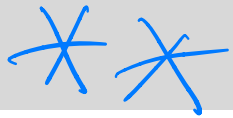
Linear Discriminant Analysis and Quadratic Discriminant Analysis



Introduction

- Assume we have a dataset of instances $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ with sample size n and dimensionality $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
- The y_i 's are the class labels.
- Linear Discriminant Analysis (LDA) and Quadratic discriminant Analysis (QDA) are two well-known supervised classification methods in statistical and probabilistic learning.

Optimization for the Boundary of Classes

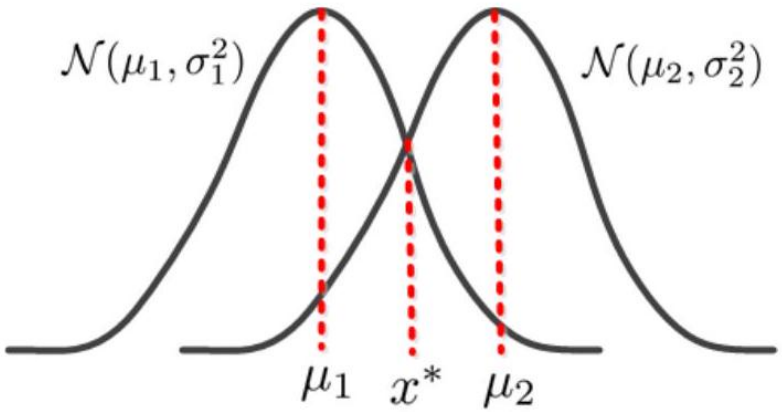
- First suppose the data is one dimensional, $x \in \mathbb{R}$.
- We assume that the two classes have normal (Gaussian) distribution which is the most common and default distribution in the real-world applications.
- An instance $x \in \mathbb{R}$ belongs to one of these two classes:

$$x \sim \begin{cases} \mathcal{N}\left(\mu_1, \sigma_1^2\right), & \text{if } x \in \mathcal{C}_1 \\ \mathcal{N}\left(\mu_2, \sigma_2^2\right), & \text{if } x \in \mathcal{C}_2 \end{cases}$$

  where $\mathcal{C}_1$ and $\mathcal{C}_2$ denote the first and second class, respectively.
- For an instance x, we may have an error in estimation of the class it belongs to. At a point, which we denote by x^*, the probability of the two classes are equal; therefore, the point x^* is on the boundary of the two classes.

Optimization for the Boundary of Classes

- Suppose $\mu_1 < \mu_2$.



- We can say $\mu_1 < x^* < \mu_2$ as shown in Fig. 1
- Therefore, if $x < x^*$ or $x > x^*$ the instance x belongs to the first and second class, respectively.

Optimization for the Boundary of Classes

- Estimating $x < x^*$ or $x > x^*$ to respectively belong to the second and first class is an error in estimation of the class. This probability of the error can be stated as:

$$\mathbb{P}(\text{ error }) = \mathbb{P}\left(x > x^*, x \in \mathcal{C}_1\right) + \mathbb{P}\left(x < x^*, x \in \mathcal{C}_2\right).$$

- As we have $\mathbb{P}(A, B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$, we can say:

$$\begin{aligned}\mathbb{P}(\text{ error }) =& \mathbb{P}\left(x > x^* \mid x \in \mathcal{C}_1\right)\mathbb{P}\left(x \in \mathcal{C}_1\right) \\ &+ \mathbb{P}\left(x < x^* \mid x \in \mathcal{C}_2\right)\mathbb{P}\left(x \in \mathcal{C}_2\right),\end{aligned}$$

  which is what we want to minimize.

Optimization for the Boundary of Classes

- According to the definition of CDF, we have:

$$\begin{aligned}
&\mathbb{P}\left(x < c, x \in \mathcal{C}_1\right) = F_1(c), \\
&\Longrightarrow \mathbb{P}\left(x > x^*, x \in \mathcal{C}_1\right) = 1 - F_1\left(x^*\right), \\
&\mathbb{P}\left(x < x^*, x \in \mathcal{C}_2\right) = F_2\left(x^*\right).
\end{aligned}$$

- According to the definition of PDF, we have:

$$\begin{aligned}
\mathbb{P}\left(x \in \mathcal{C}_1\right) &= f_1(x) = \pi_1, \\
\mathbb{P}\left(x \in \mathcal{C}_2\right) &= f_2(x) = \pi_2,
\end{aligned}$$

  where we denote the priors $f_1(x)$ and $f_2(x)$ by π_1 and π_2, respectively.

- Hence, the problem becomes:

$$\underset{x^*}{\text{minimize}} \left(1 - F_1\left(x^*\right)\right)\pi_1 + F_2\left(x^*\right)\pi_2.$$

Optimization for the Boundary of Classes

- We take derivative for the sake of minimization:

$$\frac{\partial \mathbb{P}(\text{ error })}{\partial x^*} = -f_1(x^*)\,\pi_1 + f_2(x^*)\,\pi_2 \overset{\text{set}}{=} 0,$$
$$\Longrightarrow f_1(x^*)\,\pi_1 = f_2(x^*)\,\pi_2.$$

Optimization for the Boundary of Classes

- Another way to obtain this expression is equating the posterior probabilities to have the equation of the boundary of classes:

$$\mathbb{P}\left(x \in \mathcal{C}_1 \mid X = x\right) \overset{\text{set}}{=} \mathbb{P}\left(x \in \mathcal{C}_2 \mid X = x\right).$$

- According to Bayes rule, the posterior is:

$$\begin{aligned}\mathbb{P}\left(x \in \mathcal{C}_1 \mid X = x\right) &= \frac{\mathbb{P}\left(X = x \mid x \in \mathcal{C}_1\right)\mathbb{P}\left(x \in \mathcal{C}_1\right)}{\mathbb{P}(X = x)} \\ &= \frac{f_1(x)\pi_1}{\sum_{k=1}^{|\mathcal{C}|}\mathbb{P}\left(X = x \mid x \in \mathcal{C}_k\right)\pi_k},\end{aligned}$$

- We can see that the boundary is also $f_1\left(x^*\right)\pi_1 = f_2\left(x^*\right)\pi_2$.

d-dimensional case

- Now let us think of data as multivariate data with dimensionality d. The PDF for multivariate Gaussian distribution, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is:

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

  where $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{\mu} \in \mathbb{R}^d$ is the mean, $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is the covariance matrix, and $|\cdot|$ is the determinant of matrix.

- The decision boundary becomes:

$$\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_1|}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\right) \pi_1$$

$$= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_2|}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\right) \pi_2,$$

LDA for Binary Classification

- In Linear Discriminant Analysis (LDA), we assume that the two classes have equal covariance matrices:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$$

- The decision boundary becomes:

$$\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\right) \pi_1$$

$$= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\right) \pi_2$$

$$\implies \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\right) \pi_1 = \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\right) \pi_2$$

$$\overset{(a)}{\implies} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1) + \ln(\pi_1)$$

$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln(\pi_2)$$

LDA for Binary Classification

- Thus, we have:

$$
\begin{aligned}
&-\frac{1}{2}\boldsymbol{x}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_1^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_1^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\ln\left(\pi_1\right)\\
&=-\frac{1}{2}\boldsymbol{x}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_2^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\boldsymbol{\mu}_2^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\ln\left(\pi_2\right).
\end{aligned}
$$

- Therefore, if we multiply the sides of equation by 2 , we have:

$$
\begin{aligned}
&2\left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\right)\right)^{\top}\boldsymbol{x}\\
&\quad+\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\Big)+2\ln\left(\frac{\pi_2}{\pi_1}\right)=0
\end{aligned}
$$

- It is the equation of a line in the form of $\boldsymbol{a}^{\top}\boldsymbol{x}+b=0$.

LDA for Binary Classification

- Let

$$\delta(\boldsymbol{x}) := 2\left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)\right)^{\top}\boldsymbol{x} + \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\Big) + 2\ln\left(\frac{\pi_2}{\pi_1}\right),$$

- The class of an instance $\boldsymbol{x}$ is estimated as:

$$\widehat{\mathcal{C}}(x) = \begin{cases} 1, & \text{if } \delta(\boldsymbol{x}) < 0 \\ 2, & \text{if } \delta(\boldsymbol{x}) > 0 \end{cases}$$

QDA for Binary Classification

- we relax the assumption of equality of the covariance matrices:

$$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$$

- By similar calculation as before, we can show that the decision boundary is:

$$\begin{aligned}
&\boldsymbol{x}^\top \left(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\right)^{-1} \boldsymbol{x} + 2\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right)^\top \boldsymbol{x} \\
&+ \left(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) + \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) \\
&+ 2\ln\left(\frac{\pi_2}{\pi_1}\right) = 0,
\end{aligned}$$

- It is in the quadratic form $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c = 0$.

LDA and QDA for Multi-class Classification

- Now we consider multiple classes, which can be more than two, indexed by $k \in \{1, \ldots, |\mathcal{C}|\}$.
- the scaled posterior of the k-th class becomes:

$$\begin{aligned}\delta_k(\boldsymbol{x}) := &-\frac{1}{2}\ln\left(|\boldsymbol{\Sigma}_k|\right) \\ &-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)+\ln\left(\pi_k\right).\end{aligned}$$

In QDA, the class of the instance $\boldsymbol{x}$ is estimated as:

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x})$$

LDA and QDA for Multi-class Classification

- In LDA, we assume that the covariance matrices of the k classes are equal:

$$\Sigma_1 = \cdots = \Sigma_{|\mathcal{C}|} = \Sigma$$

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= -\frac{1}{2}\ln(|\boldsymbol{\Sigma}|) \\ &- \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right) + \ln\left(\pi_k\right) \\ &= -\frac{1}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \ln\left(\pi_k\right).\end{aligned}$$

Estimation for LDA and QDA

- Usually, the prior of the k-th class is estimated according to the sample size of the k-th class:

$$\widehat{\pi}_k = \frac{n_k}{n}$$

where n_k and n are the number of training instances in the k-th class and in total, respectively.

- The mean of the k-th class can be estimated using the Maximum Likelihood Estimation (MLE), or Method of Moments (MOM), for the mean of a Gaussian distribution:

$$\mathbb{R}^d \ni \widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i=1}^{n} \boldsymbol{x}_i \mathbb{I}\left(\mathcal{C}\left(\boldsymbol{x}_i\right) = k\right).$$

Estimation for LDA and QDA

- In QDA, the covariance matrix of the k-th class is estimated using MLE:

$$\mathbb{R}^{d\times d} \ni \widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k\right)\left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k\right)^\top \mathbb{I}\left(\mathcal{C}\left(\boldsymbol{x}_i\right) = k\right).$$

Or we can use the unbiased estimation of the covariance matrix:

$$\mathbb{R}^{d\times d} \ni \widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k\right)\left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k\right)^\top \mathbb{I}\left(\mathcal{C}\left(\boldsymbol{x}_i\right) = k\right).$$

Estimation for LDA and QDA

- In LDA, we assume that the covariance matrices of the classes are equal; therefore, we use the weighted average of the estimated covariance matrices as the common covariance matrix in LDA:

$$\mathbb{R}^{d\times d} \ni \widehat{\boldsymbol{\Sigma}} = \frac{\sum_{k=1}^{|\mathcal{C}|} n_k \widehat{\boldsymbol{\Sigma}}_k}{\sum_{r=1}^{|\mathcal{C}|} n_r} = \frac{\sum_{k=1}^{|\mathcal{C}|} n_k \widehat{\boldsymbol{\Sigma}}_k}{n},$$

where the weights are the cardinality of the classes.

LDA and QDA are Metric Learning!

- Assume that the covariance matrices are all equal (as we have in LDA) and they all are the identity matrix:

$$\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_{|\mathcal{C}|} = \boldsymbol{I}$$

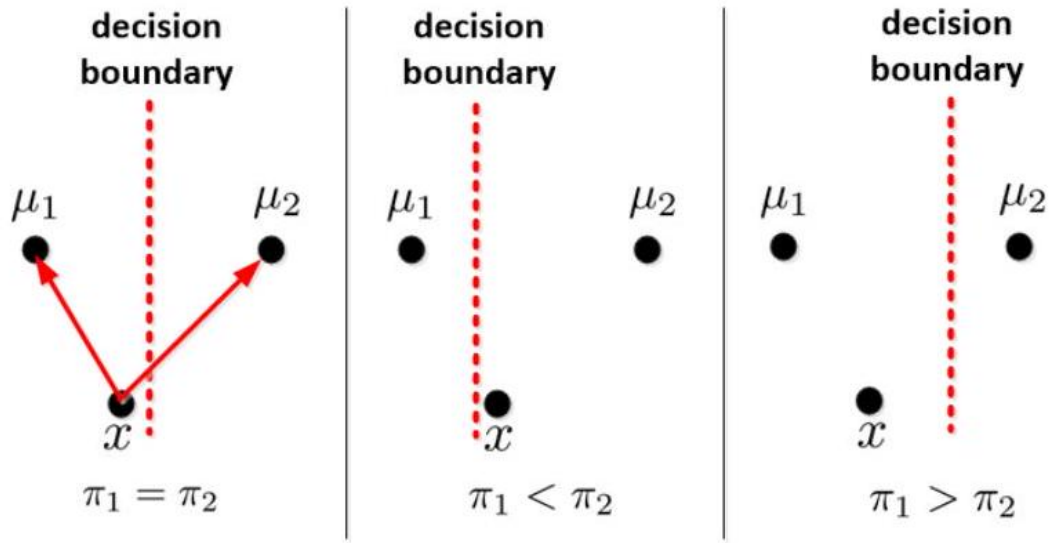
  which means that all the classes are assumed to be spherically distributed in the d dimensional space.

- $$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)^\top \left(\boldsymbol{x} - \boldsymbol{\mu}_k\right) + \ln\left(\pi_k\right),$$

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)^\top \left(\boldsymbol{x} - \boldsymbol{\mu}_k\right) = -\frac{1}{2}d_k^2,$$

$$d_k = \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2 = \sqrt{\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)^\top \left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)}.$$

Thus, the QDA or LDA reduce to simple Euclidean distance from the means of classes if the covariance matrices are all identity matrix and the priors are equal. Simple



- For general case when $\boldsymbol{\Sigma} \neq \boldsymbol{I}$, we use Mahalanobis distance:

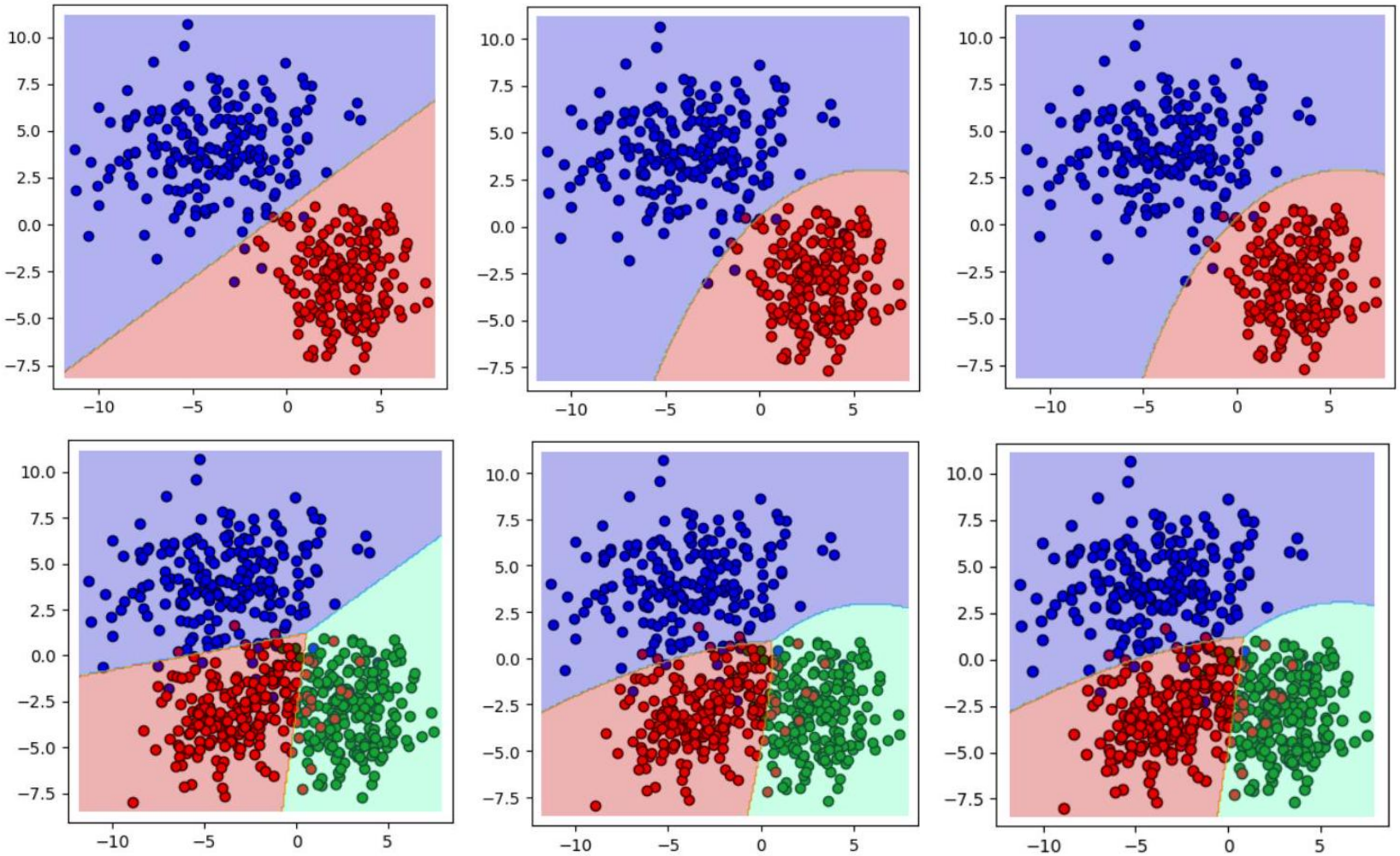
$$d_M^2(\boldsymbol{x}, \boldsymbol{\mu}) := \|\boldsymbol{x} - \boldsymbol{\mu}\|_M^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}),$$

.

Relationship with Naive Bayes classifier

- Gaussian naive Bayes is equivalent to QDA where the covariance matrices are di agonal, i.e., the off-diagonal of the covariance matrices are ignored.
- Therefore, we can say that QDA is more powerful than Gaussian naive Bayes because Gaussian naive Bayes is a simplified version of QDA.
- Gaussian naive Bayes and QDA are equivalent for one dimensional data.

Experiments



Experiments

Experiments with equal class sample sizes: (1) LDA for two classes, (2) QDA for two classes, (3) Gaussian naive Bayes for two classes, (4) LDA for three classes, (5) QDA for three classes, (6) Gaussian naive Bayes for three classes.

High-dimensional consideration

- For K classes case, the decision rule for LDA can be equivalently written as

$$\hat{Y} = \arg\max_k \left\{ \left(\mathbf{X} - \frac{\boldsymbol{\mu}_k}{2} \right)^{\mathrm{T}} \boldsymbol{\beta}_k + \log \pi_k \right\},$$

  where $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$ for $k = 1, \ldots, K$.
- To estimate $\boldsymbol{\beta}_k$, traditional approach estimates $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}^{-1}$ separately.
- In high-dimensions, like LASSO, we can assume both $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}^{-1}$ are sparse.
- This assumption can somehow decrease the estimation errors for the parameters and make $\boldsymbol{\Sigma}^{-1}$ estimable.

High-dimensional consideration

- A sparse estimation for $\boldsymbol{\mu}_k$ is easy to construct; sparse estimation for $\boldsymbol{\Sigma}^{-1}$ is more difficult and computationally expensive.
- What if we direct assume that $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$ is sparse?
- Mai Q, Yang Y, Zou H. Multiclass sparse discriminant analysis[J]. Statistica Sinica, 2019, 29(1): 97-111.

High-dimensional consideration

- Let $\boldsymbol{\theta}_k^{\text{Bayes}} = \boldsymbol{\beta}_k - \boldsymbol{\beta}_1$ for $k = 1, \ldots, K$. Then the Bayes decision rule can be written as

$$\hat{Y} = \arg\max_k \left\{ \left(\boldsymbol{\theta}_k^{\text{Bayes}} \right)^{\mathrm{T}} \left(\mathbf{X} - \frac{\boldsymbol{\mu}_k}{2} \right) + \log \pi_k \right\}.$$

- We refer to the directions $\boldsymbol{\theta}^{\text{Bayes}} = \left(\boldsymbol{\theta}_2^{\text{Bayes}}, \ldots, \boldsymbol{\theta}_K^{\text{Bayes}} \right) \in \mathbb{R}^{p \times (K-1)}$ as the discriminant directions.
- On the population level, we have

$$\left(\boldsymbol{\theta}_2^{\text{Bayes}}, \ldots, \boldsymbol{\theta}_K^{\text{Bayes}} \right)$$

$$= \arg \min_{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\theta}_k - \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1 \right)^{\mathrm{T}} \boldsymbol{\theta}_k \right\}.$$

High-dimensional consideration

- In the classical low-dimension-large-sample-size setting, we can estimate $\left(\boldsymbol{\theta}_2^{\text{Bayes}}, \ldots, \boldsymbol{\theta}_K^{\text{Bayes}}\right)$ via an empirical version

$$\left(\hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_K\right)$$
$$= \arg \min_{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta}_k - \left(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1\right)^{\mathrm{T}} \boldsymbol{\theta}_k \right\}.$$

- Still cannot handle high-dimensional data sets.

High-dimensional consideration

- Write $\boldsymbol{\theta}_{\cdot j} = (\theta_{2j}, \ldots, \theta_{Kj})^{\mathrm{T}}$ and define $\|\boldsymbol{\theta}_{\cdot j}\| = \left(\sum_{i=2}^{K} \theta_{ij}^2\right)^{1/2}$.
- For the high-dimensional case, consider the following penalized formulation for multiclass sparse discriminant analysis.

$$\begin{aligned}
&\left(\hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_K\right) \\
&= \arg \min_{\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K} \sum_{k=2}^{K} \left\{ \frac{1}{2} \boldsymbol{\theta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta}_k - \left(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1\right)^{\mathrm{T}} \boldsymbol{\theta}_k \right\} \\
&\quad + \lambda \sum_{j=1}^{p} \|\boldsymbol{\theta}_{\cdot j}\|,
\end{aligned}$$

Algorithm

- Given $\left\{\boldsymbol{\theta}_{\cdot j'}, j' \neq j\right\}$, the solution of $\boldsymbol{\theta}_{\cdot j}$ to (10) is defined as

$$\arg\min_{\boldsymbol{\theta}_{\cdot j}} \sum_{k=2}^{K} \frac{1}{2}\left(\theta_{kj} - \tilde{\theta}_{kj}\right)^2 + \frac{\lambda}{\hat{\sigma}_{jj}} \left\|\boldsymbol{\theta}_{\cdot j}\right\|$$

  where $\tilde{\boldsymbol{\theta}}_{k,j} = \frac{\hat{\delta}_j^k - \sum_{l \neq j} \hat{\sigma}_{lj}\theta_{kl}}{\hat{\sigma}_{jj}}$.

- Let $\tilde{\boldsymbol{\theta}}_{\cdot j} = \left(\tilde{\theta}_{2j}, \ldots, \tilde{\theta}_{Kj}\right)^{\mathrm{T}}$ and $\left\|\tilde{\boldsymbol{\theta}}_{\cdot j}\right\| = \left(\sum_{k=2}^{K} \tilde{\theta}_{kj}^2\right)^{1/2}$. The solution is given by

$$\hat{\boldsymbol{\theta}}_{\cdot j} = \tilde{\boldsymbol{\theta}}_{\cdot j}\left(1 - \frac{\lambda}{\left\|\tilde{\boldsymbol{\theta}}_{\cdot j}\right\|}\right)_+$$

Algorithm

- write $\hat{\boldsymbol{\delta}}^k = \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1$.
  1. Compute $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\delta}}^k, k = 1, 2, \ldots, K$;
  2. Initialize $\hat{\boldsymbol{\theta}}_k^{(0)}$ and compute $\tilde{\boldsymbol{\theta}}_k^{(0)}$ accordingly;
  3. For $m = 1, \ldots,$ do the following loop until convergence: for $j = 1, \ldots, p$, (a) compute

$$\hat{\boldsymbol{\theta}}_{\cdot j}^{(m)} = \tilde{\boldsymbol{\theta}}_{\cdot j}^{(m-1)} \left(1 - \frac{\lambda}{\left\| \tilde{\boldsymbol{\theta}}_{\cdot j}^{(m-1)} \right\|} \right)_+$$

     (b) update

$$\tilde{\theta}_{kj} = \frac{\hat{\delta}_j^k - \sum_{l \neq j} \hat{\sigma}_{lj} \hat{\theta}_{kl}^{(m)}}{\hat{\sigma}_{jj}}.$$

  4. Let $\hat{\boldsymbol{\theta}}_k$ be the solution at convergence. The output classifier is the usual linear discriminant classifier on $\left(\mathbf{X}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_2, \ldots, \mathbf{X}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_K \right)$.

Theorem

Under some technical conditions , there exists a generic constant M such that, if $\lambda < \min\left\{\frac{\theta_{\min}}{8\varphi}, M(1-\kappa)\right\}$, then with a probability greater than

$$1-Cpd\exp\left(-Cn\frac{\epsilon^2}{Kd^2}\right)-CK\exp\left(-C\frac{n}{K^2}\right)-Cp(K-1)\exp\left(-Cn\frac{\epsilon^2}{K}\right)$$

we have that $\hat{\mathcal{D}} = \mathcal{D}$, and $\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^{\text{Bayes}}\right\|_\infty \leq 4\varphi\lambda$ for $k = 2, \ldots, K$. 2. If we further assume conditions (C2)-(C3), we have that if $\left\{\frac{d^2\log(pd)}{n}\right\}^{1/2} \ll \lambda \ll \theta_{\min}$, then with probability tending to 1 , we have $\hat{\mathcal{D}} = \mathcal{D}$, and $\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^{\text{Bayes}}\right\|_\infty \leq 4\varphi\lambda$ for $k = 2, \ldots, K$.

- Waht about high-dimensional QDA?
- It is more technically involved since we need to give a sparse estimation for $\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1}$.
- Cai T T, Zhang L. A convex optimization approach to high-dimensional sparse quadratic discriminant analysis[J]. The Annals of Statistics, 2021, 49(3): 1537-1568.
- Jiang B, Wang X, Leng C. A direct approach for sparse quadratic discriminant analysis[J]. The Journal of Machine Learning Research, 2018, 19(1): 1098-1134.