ML HW5

Date: 10/01/2025
Week: 5
Author: Alvin B. Lin
Student ID: 112652040

Problem 1: Integral of Multivariable Normal Distribution

Given

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)},$$

where $\mathbf{x}, \mu \in \mathbb{R}^k$, $\mathbf{\Sigma}$ is a k-by-k positive definite matrix and $|\mathbf{\Sigma}|$ is its determinant. Show that

$$\int_{\mathbb{R}^k} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1.$$

Solution:

We first do a substitution:

$$\mathbf{y} = \mathbf{x} - \mu; \qquad \Omega = \mathbb{R}^k \Longrightarrow \mathbb{R}^k = \Omega'.$$

We need the following lemma:

Theorem 1: Symmetric $\iff$ Orthogonally Diagonalisable

Given a matrix $\mathbf{A} \in \mathrm{M}_{k\times k}$, then it follows that

$$\mathbf{A} \text{ is symmetric} \iff \mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$$

for some **orthogonal** $\mathbf{Q} \in \mathrm{M}_{k\times k}$, diagonal $\mathbf{D} \in \mathrm{M}_{k\times k}$.

Combine **Theorem 1** and the **positive definite** property on $\mathbf{\Sigma}$, we know that:

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top = \left(\mathbf{Q}\sqrt{\mathbf{D}}\right)\left(\mathbf{Q}\sqrt{\mathbf{D}}\right)^\top := \left(\mathbf{\Sigma}^{1/2}\right)\left(\mathbf{\Sigma}^{1/2}\right)^\top = \left(\mathbf{\Sigma}^{1/2}\right)\left(\mathbf{\Sigma}^{1/2}\right),$$

where $\mathbf{Q}$ is orthogonal, $\mathbf{D}$ is a diagonal matrix with all diagonal entries positive, $\sqrt{\mathbf{D}}^2 = \mathbf{D}$.

So now we have our exponent part becomes:

$$-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) = -\frac{1}{2}\left(\mathbf{\Sigma}^{-1/2}\mathbf{y}\right)^\top \left(\mathbf{\Sigma}^{-1/2}\mathbf{y}\right) = -\frac{1}{2}\left\|\mathbf{\Sigma}^{-1/2}\mathbf{y}\right\|^2$$

By letting $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{y}$, we get $\mathrm{d}\mathbf{z} = \left|\mathbf{\Sigma}^{-1/2}\right|\mathrm{d}\mathbf{y} = \frac{1}{\sqrt{|\mathbf{\Sigma}|}}\mathrm{d}\mathbf{y}$, where $\left|\mathbf{\Sigma}^{-1/2}\right|$ is the **Jacobian**.

Notice that $\|\mathbf{z}\|^2 = z_1^2 + z_2^2 + \cdots + z_k^2$, and by **Fubini's Theorem**, we have:

$$\int_{\mathbb{R}^k} \varphi_1(x_1)\cdot\varphi_2(x_2)\cdots\varphi_k(x_k)\,\mathrm{d}\mathbf{x} = \left(\int_{\mathbb{R}} \varphi_1(x_1)\,\mathrm{d}x_1\right)\left(\int_{\mathbb{R}} \varphi_2(x_2)\,\mathrm{d}x_2\right)\cdots\left(\int_{\mathbb{R}} \varphi_k(x_k)\,\mathrm{d}x_k\right)$$

Also, there is a well-known result that $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x = 1$. We are all set now.

Combine everything together, we get:

$$\begin{aligned}\int_{\mathbb{R}^k} f(\mathbf{x})\ \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{y}\|^2}\ \mathrm{d}\mathbf{y} &= \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k}} e^{-\frac{1}{2}\sum_{i=1}^{k} z_i^2}\ \mathrm{d}\mathbf{z} \\ &= \prod_{i=1}^{k}\left(\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2}\ \mathrm{d}z_i\right) \\ &= 1^k \\ &= 1\end{aligned}$$

Hence proved.

Problem 2: MLE of Multivariate Gaussian

(a) Show that $\dfrac{\partial}{\partial \mathbf{A}} \mathrm{Tr}(\mathbf{AB}) = \mathbf{B}^\top$.

(b) Show that $\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathrm{Tr}(\mathbf{x}\mathbf{x}^\top \mathbf{A})$.

(c) Derive the maximum likelihood estimator for the multivariate Gaussian.

Solution:

(a) Given $\mathbf{M} \in \mathrm{M}_{n\times n}$, recall that the definition of the **trace** is stated as:

$$\mathrm{Tr}(\mathbf{M}) = \sum_{i=1}^{n} m_{ii}.$$

Also, for matrices $\mathbf{A} \in \mathrm{M}_{n\times m}$ and $\mathbf{B} \in \mathrm{M}_{m\times n}$, the matrix multiplication gives:

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj} \qquad\qquad \mathrm{Tr}(\mathbf{AB}) = \sum_{i=1}^{n}\sum_{k=1}^{m} a_{ik}b_{ki}$$

Since all the entries of $\mathbf{A}$ are pairwise independent, the derivative rule is given as:

$$\frac{\partial}{\partial a_{lm}}\left(a_{ik}b_{ki}\right) = \begin{cases} b_{ki} = b_{ml}, & \text{if } l = i \text{ and } m = k. \\ 0, & \text{otherwise} \end{cases}$$

Recall that the matrix derivative is given as:

$$\frac{\partial}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial}{\partial a_{11}} & \cdots & \frac{\partial}{\partial a_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial a_{n1}} & \cdots & \frac{\partial}{\partial a_{nm}} \end{pmatrix}$$

Combine everything, we get:

$$\frac{\partial \mathrm{Tr}(\mathbf{AB})}{\partial \mathbf{A}} = \sum_{i=1}^{n}\sum_{k=1}^{m} \begin{pmatrix} \frac{\partial}{\partial a_{11}} & \cdots & \frac{\partial}{\partial a_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial a_{n1}} & \cdots & \frac{\partial}{\partial a_{nm}} \end{pmatrix} a_{ik}b_{ki} = \begin{pmatrix} b_{11} & b_{21} & \cdots & b_{m1} \\ b_{12} & b_{22} & \cdots & b_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \cdots & b_{mn} \end{pmatrix} = \mathbf{B}^\top$$

Hence proved.

(b) For $\mathbf{A} \in \mathrm{M}_{n\times n}$ and $\mathbf{x} = (x_1,\ x_2, \cdots,\ x_n)^\top \in \mathbb{R}^n$, we have:

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top \begin{pmatrix} \text{---}\ \mathbf{A}_1\mathbf{x}\ \text{---} \\ \text{---}\ \mathbf{A}_2\mathbf{x}\ \text{---} \\ \vdots \\ \text{---}\ \mathbf{A}_n\mathbf{x}\ \text{---} \end{pmatrix} = \sum_{i=1}^n x_i\mathbf{A}_i\mathbf{x} = \sum_{i=1}^n\sum_{j=1}^n x_i a_{ij} x_j \tag{1}$$

For $\mathbf{A}_i$ is the ith row of the matrix $\mathbf{A}$; also:

$$\mathbf{x}\mathbf{x}^\top\mathbf{A} = \mathbf{x} \begin{pmatrix} | & | & & | \\ \mathbf{x}^\top\mathbf{A}'_1 & \mathbf{x}^\top\mathbf{A}'_2 & \cdots & \mathbf{x}^\top\mathbf{A}'_n \\ | & | & & | \end{pmatrix},$$

where $\mathbf{A}'_j$ is the jth column of matrix $\mathbf{A}$. Therefore,

$$\mathrm{Tr}(\mathbf{x}\mathbf{x}^\top\mathbf{A}) = \sum_{j=1}^n \left(\mathbf{x}\mathbf{x}^\top\mathbf{A}\right)_{jj} = \sum_{j=1}^n x_j \sum_{i=1}^n x_i a_{ij} = \sum_{i=1}^n\sum_{j=1}^n x_i a_{ij} x_j \tag{2}$$

We obtain the same result in (1) and (2), meaning that:

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} = \sum_{i=1}^n\sum_{j=1}^n x_i a_{ij} x_j = \mathrm{Tr}(\mathbf{x}\mathbf{x}^\top\mathbf{A})$$

Hence proved.

Or simply with "cyclic" property on $\mathrm{Tr}(\cdot)$: $\mathrm{Tr}(\mathbf{x}\mathbf{x}^\top\mathbf{A}) = \mathrm{Tr}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ $\square$

(c) The likelihood function of N k-dimensional multivariate Gaussian $\mathbf{X}_i \in \mathcal{N}(\mu, \mathbf{\Sigma})$ is:

$$\mathrm{L}(\mu, \mathbf{\Sigma}) = \prod_{i=1}^N \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}^{(i)}-\mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)}$$

After taking log, the log-likelihood function is:

$$\ell(\mu, \Sigma) = \sum_{i=1}^N \left(-\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)\right).$$

If we take the derivative with respective to μ,

$$\begin{aligned} \frac{\partial\ell}{\partial\mu} = \sum_{i=1}^N \frac{1}{2}\left(\Sigma^{-1}(\mathbf{x}^{(i)} - \mu)\right)^\top + \frac{1}{2}(\mathbf{x}^{(i)} - \mu)^\top\Sigma^{-1} &= \sum_{i=1}^N \frac{1}{2}(\mathbf{x}^{(i)} - \mu)^\top\left(\mathbf{\Sigma}^{-1} + (\mathbf{\Sigma}^{-1})^\top\right) \\ &= \sum_{i=1}^N (\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1} \end{aligned}$$

Letting $\frac{\partial\ell}{\partial\mu} = \mathbf{0}$, we must have $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}^{(i)}$.

Now we take the derivative with respect to $\mathbf{\Sigma}$:

$$\frac{\partial\ell}{\partial\mathbf{\Sigma}} = -\frac{N}{2}\underbrace{\frac{\partial\ln|\mathbf{\Sigma}|}{\partial\mathbf{\Sigma}}}_{(i)} - \frac{1}{2}\sum_{i=1}^N \underbrace{\frac{\partial}{\partial\mathbf{\Sigma}}\left((\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}^{(\mathrm{i})} - \mu)\right)}_{(\mathrm{ii})}$$

In the equation, we have two derivatives to deal with:

(i) For this part, we need several lemmas to cover:

Lemma 1: Jacobi's Formula

For $\mathbf{A} \in \mathrm{M}_{n\times n}$ invertible matrix, we have the following identity:

$$\frac{\mathrm{d}}{\mathrm{d}t}|\mathbf{A}(t)| = \mathrm{tr}\left((\mathbf{A}^*(t))\frac{\mathrm{d}\mathbf{A}(t)}{\mathrm{d}t}\right) = |\mathbf{A}(t)| \cdot \mathrm{tr}\left(\mathbf{A}(t)^{-1} \cdot \frac{\mathrm{d}\mathbf{A}(t)}{\mathrm{d}t}\right)$$

In special case:

$$\frac{\partial|\mathbf{A}|}{\partial\mathbf{A}} = \frac{\partial|\mathbf{A}|}{\partial\mathbf{A}_{ij}} = \mathbf{A}^*_{ji} = (\mathbf{A}^*)^\top$$

Where $\mathbf{A}^*$ be the **adjoint matrix** of $\mathbf{A}$.

Lemma 2: Adjoint Matrix Identity

For a **invertible** matrix $\mathbf{A} \in \mathrm{M}_{n\times n}$, and $\mathbf{A}^*$ be $\mathbf{A}$'s adjoint, we have:

$$\mathbf{A}\mathbf{A}^* = |\mathbf{A}| \cdot \mathbf{I} \quad \text{and} \quad \mathbf{A}^* = |\mathbf{A}| \cdot \mathbf{A}^{-1}$$

With the assist of **lemma 1, 2**, we have:

$$\frac{\partial \ln|\mathbf{\Sigma}|}{\partial\mathbf{\Sigma}} = \frac{\partial \ln|\mathbf{\Sigma}|}{\partial|\mathbf{\Sigma}|} \cdot \frac{\partial|\mathbf{\Sigma}|}{\partial\mathbf{\Sigma}} = \frac{1}{|\mathbf{\Sigma}|} \cdot \left(|\mathbf{\Sigma}| \cdot \mathbf{\Sigma}^{-1}\right)^\top = \frac{1}{|\mathbf{\Sigma}|} \cdot |\mathbf{\Sigma}| \cdot \mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{-1}$$

(ii) We use the result in (b), we get:

$$\frac{\partial}{\partial\mathbf{\Sigma}}\left((\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)\right) = \frac{\partial}{\partial\mathbf{\Sigma}}\mathrm{Tr}\left((\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1}\right)$$

Recall that $\mathrm{Tr}(\cdot,\cdot)$ is **reflexive**, *i.e.* $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$:

$$\frac{\partial}{\partial\mathbf{\Sigma}}\mathrm{Tr}\left((\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1}\right) = \frac{\partial}{\partial\mathbf{\Sigma}}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top\right)$$

Lemma 3: Inverse Matrix Differential

For $\mathbf{A} \in \mathrm{M}_{n\times n}$, and $\mathbf{A}^{-1}$ be its inverse:

$$\mathrm{d}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}(\mathrm{d}\mathbf{A})\mathbf{A}^{-1} \iff \mathrm{d}\mathbf{A} = -\mathbf{A}\mathrm{d}(\mathbf{A}^{-1})\mathbf{A}$$

With **lemma 3**, we get the following:

$$\begin{aligned}
\partial\mathrm{Tr}(\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top) &= \mathrm{Tr}(\partial(\mathbf{\Sigma}^{-1})(\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top)) \\
&= \mathrm{Tr}(-\mathbf{\Sigma}^{-1}(\partial\mathbf{\Sigma})\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top) \\
&= \mathrm{Tr}(-(\partial\mathbf{\Sigma})\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top\mathbf{\Sigma}^{-1})
\end{aligned}$$

Therefore, with the result in (a),

$$\begin{aligned}\frac{\partial}{\partial \mathbf{\Sigma}}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)(\mathbf{x}^{(i)}-\mu)^\top\right) &= -\left(\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)(\mathbf{x}^{(i)}-\mu)^\top\mathbf{\Sigma}^{-1}\right)^\top \\ &= -\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)(\mathbf{x}^{(i)}-\mu)^\top\mathbf{\Sigma}^{-1}\end{aligned}$$

Bring everything back to the original identity:

$$\frac{\partial \ell}{\partial \mathbf{\Sigma}} = -\frac{N}{2}\underbrace{\mathbf{\Sigma}^{-1}}_{\text{(i)}} + \frac{1}{2}\sum_{i=1}^{N}\underbrace{\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)(\mathbf{x}^{(i)}-\mu)^\top\mathbf{\Sigma}^{-1}}_{\text{(ii)}}$$

By letting $\frac{\partial \ell}{\partial \mathbf{\Sigma}} = \mathbf{0}$, we multiply the equations on the left and right by $\mathbf{\Sigma}$:

$$\mathbf{0} = -N\mathbf{\Sigma} + \sum_{i=1}^{N}(\mathbf{x}^{(i)}-\mu)(\mathbf{x}^{(i)}-\mu)^\top \implies \widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)}-\hat{\mu})(\mathbf{x}^{(i)}-\hat{\mu})^\top$$

Since $\log\cdot$ is an increasing function, likelihood function and log-likelihood function will share the same **MLE**.

Hence, finally, we get our maximum likelihood estimators are:

$$\begin{cases}\hat{\mu} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathbf{x}^{(i)} \\ \widehat{\mathbf{\Sigma}} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(\mathbf{x}^{(i)}-\hat{\mu})(\mathbf{x}^{(i)}-\hat{\mu})^\top\end{cases}$$

My Question 1: Theoretical Limitations of Logistic Regression on Imbalanced Data

In classification problems with **severe class imbalance**, standard Logistic Regression models often exhibit a bias toward the majority class. Could you elaborate on how the **Maximum Likelihood Estimation (MLE)** objective function, which underlies Logistic Regression, fundamentally contributes to this bias, and what the theoretical justification is for its reduced effectiveness in maximizing a performance metric like **F1-score** or **Recall** for the minority class? (This question is refined by `Gemini`.)

Example:

Suppose I have a data set with data **99%** in the class 0 and merely **1%** is in the class 1 and we aim to predict the class for each data. Logistic regression is used for classifying.

What I Expect To Get: The hypothesis function is very close to my data structure, *i.e.* the F1 score is high and accurate for each class.

What I End Up Getting: A high F1 score, but **nearly all (>99.9%)** the data have been classified into class 0, even if the data themself are in class 1 originally, **FN** is high, but does not affect the overall score.