# Review of Multinomial Logit Model and Discriminant Analysis

Teaching team of STA314

Department of Statistical Sciences
University of Toronto

Monday Novemeber 11th, 2024

# Linear Discriminant Analysis – Review

Suppose we have $K$ classes, $C = \{0, 1, 2, \ldots, K-1\}$. For any $k \in C$, recall:

- we write
$$\pi_k := \mathbb{P}(Y = k)$$
as the **prior probability** that a randomly chosen observation comes from the $k$-th class.
- Define
$$f_k(x) := \mathbb{P}(X = x \mid Y = k)$$
as the **conditional density function** of $X = x \in \mathbb{R}$ from class $k$.
- In discriminant analysis, a **parametric assumption** is made on $f_k(x)$.

# Linear Discriminant Analysis – Review

- According to the Bayes classifier, we should classify a new point $X = x$ according to

$$\arg\max_{k \in C} p_k(x) := \arg\max_{k \in C} \frac{\pi_k f_k(x)}{\sum_{\ell \in C} \pi_\ell f_\ell(x)} = \arg\max_{k \in C} \pi_k f_k(x).$$

- Assume that

$$X \mid Y = k \sim \mathcal{N}(\mu_k, \sigma_k^2), \quad \forall k \in C,$$

namely,

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right).$$

- **Linear Discriminant Analysis (LDA)** further assumes

$$\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_{K-1}^2 = \sigma^2.$$

# Linear Discriminant Analysis – Continued

- As a result, the Bayes rule classifies $X = x$ as

$$
\begin{aligned}
\arg\max_{k \in C} p_k(x) &= \arg\max_{k \in C} \log(p_k(x)) \\
&= \arg\max_{k \in C} \log\left(\frac{\pi_k f_k(x)}{\sum_{\ell \in C} \pi_\ell f_\ell(x)}\right) \\
&= \arg\max_{k \in C} \log(\pi_k f_k(x)) \\
&= \arg\max_{k \in C} \log\left(\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)\right) \\
&= \arg\max_{k \in C} \left(\log(\pi_k) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(x^2 + \mu_k^2 - 2x\mu_k)}{2\sigma^2}\right)
\end{aligned}
$$

with the goal of **maximizing with respect to** $k$

$$
= \arg\max_{k \in C} \left(\frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)\right)
$$

# Linear Discriminant Analysis - MLE

If we know $\mu_0, \ldots, \mu_{K-1}$, $\sigma^2$, and $\pi_0, \ldots, \pi_{K-1}$, then we can construct the Bayes rule. However, we typically dont know these parameters and need to estimate them from the training data!

**Question:**

Given training data $(x_1, y_1), \ldots, (x_n, y_n)$ for all $k \in C$, we have three parameters to estimate: $\pi_k$, $\mu_k$, and $\sigma^2$. How can you find them using **maximum likelihood estimation (MLE)**?

# Linear Discriminant Analysis - MLE

Let's start with the likelihood function. Given pairs of data $(x_i, y_i)$ for $i = 1, \ldots, n$, we have:

$$L(\mu_0, \ldots, \mu_{K-1}, \pi_0, \ldots, \pi_{K-1}, \sigma) = \prod_{i=1}^{n} L(Y_i = y_i, X_i = x_i).$$

This gives us the log-likelihood function:

$$\begin{aligned}
\ell &:= \log L(\mu_0, \ldots, \mu_{K-1}, \pi_0, \ldots, \pi_{K-1}, \sigma) \\
&= \sum_{i=1}^{n} \log(L(Y_i = y_i, X_i = x_i)) \\
&= \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} \log(L(Y_i = k, X_i = x_i)) \\
&= \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} \log(L(X_i = x_i \mid Y_i = k) L(Y_i = k)) \\
&= \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} \left( \log(\pi_k) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu_k)^2}{2\sigma^2} \right).
\end{aligned}$$

# MLE for $\pi_k$

- Notice that you **cannot** take direct derivatives with respect to $\pi_k$ because they are constrained by $\sum_{k=0}^{K} \pi_k = 1$.
- If $K = 1$, the response variable is **binary**. Then, with $\pi_1 = 1 - \pi_0$, the analysis follows the Bernoulli distribution MLE covered in Tutorial 6.
- If $K \geq 2$, we need to use Lagrange multipliers on a Multinomial distribution to find the solution. For more information, you can read here.

# Linear Discriminant Analysis - MLE for $\mu_k$, $\sigma^2$

Now, for $\mu_k$ and $\sigma^2$, we can take the partial derivatives as follows:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k} \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} \left( \log(\pi_k) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i - \mu_k)^2}{2\sigma^2} \right) \\ &= \sum_{1 \le i \le n, y_i = k} \frac{(x_i - \mu_k)}{\sigma^2}.\end{aligned}$$

which gives us the **Maximum Likelihood Estimator** for $\mu_k$:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{1 \le i \le n, y_i = k} x_i.$$

Now for $\sigma^2$, we take the partial derivatives:

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} \left( \log(\pi_k) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu_k)^2}{2\sigma^2} \right)$$
$$= \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} \left( -\frac{1}{2\sigma^2} + \frac{(x_i - \mu_k)^2}{2\sigma^4} \right).$$

Setting this derivative to zero and solving, we obtain the maximum likelihood estimator for $\sigma^2$:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=0}^{K-1} \sum_{1 \le i \le n, y_i = k} (x_i - \hat{\mu}_k)^2.$$

# Review of Multinomial Logit Model (Optional)

Recall the question on Multinomial Logit Model during Midterm 2:

**Problem 4 Midterm 2**

The Multinomial Logit Model (MLM) is a popular model for multi-class classification problems. Imagine a study where individuals are asked to choose their preferred product among a list of $K+1$ items. For each product, we have a measurement of its attributes. Here, we consider only one attribute, such as price. The prices of each product are $x_0, x_1, \ldots, x_K$. The MLM assumes that the customer makes their choice $Y$ according to:

$$\log \frac{\mathbb{P}(Y=k)}{\mathbb{P}(Y=0)} = \beta_0^* + \beta_1^* x_k, \quad k \in \{1, \ldots, K\}$$

Product 0 is chosen as the baseline. We write $Y = k$ if the customer chooses product $k$. The unknown coefficients $\beta_0^*$ and $\beta_1^*$ represent the customer's "taste" for price. Suppose we observe $n$ i.i.d. choices $y_1, \ldots, y_n$ of a chosen customer according to the above model.

# Probability Mass Function for Each Class $k$ (Optional)

We start by calculating the probability mass function for each class $0, \dots, K$:
By definition, for all $1 \le k \le K$,

$$\mathbb{P}(Y = k) = e^{\beta_0^* + \beta_1^* x_k} \mathbb{P}(Y = 0) \tag{1}$$

Since:

$$1 = \sum_{k=1}^{K} \mathbb{P}(Y = k) + \mathbb{P}(Y = 0) = \mathbb{P}(Y = 0) \left( \sum_{k=1}^{K} e^{\beta_0^* + \beta_1^* x_k} + 1 \right),$$

we have:

$$\mathbb{P}(Y = 0) = \frac{1}{\sum_{k=1}^{K} e^{\beta_0^* + \beta_1^* x_k} + 1}.$$

Plugging in equation (1), we obtain:

$$\mathbb{P}(Y = k) = \frac{e^{\beta_0^* + \beta_1^* x_k}}{\sum_{j=1}^{K} e^{\beta_0^* + \beta_1^* x_j} + 1}.$$

# Log-likelihood Function at any $\beta_0, \beta_1$ (Optional)

Let $n_k = \sum_{i=1}^{n} 1\{y_i = k\}$, for all $k \in \{0, 1, \ldots, K\}$.
The likelihood of $y_1$ is:

$$L(\beta_0, \beta_1; y_1) = \prod_{k=0}^{K} \mathbb{P}(y_1 = k)^{1\{y_1 = k\}}$$

so that the log-likelihood of $y_1, \ldots, y_n$ at any $\beta_0, \beta_1$ is:

$$\begin{aligned}
\ell(\beta_0, \beta_1) &= \sum_{i=1}^{n} \sum_{k=0}^{K} 1\{y_i = k\} \log \left[\mathbb{P}(y_i = k)\right] \\
&= \sum_{i=1}^{n} 1\{y_i = 0\} \left( - \log \left( 1 + \sum_{k=1}^{K} \exp(\beta_0 + \beta_1 x_k) \right) \right) \\
&\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} 1\{y_i = k\} \left( \beta_0 + \beta_1 x_k - \log \left( 1 + \sum_{k=1}^{K} \exp(\beta_0 + \beta_1 x_k) \right) \right) \\
&= \sum_{k=1}^{K} n_k (\beta_0 + \beta_1 x_k) - n \log \left( 1 + \sum_{k=1}^{K} \exp(\beta_0 + \beta_1 x_k) \right) \qquad (2)
\end{aligned}$$

# Gradient Descent for $\beta_1$ (Optional)

Suppose we know $\beta_0^* = 0$ and we only maximize the log-likelihood function $\ell(\beta_1) := \ell(\beta_0 = 0, \beta_1)$ in equation (2) over $\beta_1 \in \mathbb{R}$ to compute the MLE of $\beta_1^*$.

**Question:**

Write

$$p_0(\beta_0, \beta_1) = \frac{1}{1 + \sum_{k=1}^{K} e^{\beta_0 + \beta_1 x_k}}$$

and

$$p_k(\beta_0, \beta_1) = \frac{e^{\beta_0 + \beta_1 x_k}}{1 + \sum_{k=1}^{K} e^{\beta_0 + \beta_1 x_k}}, \quad k \in \{1, \ldots, K\}.$$

Starting from a given initialization $\hat{\beta}_1^{(0)}$ with a given step size (learning rate) $\alpha$, **state the gradient descent iterates for computing the MLE of $\beta_1^*$.** (You need to derive the expression of the gradient).

# Gradient Descent – Continued (Optional)

Since

$$\frac{\partial \ell(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{k=1}^{K} n_k x_k - n \frac{\sum_{k=1}^{K} e^{\beta_0 + \beta_1 x_k} x_k}{1 + \sum_{k=1}^{K} e^{\beta_0 + \beta_1 x_k}} = \sum_{k=1}^{K} \left[ n_k - n p_k(\beta_0, \beta_1) \right] x_k,$$

the **Gradient Descent Update** for $\hat{\beta}_1^{(t)}$ follows as

$$\hat{\beta}_1^{(t+1)} = \hat{\beta}_1^{(t)} - \alpha \sum_{k=1}^{K} \left[ n_k - n p_k(0, \hat{\beta}_1^{(t)}) \right] x_k.$$

Specifically,

$$p_k(0, \hat{\beta}_1^{(t)}) = \frac{e^{\hat{\beta}_1^{(t)} x_k}}{1 + \sum_{k=1}^{K} e^{\hat{\beta}_1^{(t)} x_k}}.$$

# Convexity of the log-likelihood function when K = 1 (Optional)

**Convexity:** Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all $x, y \in \mathbb{R}$ and all $\lambda \in [0,1]$. A sufficient condition for $f(x)$ to be convex is $f''(x) \geq 0$ for all $x$.

**Question:**

Suppose $K = 1$. Prove that the negative log-likelihood, $-\ell(\beta_1)$, in the previous subquestion is a **convex** function of $\beta_1$. Reason whether or not the MLE of $\beta_1$ can be computed via the gradient descent you derived above with a suitable step size.

From the previous part,

$$-\frac{\partial^2 \ell(\beta_0, \beta_1)}{\partial \beta_1^2} = n\left\{\frac{\sum_{k=1}^K x_k^2 e^{\beta_0+\beta_1 x_k}}{1+\sum_{k=1}^K e^{\beta_0+\beta_1 x_k}} - \left(\frac{\sum_{k=1}^K x_k e^{\beta_0+\beta_1 x_k}}{1+\sum_{k=1}^K e^{\beta_0+\beta_1 x_k}}\right)^2\right\}.$$

For $K = 1$ and $\beta_0 = 0$, this simplifies to

$$\frac{n}{(1+e^{\beta_1 x_1})^2}\left(x_1^2 e^{\beta_1 x_1}(1+e^{\beta_1 x_1}) - (x_1 e^{\beta_1 x_1})^2\right) = n\frac{x_1^2 e^{\beta_1 x_1}}{(1+e^{\beta_1 x_1})^2} \geq 0.$$

Therefore, we know that

$$\frac{\partial^2 \ell(\beta_0, \beta_1)}{\partial \beta_1^2} \geq 0$$

for all $\beta_0$ and $\beta_1$, hence $\ell(\beta_1)$ is convex.
As a result of the convexity of $\ell(\beta_1)$, and since the minimization is over $\beta_1 \in \mathbb{R}$, which is a convex space, gradient descent with a suitable step size guarantees finding the MLE.

# Convexity of the log-likelihood function when $K \geq 2$ (Optional)

**Bonus Question:**

Can you extend the result of the previous subquestion to $K \geq 2$?

**Hint**: For any two sequences $\{a_1, \ldots, a_n\}$ and $\{b_1, \ldots, b_n\}$, the CauchySchwarz inequality states that

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

# Convexity for $K \geq 2$ – Continued (Optional)

For general $K \geq 2$, we have the claim by noting that

$$\frac{\sum_{k=1}^{K} x_k^2 e^{\beta_0+\beta_1 x_k}}{1+\sum_{k=1}^{K} e^{\beta_0+\beta_1 x_k}} - \left(\frac{\sum_{k=1}^{K} x_k e^{\beta_0+\beta_1 x_k}}{1+\sum_{k=1}^{K} e^{\beta_0+\beta_1 x_k}}\right)^2$$

can be rewritten as

$$\frac{\left(1+\sum_{k=1}^{K} e^{\beta_0+\beta_1 x_k}\right)\sum_{k=1}^{K} x_k^2 e^{\beta_0+\beta_1 x_k} - \left(\sum_{k=1}^{K} x_k e^{\beta_0+\beta_1 x_k}\right)^2}{\left(1+\sum_{k=1}^{K} e^{\beta_0+\beta_1 x_k}\right)^2}. \qquad (3)$$

Applying the CauchySchwarz inequality to the numerator, we get

$$\left(\sum_{k=1}^{K} x_k e^{\beta_0+\beta_1 x_k}\right)^2 \leq \left(\sum_{k=1}^{K} x_k^2 e^{\beta_0+\beta_1 x_k}\right)\left(\sum_{k=1}^{K} e^{\beta_0+\beta_1 x_k}\right).$$

Therefore, the equation (3) is further fimplified to be $\geq 0$.

Thus, the inequality holds due to the CauchySchwarz inequality, confirming the convexity of the expression.

# Quiz Time

Please go to Quercus and start the quiz.

The passcode for the quiz is **sta314qq**.