INFR10086 Machine Learning (MLG) Semester 2, 2025/6

Exercises 2

Lecturer: Hao Tang

Exercise 1. When we solve a maximum likelihood problem, we tend to take the log and maximize the log likelihood. The logorithm function $f(x) = \log(x)$ is a monotonic function. Specifically, if $x \geq x'$ then $f(x) \geq f(x')$ for any x and x'. Use the monotonicity and show that

$$\operatorname*{argmax}_{\mathbf{w}} L(\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \log L(\mathbf{w}), \tag{1}$$

where L is a likelihood function of $\mathbf{w}$.

Exercise 2. Show that $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = (A + A^\top)\mathbf{x}$.

Exercise 3. Show that the Hessian of $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$ is $A + A^\top$.

Exercise 4. Given a data set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, show that the log likelihood of the Gaussian mean

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{n} \left[-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \tag{2}$$

has a Hessian $-n\Sigma^{-1}$ and conclude that the log likelihood is concave in μ.

Exercise 5. Given a data set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, show that

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^n. \tag{3}$$

is the maximum likelihood estimator of the Gaussian mean

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{n} \left[-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]. \tag{4}$$

In other words,

$$\hat{\boldsymbol{\mu}} = \operatorname*{argmax}_{\boldsymbol{\mu}} L(\boldsymbol{\mu}). \tag{5}$$

Exercise 6. Show that if f is convex, then $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ is also convex.

Exercise 7. Show that if $f_1, f_2, \ldots, f_k$ is convex, then $f = \beta_1 f_1 + \cdots + \beta_k f_k$ is also convex for $\beta_1, \ldots, \beta_k \geq 0$.

Exercise 8. In this question, we will solve the Lagrangian to study the following problem

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top X^\top X \mathbf{w}}{\mathbf{w}^\top \mathbf{w}}. \tag{6}$$

This is actually the first principle component in principal component analysis (PCA), and we will say more about this in class.

1. First, show that

$$F(\mathbf{w}) = \frac{\mathbf{w}^\top X^\top X \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \tag{7}$$

is invariant to scaling. In other words, $F(\alpha \mathbf{w}) = F(\mathbf{w})$ for any $\alpha \neq 0$.

2. Given that F is invariant to scaling, we can rewrite the problem

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top X^\top X \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \tag{8}$$

into

$$\max_{\mathbf{w}} \quad \mathbf{w}^\top X^\top X \mathbf{w} \tag{9}$$

$$\text{s.t.} \quad \|\mathbf{w}\|^2 = 1 \tag{10}$$

The constraint $\|\mathbf{w}\|^2 = 1$ might seem unnecessary, given that F is invariant to scaling. What happens if we remove the constraint? Argue why the constraint is necessary.

3. Show that the Lagrangian of the problem

$$\max_{\mathbf{w}} \quad \mathbf{w}^\top X^\top X \mathbf{w} \tag{11}$$

$$\text{s.t.} \quad \|\mathbf{w}\|^2 = 1 \tag{12}$$

is

$$L(\mathbf{w}, \lambda) = \mathbf{w}^\top X^\top X \mathbf{w} + \lambda(1 - \|\mathbf{w}\|^2) \tag{13}$$

for $\lambda \geq 0$.

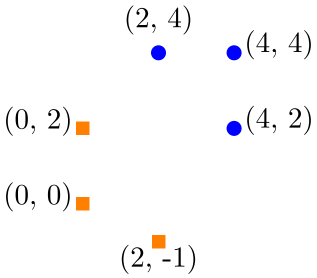
4. Show that the solution of the Lagrangian is

$$(X^\top X)\mathbf{w} = \lambda \mathbf{w}. \tag{14}$$

In particular, λ is an eigenvalue of $X^\top X$.

5. Now plugging $(X^\top X)\mathbf{w} = \lambda \mathbf{w}$ back to the objective $\mathbf{w}^\top X^\top X \mathbf{w}$, and show that λ is actually the *largest* eigenvalue of $X^\top X$.

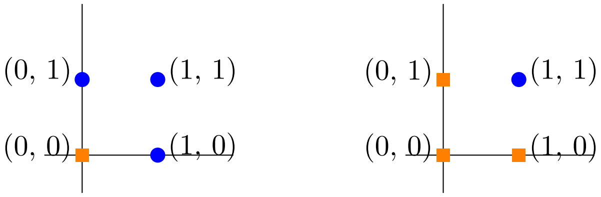
Exercise 9. Suppose we have the following 2D data set, where the blue points (●) are class +1 and the orange points (■) are class -1.



If we train a support vector machine on this data set, what is the decision boundary and what is the margin?

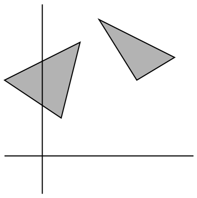
Exercise 10. Given the kernel $k(\mathbf{a}, \mathbf{b}) = (\mathbf{a}^\top \mathbf{b})^2$, what is the feature function ϕ such that $k(\mathbf{a}, \mathbf{b}) = \phi(\mathbf{a})^\top \phi(\mathbf{b})$?

Exercise 11. Suppose we have the following 2D data sets, where the blue points (●) are class +1 and the orange points (■) are class -1.



In each data set, find a decision boundary that separate the two classes.

Exercise 12. Design a multi-layer perceptron to classify the gray region in the following 2D figure. The input are the 2D coordinates, and the output is binary, whether it's inside the gray region or not.



How many nodes do you need in each layer? Explain your answer.