end{aligned}$$

Take the partial derivative of L w.r.t to each w_i:

$$\begin{aligned} \frac{\partial L}{\partial w_i} &= \frac{1}{2} \frac{\partial}{\partial w_i} (\sum_{i=1}^{m} (w_1^2 + w_2^2 + ... + w_m^2)) - \frac{\partial}{\partial w_i} (\sum_{n=1}^{N} \lambda_n (y_n(\mathbf{w^T x_n} + b) - 1)) \\ &= w_i + (\sum_{n=1}^{N} \lambda_n y_n x_n^i) \end{aligned}$$

Then we can derive w_i and $\mathbf{w}$, which is dependent on the Lagrange multipliers and data points:

$$\begin{aligned} &\frac{\partial L}{\partial w_i} = 0 \\ &\rightarrow w_i + \sum_{n=1}^{N} \lambda_n y_n x_n^i = 0 \\ &\rightarrow w_i = \sum_{n=1}^{N} \lambda_n y_n x_n^i \\ &\rightarrow \mathbf{w} = \sum_{n=1}^{N} \lambda_n y_n \mathbf{x_n} \end{aligned}$$

We derive b by making use of the KKT-conditions and the fact that b depends on data points for which $\lambda_n \neq 0$.

$$\begin{aligned} &\lambda_n (y_n(\mathbf{w^T x_n} + b) - 1) = 0 \\ &y_n(\mathbf{w^T x_n} + b) - 1 = 0 \quad (\lambda_n \neq 0) \\ &y_n(\mathbf{w^T x_n} + b) = 1 \\ &y_n^2(\mathbf{w^T x_n} + b) = y_n \\ &\mathbf{w^T x_n} + b = y_n \quad (y_n^2 = 1) \\ &b = y_n - \mathbf{w^T x_n} \quad (y_n^2 = 1) \\ &= y_n - \sum_{i=1}^{N} (\lambda_i y_i \mathbf{x_i^T}) \mathbf{x_n} \end{aligned}$$

1.11b)

Our primal formulation can sometimes be tricky to solve. Therefore, we would like to switch to the dual formulation. Please derive the dual formulation and define the corresponding optimization problem. You do not need to solve it!

Solution:

Substitute $\mathbf{w}$ that we derived in the previous task in L:

$$\begin{aligned}
L(\mathbf{w}, \boldsymbol{\lambda}) &= G(\boldsymbol{\lambda}) \\
&= \frac{1}{2}\sum_{k=1}^{N}\lambda_k y_k \mathbf{x_k}\sum_{l=1}^{N}\lambda_l y_l \mathbf{x_l} - \sum_{n=1}^{N}\lambda_n(y_n(\sum_{i=1}^{N}(\lambda_i y_i \mathbf{x_i})\mathbf{x_n} + b) - 1) \\
&= \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\lambda_k\lambda_l y_k y_l \mathbf{x_k}\mathbf{x_l} - \sum_{n=1}^{N}\sum_{i=1}^{N}\lambda_n\lambda_i y_n y_i \mathbf{x_n}\mathbf{x_i} - b\sum_{n=1}^{N}\lambda_n y_n + \sum_{n=1}^{N}\lambda_n \\
&= \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\lambda_k\lambda_l y_k y_l \mathbf{x_k}\mathbf{x_l} - b\sum_{n=1}^{N}\lambda_n y_n
\end{aligned}$$

The partial derivative of L w.r.t to b is 0:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{1}{2}\frac{\partial}{\partial b}(\sum_{i=1}^{m}(w_1^2 + w_2^2 + ... + w_m^2)) - \frac{\partial}{\partial b}(\sum_{n=1}^{N}\lambda_n(y_n(\mathbf{w^T}\mathbf{x_n} + b) - 1)) \\
&= \sum_{n=1}^{N}\lambda_n y_n = 0
\end{aligned}$$

Substitute this in $G(\boldsymbol{\lambda})$, we have

$$G(\boldsymbol{\lambda}) = \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\lambda_k\lambda_l y_k y_l \mathbf{x_k}\mathbf{x_l}$$

The dual problem is then formulated as

$$\begin{aligned}
&\max_{\boldsymbol{\lambda}} \quad G(\boldsymbol{\lambda}) = \sum_{n=1}^{N}\lambda_n - \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\lambda_k\lambda_l y_k y_l \mathbf{x_k}\mathbf{x_l} \\
&\text{s.t.} \quad \lambda_n \geq 0 \quad \text{for} \quad n = 1, ..., N
\end{aligned}$$

Bayesian Decision Making

Task 1.12: Bayes' Theorem

Suppose you work for a tech company that produces two types of computer chips, A and B. The company has two factories, X and Y, that produce the chips. Factory X produces 55% of the chips, while factory Y produces the remaining 45%. The quality of the chips produced by factory X is such that 4% of the chips produced by X are defective, while the quality of the chips produced by factory Y is such that 3% of the chips produced by Y are defective.

A customer purchases one of these chips from your company, but you don't know which factory produced the chip. The customer reports that the chip is defective. What is the probability that the chip was produced by factory X? Please derive your answer.

Solution:

In order to keep the solution clear and concise, we define the following events for the subtask:

- $defective$: event when the chip is defective, regardless which factory produced it
- X: event when the factory X produced the chip
- Y: event when the factory Y produced the chip

With our notation defined, we want to calculate the probability that the chip which was purchased by the customer is defective, given that the chip was produced by factory X. This probability is namely

$$P(X|defective) \tag{1.12.1}$$

.

Given are the following probabilities:

- $P(X)$ = 0.55
- $P(Y)$ = 0.45
- $P(defective|X)$ = 0.04
- $P(defective|Y)$ = 0.03

In order to calculate our goal in (1.12.1), we can use the Bayes Theorem as following:

$$P(X|defective) = \frac{P(defective|X).P(X)}{P(defective)} \tag{1.12.2}$$

Since the probabilities $P(defective|X)$ and $P(X)$ are already given, we only have to calculate the marginal probability $P(defective)$.

$$\begin{aligned} P(defective) &= P(defective, X) + P(defective, Y) \\ &= P(defective|X)P(X) + P(defective|Y)P(Y) \\ &= 0.55 * 0.04 + 0.45 * 0.03 = 0.0355 \end{aligned}$$

Apply this result to (1.12.1) we have

$$P(X|defective) = 0.04 * 0.55/0.0355 \approx 0.62$$

So the probability that the purchased chip was produced by factory X is around 62%

Task 1.13: Misclassification

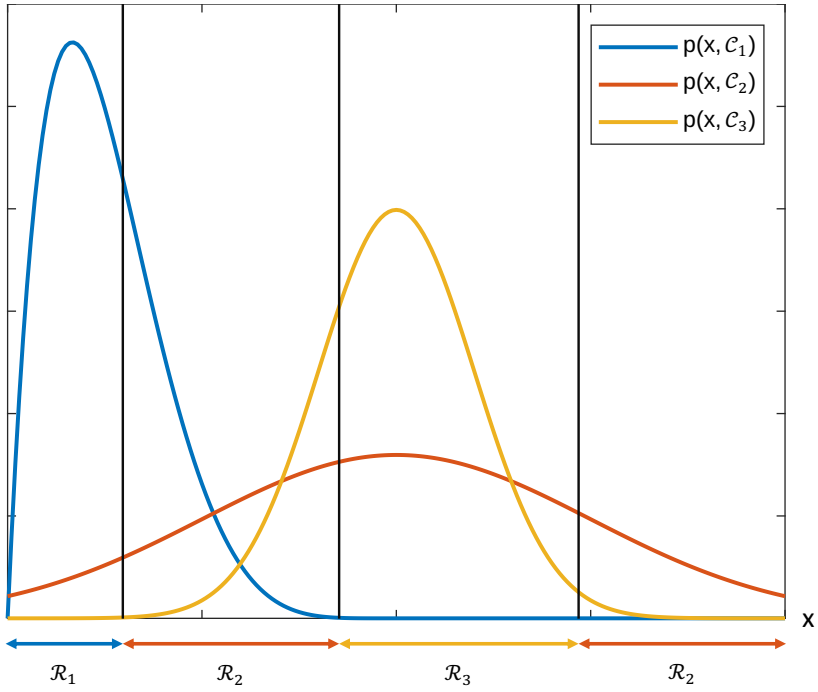
Within the figure, the three joint probabilities $p(x, \mathcal{C}_1)$, $p(x, \mathcal{C}_2)$ and $p(x, \mathcal{C}_3)$ are given. The initial decision regions are provided underneath as $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$. For instance, a data point x within the range of decision region $\mathcal{R}_2$ will be classified as $\mathcal{C}_2$.
We are now going to minimize the probability of misclassification, $p(\text{error})$.

1.13a)

Within the figure, mark the regions under the probability distributions that will no longer contribute to $p(\text{error})$.

1.13b)

Is it possible to reduce the probability of misclassification to zero in this example? Please explain your answer.



Solution:

1.13a)
In the figure "Solution to 1.13.1", we marked the green region R_x to be the region which doesn't contribute to $p(error)$. Every data point in this region is classified to class C_2 and the area which the graph of $p(x, C_1)$ and $p(x, C_3)$ span with the x-axis is zero. Therefore, there's no contribution to the $p(error)$.

1.13b)
In this example it is not possible to reduce the probability of misclassification to zero. When the optimal decision boundaries are chosen as interceptions of $p(x, C_x)$ with each others in figure 2, the misclassification is reduced to its minimum. However, even in this case, within an region R_i, the graphs of $p(x, C_j)$ and $p(x, C_k)$ with $i \neq j$ and $i \neq k$ still span non-zero areas with x-axis (see the marked green area in figure 2), hence make the $p(error)$ greater than 0.

Task 1.14: Optimal Decision Boundary

We consider discrete data of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ that is being generated by two corresponding Poisson distributions with parameters λ_1 and λ_2. For instance, class $\mathcal{C}_1$ is represented by

$$\text{Poi}(x|\lambda_1) = e^{-\lambda_1}\frac{\lambda_1^x}{x!}.$$

Derive the optimal decision boundary x^* analytically as a function of λ_1 and λ_2. The prior for $\mathcal{C}_1$ is given as $p(\mathcal{C}_1) = 0.2$.

Solution:

At the optimal decision boundary x^*, it holds:

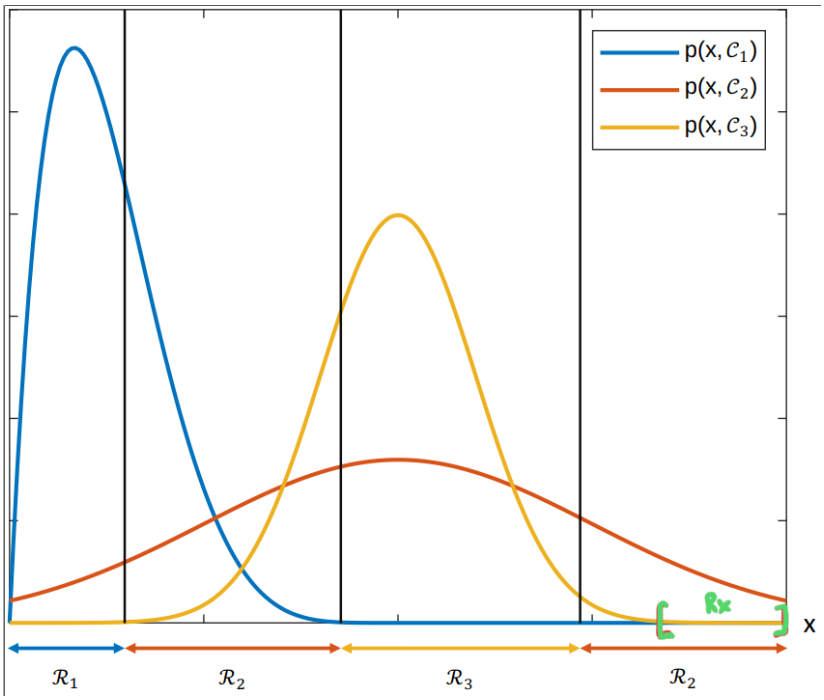


Figure 2: Solution to 1.13.1

$$
\begin{aligned}
p\left(C_1 \mid x^*\right) &= p\left(C_2 \mid x^*\right) \\
p\left(x^* \mid C_1\right) p\left(C_1\right) &= p\left(x^* \mid C_2\right) p\left(C_2\right) \\
p\left(x^* \mid C_1\right) 0,2 &= p\left(x^* \mid C_2\right) p\left(C_2\right) \\
p\left(x^* \mid \lambda_1\right) 0,2 &= p\left(x^* \mid \lambda_2\right) p\left(C_2\right) \\
e^{-\lambda_1} \frac{\lambda_1^{x^*}}{x^*!} 0,2 &= e^{-\lambda_2} \frac{\lambda_2^{x^*}}{x^*!} p\left(C_2\right) \\
\frac{0,2\left(e^{-\lambda_1+\lambda_2}\right)}{p\left(C_2\right)} &= \left(\frac{\lambda_2}{\lambda_1}\right)^{x^*} \\
x^* &= \log_{\frac{\lambda_2}{\lambda_1}}\left(\frac{0,2e^{-\lambda_1+\lambda_2}}{p\left(C_2\right)}\right) \\
x^* &= \frac{\ln\left(\frac{0,2e^{-\lambda_1+\lambda_2}}{p(C_2)}\right)}{\ln\left(\frac{\lambda_2}{\lambda_1}\right)} \\
x^* &= \frac{\ln\left(\frac{0.2}{p(C_2)}\right)+(-\lambda_1+\lambda_2)}{\ln\left(\frac{\lambda_2}{\lambda_1}\right)} \\
x^* &= \frac{\ln(0.2)-\ln(p(C_2))+(-\lambda_1+\lambda_2)}{\ln\left(\lambda_2\right)-\ln\left(\lambda_1\right)}
\end{aligned}
$$

With $p(C_2)$ = 1 - $p(C_1)$ = 0.8 we achieve:

$x^* = \frac{-\ln(4)+(-\lambda_1+\lambda_2)}{\ln(\lambda_2)-\ln(\lambda_1)}$

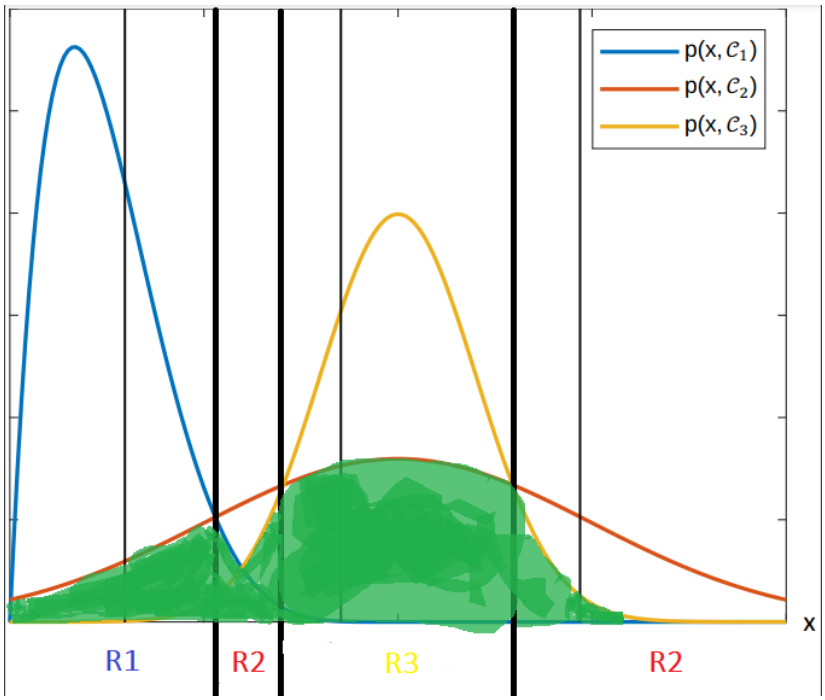


Figure 3: Solution to 1.13.2

Task 1.15: Risk Minimization

Given is a classification problem of N classes, $C = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_N\}$. We additionally introduce the option to assign a sample x to neither of the classes which is known as $\mathcal{C}_{\text{rej}}$. When the rejection risk is lower than the risk of assignment to each class $\mathcal{C}_k \in C$, rejection can be a desirable action. Let the true class of a sample x be C_k and the class it gets assigned to during classification be C_j. For any $k \in \{1, 2, ..., N\}$ and $j \in \{1, 2, ..., N+1\}$, the loss is defined as

$$\lambda_{jk} = \begin{cases} 0 & \text{, if} \quad j = k \\ \text{l}_{\text{reject}} & \text{, if} \quad j = N+1 \\ \text{l}_{\text{misclassified}} & \text{, otherwise} \end{cases}$$

1.15a)

Derive the decision criterion that will give the minimum expected loss.
Hint:

$$\text{classify}(x) \rightarrow \begin{cases} \mathcal{C}_k & \text{, if condition 1} \wedge \text{condition 2} \\ \mathcal{C}_{\text{rej}} & \text{, otherwise} \end{cases}$$

Solution:

To keep the notation for this subtask to be clear and concise, we define the following abbreviations:
α_k: The action of classify a data point x to class $\mathcal{C}_k$ with $k \in \{1, 2, ..., N\}$
α_{rej}: The action of rejecting to classify a data point x to any class $k \in \{1, 2, ..., N\}$
$R(\alpha_k|x)$: The total loss associated with the action α_k after seeing the data point x
$R(\alpha_{rej}|x)$: The total loss associated with the action α_k after seeing the data point x

In order for a data point x to be classified to class $\mathcal{C}_k$, the two following conditions should hold:

(1) Within the classes $\mathcal{C}_i$ with $i \in \{1, 2, ..., N\}$, the loss of choosing class $\mathcal{C}_k$ should be smallest. This is equivalent to

$$k = argmin_i R(\alpha_i|x)$$

(Condition 1)

(2) The loss of choosing k must be smaller than that of choosing the rejection option. This is equivalent to

$$R(\alpha_k|x) < R(\alpha_{rej}|x)$$

(Condition 2)

Next, we will follow by examing the two mentioned conditions in further details to derive the specific decision criterion

For condition 1:

For any $i \in \{1, 2, ..., N\}$:

$$R(\alpha_i|x) = \lambda_{i1}p(\mathcal{C}_1|x) + \lambda_{i2}p(\mathcal{C}_2|x) + ... + \lambda_{iN}p(\mathcal{C}_N|x)$$

$$= l_{misclassied}p(\mathcal{C}_1|x)+l_{misclassifed}p(\mathcal{C}_2|x)+...+l_{misclassifed}p(\mathcal{C}_{i-1}|x)+0+l_{misclassifed}p(\mathcal{C}_{i+1}|x)+...+l_{misclassifed}p(\mathcal{C}_N|x)$$

$$= l_{misclassified}(p(\mathcal{C}_1|x) + p(\mathcal{C}_2|x) + ... + p(\mathcal{C}_N|x) - p(\mathcal{C}_i|x))$$

$$= l_{misclassifed}(1 - p(\mathcal{C}_i|x))$$

(1.15.a.1)

The condition 1 means for any $j \in \{1, 2, ..., N\}$, we have

$$R(\alpha_k|x) < R(\alpha_j|x)$$

Using the derivation in (1.15.a.1), we have for any $j \in \{1, 2, ..., N\}$:

$$R(\alpha_k|x) < R(\alpha_j|x)$$

$$l_{misclassified}(1 - p(\mathcal{C}_k|x)) < l_{misclassified}(1 - p(\mathcal{C}_j|x))$$

$$p(\mathcal{C}_k|x) > p(\mathcal{C}_j|x)$$

(Condition 1 - result)

For condition 2:

$$R(\alpha_{rej}|x) = \alpha_{rej1}p(\mathcal{C}_1|x) + \alpha_{rej2}p(\mathcal{C}_2|x) + ... + \alpha_{rejN}p(\mathcal{C}_N|x)$$

$$= l_{reject}(p(\mathcal{C}_1|x) + p(\mathcal{C}_2|x) + ... + p(\mathcal{C}_N|x))$$

$$= l_{reject}$$

Since $R(\alpha_k|x) < R(\alpha_{rej}|x)$, we then have:

$$l_{misclassified}(1 - p(\mathcal{C}_k|x)) < l_{reject}$$

$$p(\mathcal{C}_k|x) > \frac{l_{misclassified} - l_{reject}}{l_{misclassified}}$$

(Condition 2 - result)

If the condition 1 and 2 can't be hold, then we should reject the classification.

1.15b)

What happens if $l_{reject} = 0$?

Long Nguyen 2451013-Minh Hieu Le 2673639 -Minh Vu Pham 2746346

Solution:

When $l_{reject} = 0$, the result of condition 2 above will return:

$$p(\mathcal{C}_k|x) > 1$$

, which can never be satisfied. Therefore every data point will be rejected.

1.15c)

What happens if $l_{reject} > l_{misclassified}$?

Solution:

When $l_{reject} > l_{misclassified}$, the result of condition 2 above will always be satisfied. In order for a data point to be classified as element of one of the classes $\mathcal{C}_i$, it only has to satisfy the condition 1. Since there always exists a class among N classes that produce the least loss, no data point will be rejected.