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 $\mathbf{w}$.

If $\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} L(\mathbf{w})$, then by the definition of maximum, we have

$$L(\hat{\mathbf{w}}) \geq L(\mathbf{w}) \qquad \text{for any } \mathbf{w}. \tag{2}$$

Using the monotonicity of log, we have

$$\log L(\hat{\mathbf{w}}) \geq \log L(\mathbf{w}) \qquad \text{for any } \mathbf{w}. \tag{3}$$

Based on the definition of maximum again, $\hat{\mathbf{w}}$ is one of the maximal solutions of $\log L(\mathbf{w})$.

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

We first expand the vector-matrix multiplication into

$$\mathbf{x}^\top A \mathbf{x} = \sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j a_{ij} \tag{4}$$

$$= \sum_{\substack{j=1 \\ j \neq k}}^{d} x_k x_j a_{kj} + \sum_{\substack{i=1 \\ i \neq k}}^{d} x_i x_k a_{ik} + x_k x_k a_{kk} + \sum_{i \neq k} \sum_{j \neq k} x_i x_j a_{ij} \tag{5}$$

for some arbitrary k. It is now clear that if we take the derivative, we have

$$\frac{\partial}{\partial x_k} \mathbf{x}^\top A \mathbf{x} = \sum_{\substack{j=1 \\ j \neq k}}^{d} x_j a_{kj} + \sum_{\substack{i=1 \\ i \neq k}}^{d} x_i a_{ik} + 2 x_k a_{kk} \tag{6}$$

$$= \sum_{j=1}^{d} x_j a_{kj} + \sum_{i=1}^{d} x_i a_{ik}. \tag{7}$$

We can stack everything together into the vector form to get $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top A \mathbf{x} = (A + A^\top)\mathbf{x}$.

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

From the previous question, we know that

$$\frac{\partial}{\partial x_k}\mathbf{x}^\top A\mathbf{x} = \sum_{j=1}^{d} x_j a_{kj} + \sum_{i=1}^{d} x_i a_{ik}. \tag{8}$$

We can take another derivative to have

$$\frac{\partial}{\partial x_{k'}}\frac{\partial}{\partial x_k}\mathbf{x}^\top A\mathbf{x} = a_{kk'} + a_{k'k}. \tag{9}$$

From there, it is clear that the Hessian is $A + A^\top$.

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

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{n}\left[-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right] \tag{10}$$

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

We can expand the log likelihood into

$$L(\boldsymbol{\mu}) = -\frac{n}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu} + \ldots, \tag{11}$$

and it is clear that the Hessian of the log likelihood is

$$-\frac{n}{2}\left(\Sigma^{-1} + \Sigma^{-1^\top}\right) = -n\Sigma^{-1}. \tag{12}$$

The covariance matrix is positive semidefinite by definition, and its inverse is also positive semidefinite.[1] Because of the negative sign, the negative log likelihood has a positive semidefinite Hessian and is thus convex. We can conclude that the log likelihood is concave.

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

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^n. \tag{13}$$

is the maximum likelihood estimator of the Gaussian mean

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{n}\left[-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]. \tag{14}$$

[1] A positive semidefinite matrix always has nonnegative eigenvalues. Since the inverse of a matrix has the reciprical of its eigenvalues, the inverse of a positive semidefinite matrix also has nonnegative eigenvalues and is also positive semidefinite.

In other words,

$$\hat{\boldsymbol{\mu}} = \operatorname*{argmax}_{\boldsymbol{\mu}} L(\boldsymbol{\mu}). \tag{15}$$

We expand the log likelihood into

$$L(\boldsymbol{\mu}) = -\sum_{i=1}^{n} \frac{1}{2}\left[\mathbf{x}^\top \Sigma^{-1}\mathbf{x} - 2\mathbf{x}^\top\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\right] + \dots \tag{16}$$

Since we have already shown the the log likelihood is concave, the point where gradient is zero is the optimal solution. We take the gradient of the log likelihood and get

$$\nabla_{\boldsymbol{\mu}} L(\boldsymbol{\mu}) = -\sum_{i=1}^{n} \frac{1}{2}(-2\Sigma^{-1}\mathbf{x}_i + (\Sigma^{-1} + \Sigma^{-1^\top})\boldsymbol{\mu}) = \Sigma^{-1}\sum_{i=1}^{n}\mathbf{x}_i - n\Sigma^{-1}\boldsymbol{\mu}. \tag{17}$$

Solving $\nabla_{\boldsymbol{\mu}} L(\boldsymbol{\mu}) = 0$, we have

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^n. \tag{18}$$

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

This can be shown by the definition of convexity.

$$\begin{aligned} g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= f(A(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) + \mathbf{b}) && (19)\\ &= f(\alpha(A\mathbf{x} + \mathbf{b}) + (1-\alpha)(A\mathbf{y} + \mathbf{b})) && (20)\\ &\le \alpha f(A\mathbf{x} + \mathbf{b}) + (1-\alpha) f(A\mathbf{y} + \mathbf{b}) && (21)\\ &= \alpha g(\mathbf{x}) + (1-\alpha) g(\mathbf{y}) && (22) \end{aligned}$$

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

This can be shown by the definition of convexity.

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \beta_1 f_1(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) + \cdots + \beta_k f_k(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) && (23)\\ &\le \beta_1[\alpha f_1(\mathbf{x}) + (1-\alpha) f_1(\mathbf{y})] + \cdots + \beta_k[\alpha f_k(\mathbf{x}) + (1-\alpha) f_k(\mathbf{y})] && (24)\\ &= \alpha(\beta_1 f_1(\mathbf{x}) + \cdots + \beta_k f_k(\mathbf{x})) + (1-\alpha)(\beta_1 f_1(\mathbf{y}) + \cdots + \beta_k f_k(\mathbf{y})) && (25)\\ &= \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) && (26) \end{aligned}$$

Note that multiplying a negative number flips the direction of the inequality, so $\beta_1, \dots, \beta_k$ have to be nonnegative.

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

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top X^\top X \mathbf{w}}{\mathbf{w}^\top \mathbf{w}}. \tag{27}$$

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

$$F(\mathbf{w}) = \frac{\mathbf{w}^\top X^\top X \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \tag{28}$$

   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

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top X^\top X \mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \tag{29}$$

   into

$$\max_{\mathbf{w}} \quad \mathbf{w}^\top X^\top X \mathbf{w} \tag{30}$$
$$\text{s.t.} \quad \|\mathbf{w}\|^2 = 1 \tag{31}$$

   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

$$\max_{\mathbf{w}} \quad \mathbf{w}^\top X^\top X \mathbf{w} \tag{32}$$
$$\text{s.t.} \quad \|\mathbf{w}\|^2 = 1 \tag{33}$$

   is

$$L(\mathbf{w}, \lambda) = \mathbf{w}^\top X^\top X \mathbf{w} + \lambda(1 - \|\mathbf{w}\|^2) \tag{34}$$

   for $\lambda \geq 0$.

4. Show that the solution of the Lagrangian is

$$(X^\top X)\mathbf{w} = \lambda \mathbf{w}. \tag{35}$$

   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 $X^\top X$.

1. This is quite straightforward from the definition.

$$F(\alpha\mathbf{w}) = \frac{\alpha\mathbf{w}^\top X^\top X \alpha\mathbf{w}}{(\alpha\mathbf{w})^\top(\alpha\mathbf{w})} = \frac{\mathbf{w}^\top X^\top X\mathbf{w}}{\mathbf{w}^\top\mathbf{w}} = F(\mathbf{w}) \tag{36}$$

2. The norm constraint is important in that we can scale $\mathbf{w}$ arbitrary large to make $\mathbf{w}^\top X^\top X\mathbf{w}$ large. We won't be able to do this with the norm constraint.

3. The Lagrangian brings the constraint up to the objective.

$$L(\mathbf{w}, \lambda) = \mathbf{w}^\top X^\top X\mathbf{w} + \lambda(1 - \|\mathbf{w}\|^2) \tag{37}$$

   for some $\lambda \geq 0$.

4. We can find the solution by taking the gradient

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = (X^\top X + (X^\top X)^\top)\mathbf{w} - \lambda(2\mathbf{w}) \tag{38}$$

   and set it to 0. This gives us

$$X^\top X\mathbf{w} = \lambda\mathbf{w}. \tag{39}$$

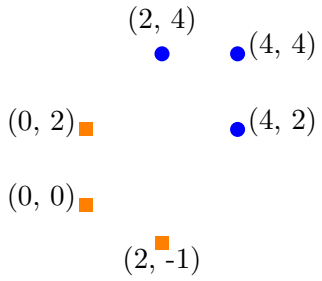
   By definition, λ is an eigenvalue of $X^\top X$.

5. Plugging $X^\top X\mathbf{w} = \lambda\mathbf{w}$ to F, we can see that

$$\frac{\mathbf{w}^\top X^\top X\mathbf{w}}{\mathbf{w}^\top\mathbf{w}} = \frac{\lambda^2\|\mathbf{w}\|^2}{\|\mathbf{w}\|^2} = \lambda^2. \tag{40}$$

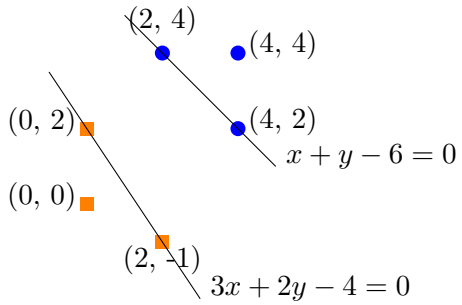
   To maximize this quantity (which is the original goal), we take λ as the largest eigenvalue of $X^\top X$.

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?

From the figure, there are two critical lines, one passes through $(0, 2)$ and $(2, -1)$ and the other passes $(2, 4)$ and $(4, 2)$.



If we choose the line $3x + 2y - 4 = 0$, we need to find the parallel line that passes through $(2, 4)$. We can plug $(2, 4)$ into $3x + 2y - b = 0$ and see that $b = 14$. The decision boundary between $3x + 2y - 4 = 0$ and $3x + 2y - 14 = 0$ is $3x + 2y - 9 = 0$. The distance from $(2, 4)$ to $3x + 2y - 9 = 0$ is $5/\sqrt{13}$. The distances from $(0, 2)$ and $(2, -1)$ to $3x + 2y - 9 = 0$ are all $5/\sqrt{13}$. It's important to check that they all have the same distance.

For the other line $x + y - 6 = 0$, the key point is $(0, 2)$ and it's $x + y - 2 = 0$ that passes through the point. It is clear that the decision boundary is $x + y - 4 = 0$. The distances from $(2, 4)$ to $x + y - 4 = 0$ is $2/\sqrt{2}$. The distances from $(4, 2)$ and $(0, 2)$ to $x + y - 4 = 0$ are all $2/\sqrt{2}$. Because $2/\sqrt{2}$ is the larger margin, our decision boundary is $x + y - 4 = 0$.

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

We can expand the kernel into

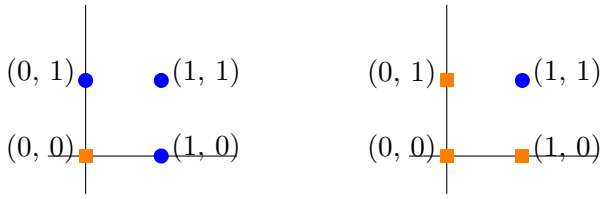
$$k(\mathbf{a}, \mathbf{b}) = (\mathbf{a}^\top \mathbf{b})^2 = \left(\sum_{i=1}^{d} a_i b_i \right) \left(\sum_{j=1}^{d} a_j b_j \right) \tag{41}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{d} a_i a_j b_i b_j = \sum_{i=1}^{d} \sum_{j=1}^{d} (a_i a_j)(b_i b_j). \tag{42}$$

We can now see that a kernel is a dot product between two vectors. The feature vector can be written as

$$\phi(\mathbf{a}) = \begin{bmatrix} a_1 a_1 \\ a_1 a_2 \\ \vdots \\ a_d a_1 \\ a_d a_2 \\ \vdots \\ a_d a_d \end{bmatrix}. \tag{43}$$

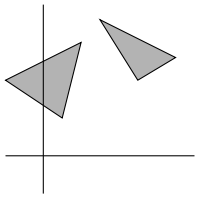
Exercise 11. Suppose we have the following 2D data sets, where the blue points (●) are class +1 and the orange points (■) are class -1.



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

There are in fact infinitely many solutions to this question. Here, I'm only listing the two typical solutions. On the left figure, we can find the line $2x + 2y - 1 = 0$ that passes through $(0, 1/2)$ and $(1/2, 0)$. On the right figure, we can find the line $2x + 2y - 3 = 0$ that passes through $(1/2, 1)$ and $(1, 1/2)$.

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.

There are multiple ways to approach this problem, and I'm listing the obvious one that takes the intersection of half-spaces. The input layer has 2 nodes. Because there are 6 lines we plan to represent, we allocate 6 nodes on the second layer. We will then need to choose the area within each triangle, so we have 2 nodes on the third layer. Finally, on the fourth layer, there is a single node that outputs one class when we are in either of the two triangles. on the second layer.