CS189/289A Spring 2023
Linear Discriminant Analysis

Lecture Note

Chawin Sitawarin *

University of California, Berkeley

Demo: https://colab.research.google.com/drive/1jPgvRdsgHa3hjeUcCA5-AobfQ8qJrk9C

1 Introduction and Logistic Example

Start by deriving the LDA decision boundary in the 2D case. Then, we will discuss the following topics to get a better intuition.

Discriminant function for class k given an input x is $\delta_k(x)$. Think of it as some kind of scores for x, and we will use this score to predict a label of x.

$$\hat{y}(x) = \arg\max_{k \in \mathcal{Y}} \delta_k(x) \tag{1}$$

- $\delta_k(x)$ can take many forms, but we usually model it after $\mathrm{P}(\mathbf{Y} = k \mid \mathbf{X} = x)$.
- When $\delta_k(x)$ is a linear function of x, the decision boundary is also linear. This is also true for any *monotonic* function of $\delta_k(x)$.
- For example, if we take the sigmoid function of some linear function, the decision boundary is still linear.

For instance, consider the following posterior probabilities:

$$f(z) = \frac{1}{1 - \exp(-z)} \tag{2}$$

$$\mathrm{P}(\mathbf{Y} = 1 \mid \mathbf{X} = x) = \frac{1}{1 + \exp(-(\beta_1^\top x + \beta_0))} \tag{3}$$

$$= \frac{\exp(\beta_1^\top x + \beta_0)}{1 + \exp(\beta_1^\top x + \beta_0)} \tag{4}$$

$$\mathrm{P}(\mathbf{Y} = 0 \mid \mathbf{X} = x) = \frac{1}{1 + \exp(\beta_1^\top x + \beta_0)} \tag{5}$$

- Note that $\mathrm{P}(\mathbf{Y} = 0 \mid \mathbf{X} = x) + \mathrm{P}(\mathbf{Y} = 1 \mid \mathbf{X} = x) = 1$ as expected for a binary classification problem.
- The monotonic function that turns $\mathrm{P}(\mathbf{Y} = 1 \mid \mathbf{X} = x)$ to a linear function is an inverse of the sigmoid or logistic function, $\log(p/(1-p))$. Sometimes, this is called a *logit function*, and its output is called *logits*.

*Corresponding email: chawins@berkeley.edu

Assume that the 0-1 loss function is used, we have a simple decision rule:

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \mathrm{P}(\mathbf{Y}=1 \mid \mathbf{X}=x) \geq \mathrm{P}(\mathbf{Y}=0 \mid \mathbf{X}=x) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$= \begin{cases} 1 & \text{if } f(x) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$f(x) \coloneqq \log\left(\frac{\mathrm{P}(\mathbf{Y}=1 \mid \mathbf{X}=x)}{\mathrm{P}(\mathbf{Y}=0 \mid \mathbf{X}=x)}\right) = \beta_1^\top x + \beta_0 \tag{8}$$

We can now recover the linear decision boundary.

2 LDA Assumptions

1. Class-conditioned density function of $\mathbf{X}$ is Gaussian.

$$\mathbf{X} \mid \mathbf{Y}=k \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{9}$$

$$\mathrm{P}(\mathbf{X}=x \mid \mathbf{Y}=k) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_k\right|}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right) \tag{10}$$

2. The covariance matrices are the same for all classes.

$$\Sigma_k = \Sigma \quad \forall k \tag{11}$$

We want to get the posterior so we are going to use Bayes rule here:

$$\begin{aligned} \mathrm{P}(\mathbf{Y}=k \mid \mathbf{X}=x) &= \frac{\mathrm{P}(\mathbf{X}=x \mid \mathbf{Y}=k)\,\mathrm{P}(\mathbf{Y}=k)}{\mathrm{P}(\mathbf{X}=x)} \\ &= \frac{\mathrm{P}(\mathbf{X}=x \mid \mathbf{Y}=k)\pi_k}{\mathrm{P}(\mathbf{X}=x)} \\ &= \frac{\mathrm{P}(\mathbf{X}=x \mid \mathbf{Y}=k)\pi_k}{\sum_{k'} \mathrm{P}(\mathbf{X}=x \mid \mathbf{Y}=k')\,\mathrm{P}(\mathbf{Y}=k')} \end{aligned}$$

Notice that the denominator is the same for all k, so we can "ignore" it. It will just be a constant number added to all discriminant functions equally, $C = \log \mathrm{P}(\mathbf{X}=x)$.

$$\begin{aligned} \delta_k(x) &\coloneqq \log \mathrm{P}(\mathbf{Y}=k \mid \mathbf{X}=x) \\ &= \log \mathrm{P}(\mathbf{X}=x \mid \mathbf{Y}=k)\pi_k + C \\ &= \log \pi_k - \frac{1}{2}\log\left|\Sigma_k\right| - \frac{1}{2}\left(\mu_k\right)^T \Sigma_k^{-1}\mu_k + \left(\mu_k\right)^T \Sigma_k^{-1} x - \frac{1}{2}x^T\Sigma_k^{-1}x + C' \end{aligned} \tag{12}$$

This discriminant is actually not linear in x! But we will use the second assumption to show that the decision boundary is linear.

Considering the binary-class problem, the decision boundary is given by $\{z \mid \delta_0(z) = \delta_1(z)\}$.

$$\delta_0(x) = \delta_1(x) \tag{13}$$

$$0 = \delta_1(x) - \delta_0(x) \tag{14}$$

$$= \log\left(\frac{\mathrm{P}(\mathbf{Y}=1 \mid \mathbf{X}=x)}{\mathrm{P}(\mathbf{Y}=0 \mid \mathbf{X}=x)}\right) \tag{15}$$

$$= \log\frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0) + \left(\mu_1-\mu_0\right)^T \Sigma^{-1}x \tag{16}$$

Next we will do some more analysis and intuition checks.

3 What happens when we don't account for priors?

1. **Demo**: Show that the LDA decision boundaries are not invariant to the priors.
2. A common mistake is forgetting the prior which will yield a wrong decision boundary when we have *imbalance* classes, i.e., $\pi_0 \neq \pi_1$.
3. *How much worse does the error become in this case in terms of π_0 and π_1 (as well as μ_0, μ_1, Σ), assuming that LDA assumptions are both true?*
4. **Demo**: Compare to linear SVM for Gaussian and uniform data. SVM is not affected by the priors when the data is **uniformly distributed and are linearly separable**.

4 What happens when variance is high?

1. **Demo**: Show high variance in data leads to also high variance in the LDA decision boundaries.

5 What happens when the covariance matrices are different between classes?

6 What happens when the covariance matrices are anisotropic?

LDA makes no assumption that the covariance matrix have to be isotropic, i.e., $\Sigma = \sigma^2 I_n$ for some $\sigma \in \mathbb{R}$.

1. **Demo**: Show that the LDA decision boundary is not *necessarily* perpendicular to the line connecting the two class centroids (means).
2. This is not true the other way around, i.e., given that the boudnary decision is perpendicular to the line connecting the centroids, it is not necessary that the covariance matrix is isotropic. *Can you give an example?*
3. You will be proving this in Problem 4 of the worksheet.

7 Next Steps

1. Work on problem 1 and 2 in the worksheet.

References

This note is, for the most parts, based on Hastie et al. [2001] and some on the course lecture note by Prof. Shewchuk (https://people.eecs.berkeley.edu/~jrs/189/lec/07.pdf).

References

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. 3