Classification

Daniel Mao

Contents

1 The First Chapter **1**

1.1 Classification 1

1.2 Linear Discriminant Analysis (LDA) 2

1.3 Fisher's Linear Discriminant Analysis (FDA) 3

Chapter 1

The First Chapter

1.1 Classification

Definition (Classification)**.** *We define* ***classification*** *to be the problem of predicting a* <u>*discrete*</u> *random variable* $\boldsymbol{y}$ *from another random variable* $\boldsymbol{x}$*.*

Definition (Error Rate)**.** *Let* h *be a classifier. We define the* ***error rate*** *of* h*, denoted by* $L(h)$*, to be a probability given by*

$$L(h) := Pr(h(\boldsymbol{x}) \neq \boldsymbol{y}).$$

Definition (Empirical Error Rate)**.** *Let* h *be a classifier. We define the* ***empirical error rate*** *of* h*, denoted by* $\hat{L}_n(h)$*, to be an average given by*

$$\hat{L}_n(h) := \frac{1}{n} \sum_{i=1}^{n} I(h(\boldsymbol{x}_i) \neq \boldsymbol{y}_i).$$

Definition (Bayes Classifier)**.** *Let* $\mathcal{X} = \mathbb{R}^d$ *for some* $d \in \mathbb{N}$*. Let* $\mathcal{Y} = \{1..K\}$*. We define the* ***Bayes classifier****, denoted by* h^**, to be a function from* $\mathcal{X}$ *to* $\mathcal{Y}$ *given by*

$$h^*(\boldsymbol{x}) := \mathrm{argmax}_{k \in \mathcal{Y}} \{\mathbb{P}(Y = k \mid \boldsymbol{X} = \boldsymbol{x})\}.$$

Definition (Decision Boudnary)**.** *Let* h *be a classifier. We define the* ***decision boundary*** *of* h*, denoted by* $D(h)$*, to be a set given by*

$$D(h) := \{\boldsymbol{x} : Pr(\boldsymbol{y} = 1 \mid \boldsymbol{x} = \boldsymbol{x}_0) = Pr(\boldsymbol{y} = 0 \mid \boldsymbol{x} = \boldsymbol{x}_0)\}.$$

Theorem 1. *The Bayes rule is optimal. i.e., if* h^* *is the Bayes rule and* h *is any other classification rule, then* $L(h^*) \leq L(h)$*.*

1.2 Linear Discriminant Analysis (LDA)

Theorem 2. *Define for each $k \in \mathcal{Y}$ the class conditional f_k by $f_k(x) := \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid Y = k)$. Define for each $k \in \mathcal{Y}$ the prior π_k by $\pi_k := \mathbb{P}(\boldsymbol{Y} = k)$. Assume that the random variables $\boldsymbol{X} \mid Y = k$ follows a Gaussian distribution. i.e.,*

$$\forall k \in \mathcal{Y}, \quad f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right).$$

Then the Bayes classifier h^* is

$$h^*(\boldsymbol{x}) = \operatorname{argmax}_{k \in \mathcal{Y}} \delta_k(\boldsymbol{x})$$

where

$$\delta_k(\boldsymbol{x}) := -\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k).$$

Proof.

$$\begin{aligned}
& \mathbb{P}(Y = i \mid \boldsymbol{X} = \boldsymbol{x}) \sim \mathbb{P}(Y = j \mid \boldsymbol{X} = \boldsymbol{x}) \\
\iff \quad & \frac{f_i(\boldsymbol{x})\pi_i}{\sum_{k=1}^K f_k(\boldsymbol{x})\pi_k} \sim \frac{f_j(\boldsymbol{x})\pi_j}{\sum_{k=1}^K f_k(\boldsymbol{x})\pi_k} \\
\iff \quad & f_i(\boldsymbol{x})\pi_i \sim f_j(\boldsymbol{x})\pi_j \\
\iff \quad & \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right)\pi_i \sim \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)\right)\pi_j \\
\iff \quad & \frac{1}{|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right)\pi_i \sim \frac{1}{|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)\right)\pi_j \\
\iff \quad & -\frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \log(\pi_i) \sim -\frac{1}{2}\log(|\Sigma_j|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j) + \log(\pi_j).
\end{aligned}$$

So

$$\begin{aligned}
h^*(\boldsymbol{x}) &= \operatorname{argmax}_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \operatorname{argmax}_{k \in \mathcal{Y}} \left(-\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k)\right) \\
&= \operatorname{argmax}_{k \in \mathcal{Y}} \delta_k(\boldsymbol{x}).
\end{aligned}$$

■

Remark (Euclidian Distance)**.** *Suppose $\forall k \in \mathcal{Y}$, $\Sigma_k = I$. Then*

$$\begin{aligned}
\delta_k(\boldsymbol{x}) &= -\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k) \\
&= -\frac{1}{2}\log(|I|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top I(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k) \\
&= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k).
\end{aligned}$$

So a new point can be classified by its distance from the center of a class, adjusted by some prior. In a two-class problem with equal prior, the discriminating function would be the perpendicular bisector of the two classes' means.

Remark (Mahalanobis Distance)**.** *Say the singular value decomposition of* Σ_k *is* $\Sigma_k = U_k S_k V_k^\top$*. Since covariance matrices are symmetric,* $U = V$*. So* $\Sigma_k = U_k S_k U_k^\top$*. Now*

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) &= (\boldsymbol{x}-\boldsymbol{\mu}_k)^\top U_k S_k^{-1} U_K^\top (\boldsymbol{x}-\boldsymbol{\mu}_k) \\ &= (U_k^\top \boldsymbol{x} - U_k^\top \boldsymbol{\mu}_k)^\top S_k^{-1}(U_k^\top \boldsymbol{x} - U_k^\top \boldsymbol{\mu}_k) \\ &= (U_k^\top \boldsymbol{x} - U_k^\top \boldsymbol{\mu}_k)^\top S_k^{-1/2} S_k^{-1/2}(U_k^\top \boldsymbol{x} - U_k^\top \boldsymbol{\mu}_k) \\ &= (S_k^{-1/2}U_k^\top \boldsymbol{x} - S_k^{-1/2}U_k^\top \boldsymbol{\mu}_k)^\top I (S_k^{-1/2}U_k^\top \boldsymbol{x} - S_k^{-1/2}U_k^\top \boldsymbol{\mu}_k) \\ &= (A\boldsymbol{x} - A\boldsymbol{\mu}_k)^\top I (A\boldsymbol{x} - A\boldsymbol{\mu}_k), \text{ where } A := S_k^{-1/2}U_k^\top.\end{aligned}$$

That is,

$$(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) = (A\boldsymbol{x} - A\boldsymbol{\mu}_k)^\top I (A\boldsymbol{x} - A\boldsymbol{\mu}_k)$$

where $A = S_k^{-1/2}U_k^\top$*.*

1.3 Fisher's Linear Discriminant Analysis (FDA)

FDA is a method of dimensionality reduction. Projects the data on the direction of maximum separation.
Goal:

- Minimize the variance of each class.
- Maximize the distance between projected means.

$$\max(\boldsymbol{w}^\top \boldsymbol{\mu}_0 - \boldsymbol{w}^\top \boldsymbol{\mu}_1)^2.$$

$$\begin{aligned}(\boldsymbol{w}^\top \boldsymbol{\mu}_0 - \boldsymbol{w}^\top \boldsymbol{\mu}_1)^2 &= (\boldsymbol{w}^\top \boldsymbol{\mu}_0 - \boldsymbol{w}^\top \boldsymbol{\mu}_1)^\top (\boldsymbol{w}^\top \boldsymbol{\mu}_0 - \boldsymbol{w}^\top \boldsymbol{\mu}_1) \\ &= (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top)\boldsymbol{w}\boldsymbol{w}^\top(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\ &= \boldsymbol{w}^\top(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}\end{aligned}$$

The variance of projected points in each class are

$$\boldsymbol{w}^\top \Sigma_0 \boldsymbol{w}, \text{ and } \boldsymbol{w}^\top \Sigma_1 \boldsymbol{w}.$$

To minimize them,

$$\min \quad \boldsymbol{w}^\top \Sigma_0 \boldsymbol{w} + \boldsymbol{w}^\top \Sigma_1 \boldsymbol{w}.$$

$$\iff \min \quad \boldsymbol{w}^\top (\Sigma_0 + \Sigma_1)\boldsymbol{w}.$$

So we can do

$$\max \quad w^T S_B w$$

subject to $w^T S_w w = 1$.

$$L(w, \lambda) = w^T S_B w - \lambda(w^T S_w w - 1).$$

$$\frac{\partial}{\partial w} L = 2 S_B w - 2\lambda S_w w$$

Setting this to zero we get

$$S_B w = \lambda S_w w.$$

$$\iff S_w^{-1} S_B w = \lambda w.$$

Note that $\text{rank}(S_B) = 1$. So $\text{rank}(S_w^{-1} S_B) = 1$. So $S_w^{-1} S_B$ has only one eigenvector.