Machine Learning Homework 1

Gausse Mael DONGMO KENFACK (董沫高斯)
Student ID: 2024403346

November 2024

Contents

Introduction

In modern machine learning, the need to handle complex, non-linear data structures has become increasingly important. Kernel methods play a critical role in this domain by enabling data transformations into higher-dimensional spaces, where linear algorithms can be applied to effectively separate and classify data. In this work, we aim to deepen our understanding of these methods by closely examining the mathematics underlying Support Vector Machines (SVMs). Additionally, we will investigate properties of the exponential family of probability distributions, maximum likelihood estimators, and the multivariate Gaussian distribution. By following the structured exercises in this assignment, we aim to gain a solid grounding in the theoretical foundations of these essential machine learning concepts.

1 Kernel Methods

Definition: We say a mapping $k : X \times X \to \mathbb{R}$ is a ***kernel*** if there exist a space F with inner product $\langle \cdot, \cdot \rangle$, and a ***feature map*** $\phi : X \to F$ such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$

1.1 Problem 1

In this section, we will determine whether a given map qualifies as a kernel by proving or disproving that it meets the required conditions.

1. Let's prove that $k(x, y) = (1 + xy)^n$ is a kernel on $X = \mathbb{R}$

   To do so, we just need to find a map ϕ and the space F that verify the conditions.

$$\begin{aligned} k(x,y) &= (1+xy)^n \\ &= \sum_{i=0}^{n} \binom{n}{i} 1^{n-i} (xy)^i \quad \text{(Newton binomial)} \\ &= \sum_{i=0}^{n} \binom{n}{i} (xy)^i \end{aligned}$$

   Now that we have this expression of $k(x, y)$ as a sum of products, it's easier to find ϕ. let's define the feature map

$$\begin{aligned} \phi : \mathbb{R} &\to \mathbb{R}^{n+1} \\ x &\mapsto (1, \sqrt{\binom{n}{1}}x, \sqrt{\binom{n}{2}}x^2, \ldots, \sqrt{\binom{n}{n}}x^n)^\top \end{aligned}$$

   One can compute $\langle \phi(x), \phi(y) \rangle$

$$\begin{aligned} \langle \phi(x), \phi(y) \rangle &= (1, \sqrt{\binom{n}{1}}x, \sqrt{\binom{n}{2}}x^2, \ldots, \sqrt{\binom{n}{n}}x^n)(1, \sqrt{\binom{n}{1}}y, \sqrt{\binom{n}{2}}y^2, \ldots, \sqrt{\binom{n}{n}}y^n)^\top \\ &= \sum_{i=0}^{n} (\sqrt{\binom{n}{i}}x^i)(\sqrt{\binom{n}{i}}y^i) \\ &= \sum_{i=0}^{n} \binom{n}{i} (xy)^i \\ &= k(x,y) \end{aligned}$$

   Therefore, we can conclude that $k(x, y) = (1 + xy)^n$ is a kernel on $X = \mathbb{R}$

2. Let's prove by contradiction that $k(x, y) = xy - 1$ is not a kernel on $X = \mathbb{R}$

   Suppose k is a kernel on $\mathbb{R}$, from the Mercer's theorem, we can deduce that k is a semi-positive definite symmetric function. This means it corresponds to a semi-positive definite symmetric Gram matrix. Therefore for any set of points $\{x_1, x_2, ..., x_n\} \subset \mathbb{R}$ and for any vector $v = (v_1, v_2, \dots, v_n)^\top \in \mathbb{R}^n$ we have $v^\top K v \geq 0$ with

$$K = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \dots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \dots & k(x_1,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \dots & k(x_n,x_n) \end{pmatrix}$$

For the set of points $\{0,1\}$ and the vector $v = (1,1)^\top$ we have

$$K = \begin{pmatrix} -1 & -1 \\ -1 & 0 \end{pmatrix} \quad \text{and} \quad v^\top K v = (1,1)K(1,1)^\top = -3 \leq 0$$

We found our contradiction, we can now conclude that $k(x,y) = xy - 1$ is not a kernel on $X = \mathbb{R}$

3. Let's prove that $k(x,y) = min(x,y)$ is a kernel on $X = [0,1]$

   To find the map ϕ corresponding, let's try to rewrite $k(x,y)$

$$\begin{aligned} k(x,y) &= min(x,y) \\ &= \int_0^1 \mathbf{1}_{[0,x]}(t)\mathbf{1}_{[0,y]}(t)dt \end{aligned}$$

where $\mathbf{1}_{[0,x]}(t)$ is the indicator function [[3][1]] of interval $[0,x]$. We can then define the feature map as

$$\begin{aligned} \phi : [0,1] &\to \mathbf{L}^2([0,1]) \\ x &\mapsto \mathbf{1}_{[0,x]}(t) \end{aligned}$$

$$\begin{aligned} \langle \phi(x), \phi(y) \rangle &= \int_0^1 \mathbf{1}_{[0,x]}(t)\mathbf{1}_{[0,y]}(t)dt \\ &= \int_0^{min(x,y)} dt \\ &= min(x,y) \\ &= k(x,y) \end{aligned}$$

One can now conclude that $k(x,y) = min(x,y)$ is a kernel on $X = [0,1]$

1.2 Problem 2

The goal of this section is to use SVM and kernel methods to solve a classification problem. Given a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^m$ be a feature map. Consider the following Soft SVM problem:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^m, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^N \xi_i \qquad (1.1) \\ \text{s.t.} \quad & \xi_i \geq 0, \quad y_i \mathbf{w}^\top \phi(\mathbf{x}_i) \geq 1 - \xi_i. \end{aligned}$$

1. The hinge loss is used in the formulation of (1.1) because the 0-1 loss function is not convex in $y\mathbf{w}^\top\mathbf{x}$, making it challenging to optimize. By contrast, hinge loss is convex, which brings several mathematical and computational advantages: it guarantees a unique global optimum that can be computed efficiently with standard optimization algorithms. This convexity makes hinge loss particularly attractive for SVM and other models where efficient and reliable optimization is crucial.
   Another feasible function to approximate the 0-1 loss is the quadratic loss define as :

   $$\ell_{quad}(yf(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))^2$$

   It's a smoother convex function and as well as hinge loss it upperbounds 0-1 loss.

2. We want to do an optimization under constraints, to solve this problem, we ar looking for $\hat{\mathbf{w}}$ verifying

   $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\boldsymbol{\xi}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{N} \xi_i$$
   $$\text{s.t.} \quad \xi_i \geq 0, \quad y_i\mathbf{w}^\top\phi(\mathbf{x}_i) \geq 1 - \xi_i.$$

   solving this is equivalent to solve this

   $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$
   $$\text{s.t.} \quad \xi_i \geq 0, \quad y_i\mathbf{w}^\top\phi(\mathbf{x}_i) \geq 1 - \xi_i.$$
   $$\text{with} \quad C = \frac{1}{\lambda}$$

   we can compute the Lagrangian of (1.1). To do so, we introduce $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ our Lagrange multipliers

   - $\boldsymbol{\alpha} \geq 0$ for the second inequality constraint
   - $\boldsymbol{\beta} \geq 0$ for the first inequality constraint

   with these multipliers, we have this Lagrangian

   $$\begin{aligned} L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\beta_i\xi_i - \sum_{i=1}^{N}\alpha_i(y_i\mathbf{w}^\top\phi(\mathbf{x}_i) - 1 + \xi_i) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\beta_i\xi_i - \sum_{i=1}^{N}\alpha_i\xi_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i y_i\mathbf{w}^\top\phi(\mathbf{x}_i) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 + C\boldsymbol{\xi}^\top\mathbf{1} - \boldsymbol{\xi}^\top(\boldsymbol{\alpha} + \boldsymbol{\beta}) + \boldsymbol{\alpha}^\top\mathbf{1} - \sum_{i=1}^{N}\alpha_i y_i\mathbf{w}^\top\phi(\mathbf{x}_i) \end{aligned}$$

3. Now that we got the Lagrangian we can minimize it to get the dual problem.

   - Taking the derivative with respect to $\mathbf{w}$, and setting it to 0, we have

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i) = 0 \\ \implies \mathbf{w} &= \sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i)\end{aligned}$$

   - Taking the derivative with respect of $\boldsymbol{\xi}$ and setting it to 0, we have

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\xi}} &= C\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta} = 0 \\ \implies \boldsymbol{\beta} &= C\mathbf{1} - \boldsymbol{\alpha} \geq 0 \\ \implies 0 &\leq \boldsymbol{\alpha} \leq C\mathbf{1}\end{aligned}$$

   - Substituting obtained values in the Lagrangian gives us the dual problem

$$L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i)\|^2 + C\boldsymbol{\xi}^\top \mathbf{1} - \boldsymbol{\xi}^\top(\boldsymbol{\alpha} + C\mathbf{1} - \boldsymbol{\alpha}) + \boldsymbol{\alpha}^\top \mathbf{1} - \sum_{i=1}^{N} \alpha_i y_i (\sum_{j=1}^{N} \alpha_j y_j \phi(\mathbf{x}_j))^\top \phi(\mathbf{x}_i)$$

$$\begin{aligned}L(\boldsymbol{\alpha}) &= \boldsymbol{\alpha}^\top \mathbf{1} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ &= \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

$$\begin{aligned}L(\boldsymbol{\alpha}) &= \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{Y} G \boldsymbol{Y} \boldsymbol{\alpha} \\ \text{with} \quad \boldsymbol{Y} &= diag(y_1, \ldots, y_N), \quad G_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad 0 &\leq \boldsymbol{\alpha} \leq C\mathbf{1}\end{aligned}$$

4. by solving the dual problem we will be able to express the prediction $f(\mathbf{x})$

$$\hat{\alpha} = \underset{0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}}{\arg\max}\, \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{Y} G \boldsymbol{Y} \boldsymbol{\alpha} \implies \hat{\mathbf{w}} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \phi(\mathbf{x}_i)$$

$$\begin{aligned}f(\mathbf{x}) &= sign(\hat{\mathbf{w}}^\top \phi(\mathbf{x}) \\ &= sign((\sum_{i=1}^{N} \hat{\alpha}_i y_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x})) \\ &= sign(\sum_{i=1}^{N} \hat{\alpha}_i y_i k(\mathbf{x}_i, \mathbf{x}))\end{aligned}$$

2 Exponential Families

In this section we take a look a the exponential family

$$p(\mathbf{x}|\eta) = h(\mathbf{x}) \exp(\eta^\top T(\mathbf{x}) - A(\eta))$$

with $T(\mathbf{x})$ a sufficient statistic and $A(\eta) = \log \int h(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})} d\mathbf{x}$ is the partition function. Let's compute the derivatives of $A(\eta)$

- first derivative of $A(\eta)$

  first let's expand $p(\mathbf{x}|\eta)$

$$\begin{aligned} p(\mathbf{x}|\eta) &= h(\mathbf{x}) \exp(\eta^\top T(\mathbf{x}) - A(\eta)) && (1) \\ &= h(\mathbf{x}) \exp(\eta^\top T(\mathbf{x})) \frac{1}{\exp(A(\eta))} && (2) \\ &= \frac{h(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})}}{\int h(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})} d\mathbf{x}} && (3) \end{aligned}$$

$$\begin{aligned} \frac{\partial A(\eta)}{\partial \eta_i} &= \frac{\int h(\mathbf{x}) T_i(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})} d\mathbf{x}}{\int h(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})} d\mathbf{x}} \qquad ((log(u))' = \frac{u'}{u}) \\ &= \int T_i(\mathbf{x}) \frac{h(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})} d\mathbf{x}}{\int h(\mathbf{x}) \mathrm{e}^{\eta^\top T(\mathbf{x})} d\mathbf{x}} \quad \text{we can replace the expression from equation (3)} \\ &= \int T_i(\mathbf{x}) p(\mathbf{x}|\eta) d\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{x}|\eta)}[T_i(\mathbf{x})] \end{aligned}$$

- now the second derivative

  first let's compute the derivative of $p(\mathbf{x}|\eta)$

$$\frac{\partial}{\partial \eta_j} p(\mathbf{x}|\eta) = p(\mathbf{x}|\eta)(T_j(\mathbf{x}) - \frac{\partial}{\partial \eta_j} A(\eta)) \qquad (4)$$

$$\begin{aligned} \frac{\partial A(\eta)}{\partial \eta_i \partial \eta_j} &= \frac{\partial}{\partial \eta_j} \mathbb{E}_{p(\mathbf{x}|\eta)}[T_i(\mathbf{x})] \\ &= \int T_i(\mathbf{x}) \frac{\partial}{\partial \eta_j} p(\mathbf{x}|\eta) d\mathbf{x} \quad \text{we can replace with the expression from equation (4)} \\ &= \int T_i(\mathbf{x}) p(\mathbf{x}|\eta)(T_j(\mathbf{x}) - \frac{\partial}{\partial \eta_j} A(\eta)) d\mathbf{x} \\ &= \int T_i(\mathbf{x}) p(\mathbf{x}|\eta) T_j(\mathbf{x}) - T_i(\mathbf{x}) p(\mathbf{x}|\eta) \frac{\partial}{\partial \eta_j} A(\eta) d\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{x}|\eta)}[T_i(\mathbf{x}) T_j(\mathbf{x})] - \mathbb{E}_{p(\mathbf{x}|\eta)}[T_i(\mathbf{x})] \mathbb{E}_{p(\mathbf{x}|\eta)}[T_j(\mathbf{x})] \\ &= Cov_{p(\mathbf{x}|\eta)}[T_i(\mathbf{x}), T_j(\mathbf{x})] \end{aligned}$$

3 Maximum Likelihood Estimators

In this section we consider the maximum likelihood estimation of the multivariate Gaussian distribution and its convergence properties. The density function of the d-dimensional multivariate Gaussian distribution with mean μ and covariance Σ is

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Given i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ from $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where μ and Σ are unknown parameters.

1. One can find the maximum likelihood estimators (MLE) $\hat{\mu}_{ML}$ and $\hat{\Sigma}_{ML}$. Since the x_i are i.i.d we can write the log-likelihood of the joint probability as

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_i) &= \log \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \log \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right) \\ &= \sum_{i=1}^{N} \left(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right) \\ l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}). \end{aligned}$$

- Deriving $\hat{\boldsymbol{\mu}}_{ML}$ [4]

$$\frac{\partial}{\partial \boldsymbol{\mu}} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_i) = \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{x}_i) = 0$$

Since $\boldsymbol{\Sigma}$ is positive definite,

$$0 = N\boldsymbol{\mu} - \sum_{i=1}^{N} \mathbf{x}_i$$

$$\implies \hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = \bar{\mathbf{x}}$$

- Deriving $\hat{\boldsymbol{\Sigma}}_{ML}$ [4]

$$\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_i) = \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top$$

Setting to zero and solving for $\boldsymbol{\Sigma}$

$$0 = N\boldsymbol{\Sigma} - \sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top$$

$$\implies \hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\boldsymbol{\mu}})(\mathbf{x}_i-\hat{\boldsymbol{\mu}})^\top$$

2. After finding the estimators, one can check if they are unbiased by computing the expectations $\mathbb{E}[\hat{\boldsymbol{\mu}}_{ML}]$ and $\mathbb{E}[\hat{\boldsymbol{\Sigma}}_{ML}]$

   - $\mathbb{E}[\hat{\boldsymbol{\mu}}_{ML}]$

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\mu}}_{ML}] &= \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathbf{x}_i] \\ &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu} \\ &= \boldsymbol{\mu}\end{aligned}$$

   Thus $\hat{\boldsymbol{\mu}}_{ML}$ is **unbiased**

   - $\mathbb{E}[\hat{\boldsymbol{\Sigma}}_{ML}]$

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\Sigma}}_{ML}] &= \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top] \\ &= \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top - (\mathbf{x}_i - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top][2] \\ &= \boldsymbol{\Sigma} - 0 - 0 - \frac{1}{N}\boldsymbol{\Sigma} \\ &= \frac{N-1}{N}\boldsymbol{\Sigma}\end{aligned}$$

   Thus $\hat{\boldsymbol{\Sigma}}_{ML}$ is **biased**

3. To conclude this section, let's prove that

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu}\|^2\right] = \frac{\mathrm{Tr}(\Sigma)}{N}.$$

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu}\|^2 &= (\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu})^\top(\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu}) \\ \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu}\|^2\right] &= \mathbb{E}\left[(\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu})^\top(\hat{\boldsymbol{\mu}}_{\mathrm{ML}} - \boldsymbol{\mu})\right] \\ &= \mathrm{Tr}\left(\frac{\boldsymbol{\Sigma}}{N}\right) = \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{N}\end{aligned}$$

Conclusion

In this work, we explored the mathematical foundations and applications of kernel methods, demonstrating their significance in handling complex data structures in machine learning. The examination of the exponential family provided further insights into the statistical foundations of probabilistic models. Kernel methods, particularly in the context of SVMs, exemplify how theoretical rigor and practical implementation can converge to solve real-world problems. Finally, This homework was the perfect exercise to have a better understanding of SVM and estimators.

References

[1] *ChatGPT Prompt 1.* https://chatgpt.com/share/672efd10-bccc-8013-86b3-dd33f71f5a9e. [Online].

[2] *ChatGPT Prompt 2.* https://chatgpt.com/share/673094e7-3df0-8013-8deb-dd2e12c83929. [Online].

[3] *Indicator Function.* https://en.wikipedia.org/wiki/Indicator_function. [Online].

[4] Michael I. Jordan. *The Multivariate Gaussian.* https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter13.pdf. [Online].