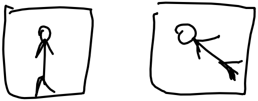
Linear and Quadratic Discriminant Analysis

Statistical Machine Learning (ENGG*6600*08)

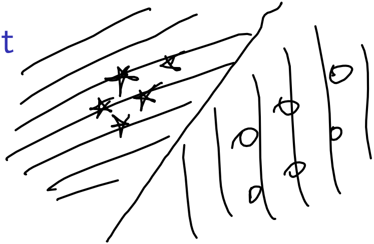
School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh
Fall 2023

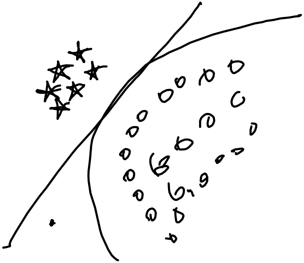


Optimization for the Boundary of Classes

Dataset



- 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.
- We would like to *classify* the space of data using these instances.
- Linear Discriminant Analysis (LDA) and Quadratic discriminant Analysis (QDA) [1] 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, $x \in \mathbb{R}$. Assume we have two classes with the Cumulative Distribution Functions (CDF) $F_1(x)$ and $F_2(x)$, respectively. Let the Probability Density Functions (PDF) of these CDFs be:

$$\begin{cases} f_1(x) = \dfrac{\partial F_1(x)}{\partial x}, & (1) \\ f_2(x) = \dfrac{\partial F_2(x)}{\partial x}, & (2) \end{cases}$$

respectively.

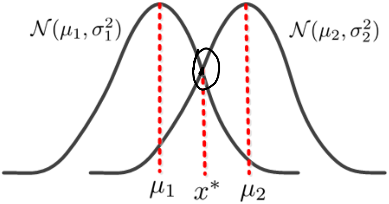
- We assume that the two classes have normal (Gaussian) distribution which is the most common and default distribution in the real-world applications. The mean of one of the two classes is greater than the other one; we assume $\mu_1 < \mu_2$. An instance $x \in \mathbb{R}$ belongs to one of these two classes:

$$x \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{if } x \in \mathcal{C}_1, \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{if } x \in \mathcal{C}_2, \end{cases} \qquad (3)$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ denote the first and second class, respectively.

Optimization for the Boundary of Classes

- 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.
- As we have $\mu_1 < \mu_2$, we can say $\mu_1 < x^* < \mu_2$ as shown in below figure. Therefore, if $x < x^*$ or $x > x^*$ the instance x belongs to the first and second class, respectively. Hence, estimating $x < x^*$ or $x > x^*$ for belonging to the second and first class, respectively, is an error in estimation of the class.


$\mathcal{N}(\mu_1, \sigma_1^2)$
$\mathcal{N}(\mu_2, \sigma_2^2)$
μ_1
x^*
μ_2

Optimization for the Boundary of Classes

- This probability of the error can be stated as:

$$\underbrace{\mathbb{P}(\text{error})} = \overbrace{\mathbb{P}(\underbrace{x > x^*}, \underbrace{x \in \mathcal{C}_1})} \oplus \overbrace{\mathbb{P}(\underbrace{x < x^*}, \underbrace{x \in \mathcal{C}_2})}. \tag{4}$$

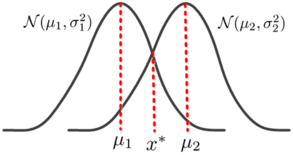
- As we have $\underbrace{\mathbb{P}(A, B)} = \underbrace{\mathbb{P}(A|B)\,\mathbb{P}(B)}$, we can say:

$$\mathbb{P}(\text{error}) = \underbrace{\underbrace{\mathbb{P}(x > x^* \,|\, x \in \mathcal{C}_1)}\,\underbrace{\mathbb{P}(x \in \mathcal{C}_1)}} + \underbrace{\underbrace{\overbrace{\mathbb{P}(x < x^* \,|\, x \in \mathcal{C}_2)}^{\text{likelihood}}}\,\underbrace{\overbrace{\mathbb{P}(x \in \mathcal{C}_2)}^{\text{prior}}}}, \tag{5}$$

which we want to minimize:

$$\underset{x^*}{\text{minimize}} \quad \boxed{\mathbb{P}(\text{error}),} \tag{6}$$

by finding the best boundary of classes, i.e., x^*.



Optimization for the Boundary of Classes

- We found:

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

- According to the definition of CDF, we have:

$$\begin{aligned} &\mathbb{P}(x < c, x \in \mathcal{C}_1) = F_1(c) \implies \mathbb{P}(x > x^*, x \in \mathcal{C}_1) = 1 - F_1(x^*), && (7)\\ &\underbrace{\mathbb{P}(x < x^*, x \in \mathcal{C}_2)} = \underbrace{F_2(x^*)}. && (8)\end{aligned}$$

$$P(x < x^*, x \in C_1) = F_1(x^*)$$

- According to the definition of PDF, we have:

$$\begin{aligned} \mathbb{P}(x \in \mathcal{C}_1) &= f_1(x) = \pi_1, && (9)\\ \mathbb{P}(x \in \mathcal{C}_2) &= f_2(x) = \pi_2, && (10)\end{aligned}$$

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

- Hence, Eqs. (5) and (6) become:

$$\underset{x^*}{\text{minimize}} \quad \underbrace{(1 - F_1(x^*))\,\pi_1 + F_2(x^*)\,\pi_2.} \qquad (11)$$

- We take derivative for the sake of minimization:

$$\frac{\partial\,\mathbb{P}(\text{error})}{\partial x^*} = -\underbrace{f_1(x^*)}\,\pi_1 + \underbrace{f_2(x^*)}\,\pi_2 \overset{\text{set}}{=} 0 \implies \boxed{f_1(x^*)\,\pi_1 = f_2(x^*)\,\pi_2.} \quad \bigstar \qquad (12)$$

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:

$$\overbrace{\mathbb{P}(x \in \mathcal{C}_1 \,|\, X = x)}^{\text{posterior}} \overset{\text{set}}{=} \overbrace{\mathbb{P}(x \in \mathcal{C}_2 \,|\, X = x)}. \tag{13}$$

- According to Bayes rule, the *posterior* is:

$$\star \quad \mathbb{P}(x \in \mathcal{C}_1 \,|\, X = x) = \frac{\overbrace{\mathbb{P}(X = x \,|\, x \in \mathcal{C}_1)}^{\text{PDF}(C_1)}\, \mathbb{P}(x \in \mathcal{C}_1)}{\mathbb{P}(X = x)} = \frac{f_1(x)\, \pi_1}{\underbrace{\sum_{k=1}^{|\mathcal{C}|} \mathbb{P}(X = x \,|\, x \in \mathcal{C}_k)\, \pi_k}}, \tag{14}$$

where $|\mathcal{C}|$ is the number of classes which is two here. The $f_1(x)$ and π_1 are the *likelihood (class conditional)* and *prior* probabilities, respectively, and the denominator is the marginal probability.

- Therefore, Eq. (13) becomes:

$$\frac{f_1(x)\, \pi_1}{\sum_{i=1}^{|\mathcal{C}|} \cancel{\mathbb{P}(X = x | x \in \mathcal{C}_i)\, \pi_i}} \overset{\text{set}}{=} \frac{f_2(x)\, \pi_2}{\sum_{i=1}^{|\mathcal{C}|} \cancel{\mathbb{P}(X = x | x \in \mathcal{C}_i)\, \pi_i}} \implies \boxed{f_1(x)\, \pi_1 = f_2(x)\, \pi_2.} \tag{15}$$

$\star$

Bayes' rule:

$$\underbrace{P(x \in C \mid \enclose{circle}{X})}_{\text{posterior}} = \frac{\overbrace{P(X \mid \enclose{circle}{C})}^{\text{likelihood}} \; \overbrace{P(C)}^{\text{prior \& class}}}{\underbrace{P(X)}_{\downarrow}}$$

prior & data

(<u>marginalized</u> prob

$$\downarrow$$

$$P(X) = \sum_{C} P(X \mid C)\, P(C)$$

Optimization for the Boundary of Classes

- 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:

$$\boxed{f(\boldsymbol{x})} = \underbrace{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\Big),}_{} \quad \bigstar \tag{16}$$

  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 $|.|$ is the determinant of matrix. The $\pi \approx 3.14$ in this equation should not be confused with the π_k (prior) in Eq. (12) or (15).

- Therefore, the Eq. (12) or (15) becomes:

$$\begin{cases} \dfrac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{\boxed{1}}|}} \exp\Big(-\dfrac{(\boldsymbol{x}-\boldsymbol{\mu}_{\boxed{1}})^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\Big)\,\pi_1 \\ = \dfrac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{\boxed{2}}|}} \exp\Big(-\dfrac{(\boldsymbol{x}-\boldsymbol{\mu}_{\boxed{2}})^\top \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\Big)\,\pi_2, \end{cases} \tag{17}$$

where the distributions of the first and second class are $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively.

Linear Discriminant Analysis

Linear Discriminant Analysis for Binary Classification

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

$$\underline{\boldsymbol{\Sigma}_1} = \underline{\boldsymbol{\Sigma}_2} = \underline{\boldsymbol{\Sigma}}. \tag{18}$$

Therefore, the Eq. (17) becomes:

$$\left\{\begin{aligned} & \cancel{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\Big)\, \pi_1 \\ & = \cancel{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\Big)\, \pi_2, \end{aligned}\right.$$

$$\implies \boxed{\exp}\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\Big)\, \pi_1 = \boxed{\exp}\overbrace{\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\Big)}\, \pi_2,$$

$$\overset{(a)}{\implies} \underbrace{-\frac{1}{2}\boxed{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}} + \underbrace{\ln(\pi_1)} = \underbrace{-\frac{1}{2}\boxed{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_2)}} + \underbrace{\ln(\pi_2)},$$

where (a) takes natural logarithm from the sides of equation.

Linear Discriminant Analysis for Binary Classification

$$(x^\top \Sigma^{-1} - \mu_1^\top \Sigma^{-1})(x - \mu_1)$$

- We can simplify this term as:

$$\star\;\left\{\begin{aligned}
\underbrace{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)} &= (\boldsymbol{x}^\top-\boldsymbol{\mu}_1^\top)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\\
&= \underline{\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}} - \boxed{\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1} - \boxed{\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}} + \underline{\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}\\
&\overset{(a)}{=} \underbrace{\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}} + \underbrace{\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1} - \underbrace{2\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}},
\end{aligned}\right. \quad \star \tag{19}$$

where (a) is because $\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ as it is a scalar and $\boldsymbol{\Sigma}^{-1}$ is symmetric so $\boldsymbol{\Sigma}^{-\top} = \boldsymbol{\Sigma}^{-1}$. Thus, we have:

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

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

$$\underbrace{\left(2\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\right)^\top\boldsymbol{x} + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + 2\ln\left(\frac{\pi_2}{\pi_1}\right) = 0}, \tag{20}$$

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

- Therefore, if we consider Gaussian distributions for the two classes where the covariance matrices are assumed to be equal, the decision boundary of classification is a line. Because of linearity of the decision boundary which discriminates the two classes, this method is named *linear discriminant* analysis.

$$(x^T \Sigma^{-1} \mu_1) = (x^T \Sigma^{-1} \mu_1)^T = \mu_1^T \underbrace{\boxed{\Sigma^{-T}}}_{\Sigma^{-1}} \underbrace{(x^T)^T}_{x}$$

$$= \mu_1^T \Sigma^{-1} x$$

Linear Discriminant Analysis for Binary Classification

- For obtaining Eq. (20), we brought the expressions to the right side which was corresponding to the second class; therefore, if we use $\delta(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ as the left-hand-side expression (function) in Eq. (20):

$$\underbrace{\delta(\boldsymbol{x})} := \underbrace{2\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right)^\top \boldsymbol{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \boxed{2\ln(\frac{\pi_2}{\pi_1})},} \tag{21}$$

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

$$\widehat{\mathcal{C}}(x) = \begin{cases} 1, & \text{if } \delta(\boldsymbol{x}) < 0, \\ \underline{2,} & \text{if } \underline{\delta(\boldsymbol{x}) > 0.} \end{cases} \tag{22}$$

- If the priors of two classes are equal, i.e., $\pi_1 = \pi_2$, the Eq. (20) becomes:

$$\underbrace{2\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right)^\top \boldsymbol{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 = 0,} \tag{23}$$

whose left-hand-side expression can be considered as $\delta(\boldsymbol{x})$ in Eq. (22).

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis for Binary Classification

- In Quadratic Discriminant Analysis (QDA), we relax the assumption of equality of the covariance matrices:

$$\underbrace{\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2,}_{} \qquad \log(a^b) = b \log a \tag{24}$$

  which means the covariances are not *necessarily* equal (if they are actually equal, the decision boundary will be linear and QDA reduces to LDA).

- Therefore, the Eq. (17) becomes:

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

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

$$\overset{(a)}{\Longrightarrow} \underbrace{-\frac{d}{2}\ln(2\pi)}_{} - \underbrace{\frac{1}{2}\ln(|\boldsymbol{\Sigma}_1|)}_{} - \underbrace{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \ln(\pi_1)}_{}$$

$$= \underbrace{-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_2|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln(\pi_2),}_{}$$

  where (a) takes natural logarithm from the sides of equation.

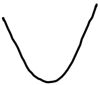
Quadratic Discriminant Analysis for Binary Classification

~~$ax^2+bx+c=0$~~

- Recall Eq. (19):

$$\star \quad \boxed{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)} = \underbrace{\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - 2\,\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}.}$$

- According to Eq. (19), we have:



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

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

$$\left\{\begin{aligned} & \overbrace{\boldsymbol{x}^\top (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{x}} + 2\,(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)^\top \underline{\boldsymbol{x}} \\ & + \underbrace{(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)} + \underbrace{\ln\Big(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\Big)} + \underbrace{2\,\ln(\frac{\pi_2}{\pi_1})} = 0, \end{aligned}\right. \tag{25}$$

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

- Therefore, if we consider Gaussian distributions for the two classes, the decision boundary of classification is quadratic. Because of quadratic decision boundary which discriminates the two classes, this method is named *quadratic discriminant* analysis.

Quadratic Discriminant Analysis for Binary Classification

- For obtaining Eq. (25), we brought the expressions to the right side which was corresponding to the second class; therefore, if we use $\delta(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ as the left-hand-side expression (function) in Eq. (25):

$$\begin{aligned} \underline{\delta(\boldsymbol{x})} := \boldsymbol{x}^\top (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^{-1} \boldsymbol{x} + 2\,(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)^\top \boldsymbol{x} \\ + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) + \ln\Big(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\Big) + \underbrace{2\,\ln\Big(\frac{\boxed{\pi_2}}{\pi_1}\Big)}, \end{aligned} \tag{26}$$

the class of an instance $\boldsymbol{x}$ is estimated as the Eq. (22):

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

- If the priors of two classes are equal, i.e., $\underline{\pi_1 = \pi_2}$, the Eq. (20) becomes:

$$\begin{aligned} \overbrace{\boldsymbol{x}^\top (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^{-1} \boldsymbol{x}} + 2\,(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)^\top \boldsymbol{x} \\ + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) + \ln\Big(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\Big) = 0, \end{aligned} \tag{27}$$

whose left-hand-side expression can be considered as $\delta(\boldsymbol{x})$ in Eq. (22).

$\underline{\text{if } \Sigma_1 = \Sigma_2}$?

LDA and QDA for Multi-class Classification

LDA and QDA for Multi-class Classification

- Now we consider multiple classes, which can be more than two, indexed by $k \in \{1, \dots, |\mathcal{C}|\}$. Recall Eq. (12) or (15) where we are using the scaled posterior, i.e., $f_k(\boldsymbol{x})\, \pi_k$.
- According to Eq. (16), we have:

$$\underbrace{f_k(\boldsymbol{x})\, \pi_k}_{} = \underbrace{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)}{2}\Big)}_{} \underbrace{\pi_k}_{}$$

Taking natural logarithm gives:

$$\underbrace{\ln(f_k(\boldsymbol{x})\, \pi_k)}_{} = \underbrace{\cancel{-\frac{d}{2}\ln(2\pi)} - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln(\pi_k).}_{}$$

- We drop the constant term $-(d/2)\ln(2\pi)$ which is the same for all classes (note that this term is multiplied before taking the logarithm). Thus, the scaled posterior of the k-th class becomes:

$$☆\ \boxed{\delta_k(\boldsymbol{x}) := -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln(\pi_k).} \tag{28}$$

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

$$\underbrace{\widehat{\mathcal{C}}(\boldsymbol{x})}_{} = \arg\max_k \underbrace{\delta_k(\boldsymbol{x})}_{}, \tag{29}$$

because it maximizes the posterior of that class. In this expression, $\delta_k(\boldsymbol{x})$ is Eq. (28).

LDA and QDA for Multi-class Classification

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

$$\underbrace{\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_{|\mathcal{C}|} = \boldsymbol{\Sigma}.} \tag{30}$$

Therefore, the Eq. (28),

$$☆ \quad \delta_k(\boldsymbol{x}) := -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln(\pi_k),$$

becomes:

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

- We drop the constant terms $\underline{-(1/2)\ln(|\boldsymbol{\Sigma}|)}$ and $-(1/2)\,\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ which are the same for all classes (note that before taking the logarithm, the term $-(1/2)\ln(|\boldsymbol{\Sigma}|)$ is multiplied and the term $-(1/2)\,\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ is multiplied as an exponential term). Thus, the scaled posterior of the k-th class becomes:

$$☆ \quad \boxed{\delta_k(\boldsymbol{x}) := \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln(\pi_k).} \tag{31}$$

- In LDA, the class of the instance $\boldsymbol{x}$ is determined by Eq. (29), where $\delta(\boldsymbol{x})$ is Eq. (31), because it maximizes the posterior of that class.

Estimation of Parameters in LDA and QDA

Estimation of Parameters in LDA and QDA

- In LDA and QDA, we have several parameters which are required in order to calculate the posteriors. These parameters are the **means** and the **covariance matrices** of classes and the **priors** of classes.
- The priors of the classes are very tricky to calculate. It is somewhat a chicken and egg problem because we want to know the class probabilities (priors) to estimate the class of an instance but we do not have the priors and should estimate them.
- Usually, the **prior** of the k-th class is estimated according to the sample size of the k-th class:

$$\boxed{\widehat{\pi}_k = \frac{n_k}{n},} \tag{32}$$

where n_k and n are the number of training instances in the k-th class and in total, respectively. This estimation considers Bernoulli distribution for choosing every instance out of the overall training set to be in the k-th class.

Estimation of Parameters in LDA and QDA

- 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:

$$\bigstar \quad \mathbb{R}^d \ni \widehat{\boldsymbol{\mu}}_k = \underbrace{\frac{1}{n_k} \sum_{i=1}^{n} \boxed{\boldsymbol{x}_i}\, \overbrace{\mathbb{I}\big(\underline{\mathcal{C}(\boldsymbol{x}_i) = k}\big)}}, \tag{33}$$

  where $\mathbb{I}(.)$ is the indicator function which is one and zero if its condition is satisfied and not satisfied, respectively.

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

$$\bigstar \quad \mathbb{R}^{d \times d} \ni \widehat{\boldsymbol{\Sigma}}_k = \underbrace{\frac{1}{\boxed{n_k}} \sum_{i=1}^{n} (\underline{\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k})(\underline{\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k})^\top \, \mathbb{I}\big(\mathcal{C}(\boldsymbol{x}_i) = k\big).} \tag{34}$$

$$\boxed{E(\hat{\Sigma}) = \frac{n_k - 1}{n_k} \Sigma}$$

  Or we can use the <u>*unbiased* estimation of the covariance matrix</u>:

$$\bigstar \quad \mathbb{R}^{d \times d} \ni \widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k \boxed{- 1}} \sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^\top \, \mathbb{I}\big(\mathcal{C}(\boldsymbol{x}_i) = k\big). \tag{35}$$

- 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:

$$\mathbb{R}^{d \times d} \ni \underbrace{\widehat{\boldsymbol{\Sigma}}} = \frac{\sum_{k=1}^{|\mathcal{C}|} \boxed{n_k}\, \widehat{\boldsymbol{\Sigma}}_k}{\underbrace{\sum_{r=1}^{|\mathcal{C}|} n_r}} = \frac{\sum_{k=1}^{|\mathcal{C}|} n_k \, \widehat{\boldsymbol{\Sigma}}_k}{n}, \tag{36}$$

  where the weights are the cardinality of the classes.

$$\overbrace{\underbrace{P(c \mid X)}_{\text{posterior}}} = \boxed{\frac{\underbrace{P(X \mid c)}_{\text{likelihood (class conditional)}}\ \underbrace{P(C)}_{\text{prior of class}}}{\underbrace{\cancel{P(X)}}_{\text{prior of data}}}}$$

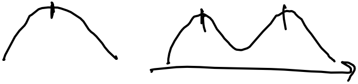
prior of data → marginal

prior of class → marginal

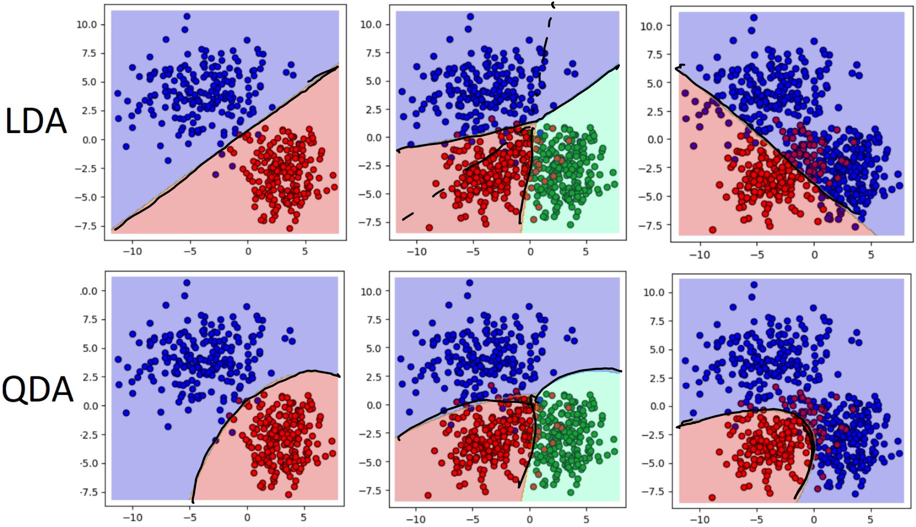
$$P(X \mid C)\, P(C) = P(X, c)$$

$$\boxed{\begin{aligned} P(X) &= \sum_c P(x, c) \\ &= \sum_c P(X \mid c)\, P(c) \end{aligned}}$$

Examples for LDA and QDA



Code of these plots and example in my GitHub:
`https://github.com/bghojogh/Linear-Quadratic-Discriminant-Analysis`



Gabor filter

LDA and QDA are Metric Learning!

central limit theorem

Poisson distribution

LDA and QDA are Metric Learning!

- Recall Eq. (28) which is the scaled posterior for the QDA:

$$★ \quad \delta_k(\boldsymbol{x}) := -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln(\pi_k).$$

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

$$\boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_{|\mathcal{C}|} = \boldsymbol{I}, \tag{37}$$

which means that all the classes are assumed to be spherically distributed in the d dimensional space. After this assumption, the Eq. (28) becomes:

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}-\boldsymbol{\mu}_k) + \cancel{\ln(\pi_k)}, \tag{38}$$

because $|\boldsymbol{I}| = 1$, $\ln(1) = 0$, and $\boldsymbol{I}^{-1} = \boldsymbol{I}$.

- If we assume that the priors are all equal, the term $\ln(\pi_k)$ is constant and can be dropped:

$$\delta_k(\boldsymbol{x}) = \boxed{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}-\boldsymbol{\mu}_k)} = -\frac{1}{2}d_k^2, \tag{39}$$

where d_k is the Euclidean distance from the mean of the k-th class:

$$★ \quad d_k = ||\boldsymbol{x}-\boldsymbol{\mu}_k||_2 = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top(\boldsymbol{x}-\boldsymbol{\mu}_k)}. \tag{40}$$

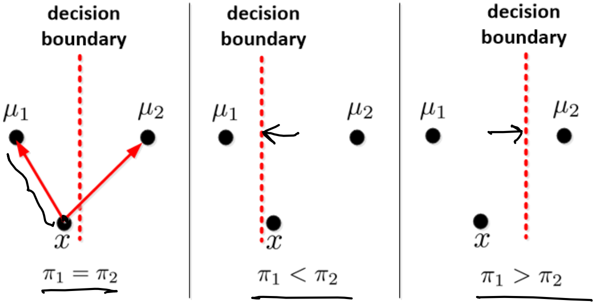
- Thus, the QDA or LDA reduce to simple Euclidean distance from the means of classes if the covariance matrices are all identity matrix and the priors are equal. Simple distance from the mean of classes is one of the simplest classification methods where the used metric is Euclidean distance.

LDA and QDA are Metric Learning!

- Now, consider the case where still the **covariance matrices are all identity matrix** but the **priors are not equal**. In this case, we have Eq. (38):

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

- If we take an exponential (inverse of logarithm) from this expression, the π_k becomes a scale factor (weight). This means that we still are using distance metric to measure the distance of an instance from the means of classes but we are **scaling the distances by the priors of classes**. If a class happens more, i.e., its **prior is larger**, it must have a larger posterior so we **reduce the distance** from the mean of its class. In other words, we **move the decision boundary** according to the prior of classes.



LDA and QDA are Metric Learning!

- As the next step, consider a more general case where the **covariance matrices are not equal** as we have in **QDA**.
- We apply Singular Value Decomposition (SVD) to the covariance matrix of the k-th class:

$$\boldsymbol{\Sigma}_k = \underline{\boldsymbol{U}_k \boldsymbol{\Lambda}_k \boldsymbol{U}_k^\top}, \leftarrow$$

where the left and right matrices of singular vectors are equal because the covariance matrix is symmetric. Therefore:

$$\boldsymbol{\Sigma}_k^{-1} = \boxed{\boldsymbol{U}_k \boldsymbol{\Lambda}_k^{-1} \boldsymbol{U}_k^\top,}$$

$$U^{-T} = U^{T^T} = U$$

where $\boldsymbol{U}_k^{-1} = \boldsymbol{U}_k^\top$ because it is an orthogonal matrix.

- Therefore, we can simplify the following term:

$$(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) = (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{U}_k \boldsymbol{\Lambda}_k^{-1} \boldsymbol{U}_k^\top (\boldsymbol{x} - \boldsymbol{\mu}_k)$$
$$= \underbrace{(\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)^\top}_{} \boldsymbol{\Lambda}_k^{-1} \underbrace{(\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)}_{}.$$

$$(x^T U - \mu^T U) = (x^T - \mu^T) U = (x - \mu)^T U$$

LDA and QDA are Metric Learning!

$A = A^{\top/2} A^{1/2}$

- We found:

$$\ast \quad (\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) = (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k^{-1} (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k).$$

- As $\boldsymbol{\Lambda}_k^{-1}$ is a diagonal matrix with non-negative elements (because it is covariance), we can decompose it as:

$$\boldsymbol{\Lambda}_k^{-1} = \underbrace{\boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{\Lambda}_k^{-1/2}}.$$

Therefore:

$$(\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k^{-1} (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k) = (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{\Lambda}_k^{-1/2} (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)$$

$$\overset{(a)}{=} (\boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{\mu}_k)^\top (\boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{\mu}_k),$$

$$= \boxed{\big(\phi_k(x) - \phi_k(\mu_k)\big)^\top \big(\phi_k(x) - \phi_k(\mu_k)\big)}$$

where (a) is because $\boldsymbol{\Lambda}_k^{-\top/2} = \boldsymbol{\Lambda}_k^{-1/2}$ because it is diagonal.

- We define the following transformation:

$$\phi_k : \boldsymbol{x} \mapsto \boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{x}, \tag{41}$$

which also results in the transformation of the mean: $\phi_k : \boldsymbol{\mu} \mapsto \boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{\mu}$.

LDA and QDA are Metric Learning!

- We had:

$$\underline{\phi_k} : \boldsymbol{x} \mapsto \boldsymbol{\Lambda}_k^{-1/2} \boldsymbol{U}_k^\top \boldsymbol{x}.$$

- Therefore, the Eq. (28),

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

can be restated as:

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2} \ln(|\boldsymbol{\Sigma}_k|) - \frac{1}{2} \big(\phi_k(\boldsymbol{x}) - \phi_k(\boldsymbol{\mu}_k)\big)^\top \big(\phi_k(\boldsymbol{x}) - \phi_k(\boldsymbol{\mu}_k)\big) + \ln(\pi_k). \tag{42}$$

- Ignoring the terms $-(1/2)\ln(|\boldsymbol{\Sigma}_k|)$ and $\ln(\pi_k)$, we can see that the **transformation** has **changed the covariance matrix of the class to identity matrix**. Therefore, the QDA (and also LDA) can be seen as **simple comparison of distances from the means of classes after applying a transformation to the data of every class**. In other words, we are **learning the metric using the SVD of covariance matrix of every class**. Thus, LDA and QDA can be seen as **metric learning** [2, 3] in a perspective.

LDA and QDA are Metric Learning!

$$(x-\mu)^{T}(x-\mu)$$

$$(x-\mu)^{T}\Sigma^{-1}(x-\mu) \rightarrow A$$

- Note that in metric learning, a valid distance metric is defined as [2]:

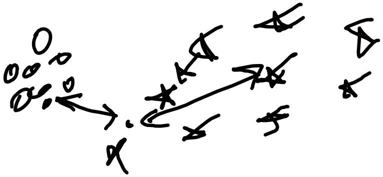
$$\bigstar\ d_{\boldsymbol{A}}^{2}(\boldsymbol{x},\boldsymbol{\mu}_k) := ||\boldsymbol{x}-\boldsymbol{\mu}_k||_{\boldsymbol{A}}^{2} = \boxed{(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\top}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu}_k),} \tag{43}$$

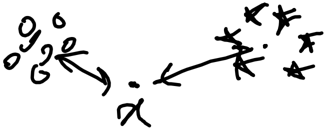
  where $\boldsymbol{A}$ is a positive semi-definite matrix, i.e., $\boldsymbol{A} \succeq 0$.
- In QDA, we are also using $(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\top}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)$.
- The covariance matrix is positive semi-definite according to the characteristics of covariance matrix. Moreover, according to characteristics of a positive semi-definite matrix, the inverse of a positive semi-definite matrix is positive semi-definite so $\boldsymbol{\Sigma}_k^{-1} \succeq 0$.
- Therefore, QDA is using metric learning.

$$(x-\mu)^{T}A(x-\mu) = (x-\mu)^{T}UU^{T}(x-\mu)$$

$$(U^{T}x-U^{T}\mu)^{T}(U^{T}x-U^{T}\mu)$$

$$A = V\Lambda V^{T} = \underbrace{V\Lambda^{\frac{1}{2}}}_{U}\Lambda^{\frac{1}{2}}V^{T} = UU^{T}$$


x


x

Acknowledgment

- For more information on LDA and QDA, refer to our tutorial paper: Benyamin Ghojogh, Mark Crowley. "Linear and quadratic discriminant analysis: Tutorial." arXiv preprint arXiv:1906.02590 (2019). [4]
- Code of my LDA and QDA:
`https://github.com/bghojogh/Linear-Quadratic-Discriminant-Analysis`
- For more information on LDA and QDA, see the book: Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, Jerome H. Friedman. "The elements of statistical learning: data mining, inference, and prediction". Vol. 2. New York: springer, 2009 [1].
- Some slides of this slide deck are inspired by teachings of Prof. Ali Ghodsi and Prof. Mu Zhu at University of Waterloo, Department of Statistics and Prof. Hoda Mohammadzade at Sharif University of Technology, Department of Electrical Engineering.

References

[1] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2.
Springer, 2009.

[2] L. Yang and R. Jin, “Distance metric learning: A comprehensive survey,” tech. rep., Department of Computer Science and Engineering, Michigan State University, 2006.

[3] B. Kulis, “Metric learning: A survey,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[4] B. Ghojogh and M. Crowley, “Linear and quadratic discriminant analysis: Tutorial,” *arXiv preprint arXiv:1906.02590*, 2019.