Linear and Quadratic Discriminant Analysis: Tutorial

Benyamin Ghojogh BGHOJOGH@UWATERLOO.CA
Department of Electrical and Computer Engineering,
Machine Learning Laboratory, University of Waterloo, Waterloo, ON, Canada

Mark Crowley MCROWLEY@UWATERLOO.CA
Department of Electrical and Computer Engineering,
Machine Learning Laboratory, University of Waterloo, Waterloo, ON, Canada

Abstract

This tutorial explains Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) as two fundamental classification methods in statistical and probabilistic learning. We start with the optimization of decision boundary on which the posteriors are equal. Then, LDA and QDA are derived for binary and multiple classes. The estimation of parameters in LDA and QDA are also covered. Then, we explain how LDA and QDA are related to metric learning, kernel principal component analysis, Mahalanobis distance, logistic regression, Bayes optimal classifier, Gaussian naive Bayes, and likelihood ratio test. We also prove that LDA and Fisher discriminant analysis are equivalent. We finally clarify some of the theoretical concepts with simulations we provide.

1. 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. We would like to *classify* the space of data using these instances. Linear Discriminant Analysis (LDA) and Quadratic discriminant Analysis (QDA) (Friedman et al., 2009) are two well-known *supervised classification* methods in statistical and probabilistic learning. This paper is a tutorial for these two classifiers where the theory for binary and multi-class classification are detailed. Then, relations of LDA and QDA to metric learning, kernel Principal Component Analysis (PCA), Fisher Discriminant Analysis (FDA), logistic regression, Bayes optimal classifier, Gaussian naive Bayes, and Likelihood Ratio Test (LRT) are explained for better understanding of these two fundamental methods. Finally, some experiments on synthetic datasets are reported and analyzed for illustration.

2. 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:

$$f_1(x) = \frac{\partial F_1(x)}{\partial x}, \tag{1}$$

$$f_2(x) = \frac{\partial F_2(x)}{\partial x}, \tag{2}$$

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:

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

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. As we have $\mu_1 < \mu_2$, 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. Hence, 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:

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

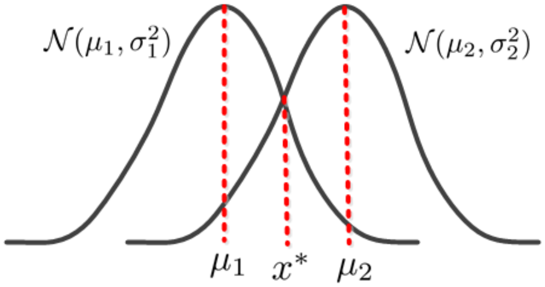


Figure 1. Two Gaussian density functions where they are equal at the point x^*.

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

$$\begin{aligned}\mathbb{P}(\text{error}) = \,&\mathbb{P}(x > x^* \,|\, x \in \mathcal{C}_1)\,\mathbb{P}(x \in \mathcal{C}_1) \\ &+ \mathbb{P}(x < x^* \,|\, x \in \mathcal{C}_2)\,\mathbb{P}(x \in \mathcal{C}_2),\end{aligned} \tag{5}$$

which we want to minimize:

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

by finding the best boundary of classes, i.e., x^*.
According to the definition of CDF, we have:

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

According to the definition of PDF, we have:

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

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:

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

We take derivative for the sake of minimization:

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

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

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

According to Bayes rule, the *posterior* is:

$$\begin{aligned}\mathbb{P}(x \in \mathcal{C}_1 \,|\, X = x) &= \frac{\mathbb{P}(X = x \,|\, x \in \mathcal{C}_1)\,\mathbb{P}(x \in \mathcal{C}_1)}{\mathbb{P}(X = x)} \\ &= \frac{f_1(x)\,\pi_1}{\sum_{k=1}^{|\mathcal{C}|} \mathbb{P}(X = x \,|\, x \in \mathcal{C}_k)\,\pi_k},\end{aligned} \tag{14}$$

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:

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

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:

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

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:

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

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.

3. Linear Discriminant Analysis for Binary Classification

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

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

Therefore, the Eq. (17) becomes:

$$\begin{aligned}
&\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\Big)\,\pi_1 \\
&= \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\Big)\,\pi_2, \\
&\implies \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\Big)\,\pi_1 \\
&\quad= \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\Big)\,\pi_2, \\
&\overset{(a)}{\implies} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \ln(\pi_1) \\
&\quad= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln(\pi_2),
\end{aligned}$$

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

We can simplify this term as:

$$\begin{aligned}
&(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) = (\boldsymbol{x}^\top-\boldsymbol{\mu}_1^\top)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) \\
&= \boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 \\
&\overset{(a)}{=} \boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - 2\,\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x},
\end{aligned} \tag{19}$$

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:

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

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

$$\begin{aligned}
&2\,\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\big)^\top\boldsymbol{x} \\
&\quad+ \big(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\big)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big) + 2\,\ln(\frac{\pi_2}{\pi_1}) = 0,
\end{aligned} \tag{20}$$

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

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

$$\begin{aligned}
\delta(\boldsymbol{x}) &:= 2\,\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\big)^\top\boldsymbol{x} \\
&\quad+ \big(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\big)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big) + 2\,\ln(\frac{\pi_2}{\pi_1}),
\end{aligned} \tag{21}$$

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

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

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

$$\begin{aligned}
&2\,\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\big)^\top\boldsymbol{x} \\
&\quad+ \big(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\big)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = 0,
\end{aligned} \tag{23}$$

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

4. Quadratic Discriminant Analysis for Binary Classification

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

$$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2, \tag{24}$$

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:

$$\begin{aligned}
&\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_1|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\Big)\,\pi_1 \\
&= \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_2|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\Big)\,\pi_2, \\
&\overset{(a)}{\implies} -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_1|) \\
&\quad - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \ln(\pi_1) \\
&\quad = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_2|) \\
&\quad - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln(\pi_2),
\end{aligned}$$

where (a) takes natural logarithm from the sides of equation. According to Eq. (19), we have:

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

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

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

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

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):

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

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

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

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

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

5. 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:

$$\begin{aligned}&f_k(\boldsymbol{x})\,\pi_k \\ &= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}{2}\Big)\pi_k.\end{aligned}$$

Taking natural logarithm gives:

$$\begin{aligned}\ln(f_k(\boldsymbol{x})\,\pi_k) = &-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_k|) \\ &- \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln(\pi_k).\end{aligned}$$

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:

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

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

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

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

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

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

Therefore, the Eq. (28) becomes:

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

We drop the constant terms $-(1/2)\ln(|\boldsymbol{\Sigma}|)$ 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:

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

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.

In conclusion, QDA and LDA deal with maximizing the *posterior* of classes but work with the *likelihoods (class conditional)* and *priors*.

6. 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:

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

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.

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:

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

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:

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

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

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

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:

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

where the weights are the cardinality of the classes.

7. LDA and QDA are Metric Learning!

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

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

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:

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

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

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

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.

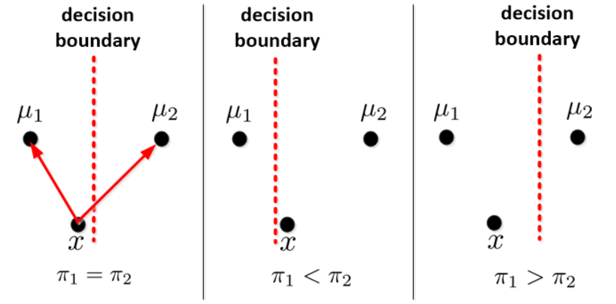


Figure 2. The QDA and LDA where the covariance matrices are identity matrix. For equal priors, the QDA and LDA reduce to simple classification using Euclidean distance from means of classes. Changing the prior modifies the location of decision boundary where even one point can be classified differently for different priors.

The Eq. (39) has a very interesting message. We know that in metric Multi-Dimensional Scaling (MDS) (Cox & Cox, 2000) and kernel Principal Component Analysis (PCA), we have (see (Ham et al., 2004) and Chapter 2 in (Strange & Zwiggelaar, 2014)):

$$\boldsymbol{K} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{D}\boldsymbol{H}, \tag{41}$$

where $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is the distance matrix whose elements are the distances between the data instances, $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix over the data instances, $\mathbb{R}^{n\times n} \ni \boldsymbol{H} := \boldsymbol{I} - (1/n)\boldsymbol{1}\boldsymbol{1}^\top$ is the centering matrix, and $\mathbb{R}^n \ni \boldsymbol{1} := [1, 1, \dots, 1]^\top$. If the elements of the distance matrix $\boldsymbol{D}$ are obtained using Euclidean distance, the MDS is equivalent to Principal Component Analysis (PCA) (Jolliffe, 2011).

Comparing Eqs. (39) and (41) shows an interesting connection between the posterior of a class in QDA and the kernel over the the data instances of the class. In this comparison, the Eq. (41) should be considered for a class and not the entire data, so $\boldsymbol{K} \in \mathbb{R}^{n_k\times n_k}$, $\boldsymbol{D} \in \mathbb{R}^{n_k\times n_k}$, and $\boldsymbol{H} \in \mathbb{R}^{n_k\times n_k}$.

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). 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 (see Fig. 2).

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:

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

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

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

where $\boldsymbol{U}_k^{-1} = \boldsymbol{U}_k^\top$ because it is an orthogonal matrix. Therefore, we can simplify the following term:

$$\begin{aligned}&(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\\ &= (\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{U}_k\, \boldsymbol{\Lambda}_k^{-1}\, \boldsymbol{U}_k^\top (\boldsymbol{x}-\boldsymbol{\mu}_k)\\ &= (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k^{-1}\, (\boldsymbol{U}_k^\top \boldsymbol{x} - \boldsymbol{U}_k^\top \boldsymbol{\mu}_k).\end{aligned}$$

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

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

Therefore:

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

where (a) is because $\boldsymbol{\Lambda}_k^{-\top/2} = \boldsymbol{\Lambda}_k^{-1/2}$ because it is diagonal. We define the following transformation:

$$\phi_k : \boldsymbol{x} \mapsto \boldsymbol{\Lambda}_k^{-1/2} \boldsymbol{U}_k^\top \boldsymbol{x}, \tag{42}$$

which also results in the transformation of the mean: $\phi_k : \boldsymbol{\mu} \mapsto \boldsymbol{\Lambda}_k^{-1/2} \boldsymbol{U}_k^\top \boldsymbol{\mu}$. Therefore, the Eq. (28) can be restated as:

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= -\frac{1}{2} \ln(|\boldsymbol{\Sigma}_k|)\\ &- \frac{1}{2}\big(\phi_k(\boldsymbol{x}) - \phi_k(\boldsymbol{\mu}_k)\big)^\top \big(\phi_k(\boldsymbol{x}) - \phi_k(\boldsymbol{\mu}_k)\big) + \ln(\pi_k).\end{aligned} \tag{43}$$

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* (Yang & Jin, 2006; Kulis, 2013) in a perspective. Note that in metric learning, a valid distance metric is defined as (Yang & Jin, 2006):

$$d_{\boldsymbol{A}}^2(\boldsymbol{x}, \boldsymbol{\mu}_k) := ||\boldsymbol{x} - \boldsymbol{\mu}_k||_{\boldsymbol{A}}^2 = (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{A}\, (\boldsymbol{x} - \boldsymbol{\mu}_k), \tag{44}$$

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 (and as will be discussed in next section, it can be seen as a *manifold learning* method, too).

It is also noteworthy that the QDA and LDA can also be seen as *Mahalanobis distance* (McLachlan, 1999; De Maesschalck et al., 2000) which is also a metric:

$$d_M^2(\boldsymbol{x}, \boldsymbol{\mu}) := ||\boldsymbol{x} - \boldsymbol{\mu}||_M^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{45}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the cloud of data whose mean is $\boldsymbol{\mu}$. The intuition of Mahalanobis distance is that if we have several data clouds (e.g., classes), the distance from the class with larger variance should be scaled down because that class is taking more of the space so it is more probable to happen. The scaling down shows in the inverse of covariance matrix. Comparing $(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)$ in QDA or LDA with Eq. (45) shows that QDA and LDA are sort of using Mahalanobis distance.

8. LDA $\overset{?}{\equiv}$ FDA

In the previous section, we saw that LDA and QDA can be seen as metric learning. We know that metric learning can be seen as a family of manifold learning methods. We briefly explain the reason of this assertion: As $\boldsymbol{A} \succeq 0$, we can say $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{U}^\top \succeq 0$. Therefore, Eq. (44) becomes:

$$\begin{aligned}||\boldsymbol{x} - \boldsymbol{\mu}_k||_{\boldsymbol{A}}^2 &= (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{U}\boldsymbol{U}^\top\, (\boldsymbol{x} - \boldsymbol{\mu}_k)\\ &= (\boldsymbol{U}^\top \boldsymbol{x} - \boldsymbol{U}^\top \boldsymbol{\mu}_k)^\top (\boldsymbol{U}^\top \boldsymbol{x} - \boldsymbol{U}^\top \boldsymbol{\mu}_k),\end{aligned}$$

which means that metric learning can be seen as comparison of simple Euclidean distances after the transformation $\phi : \boldsymbol{x} \mapsto \boldsymbol{U}^\top \boldsymbol{x}$ which is a projection into a subspace with projection matrix $\boldsymbol{U}$. Thus, metric learning is a manifold learning approach. This gives a hint that the Fisher Discriminant Analysis (FDA) (Fisher, 1936; Welling, 2005), which is a manifold learning approach (Tharwat et al., 2017), might have a connection to LDA; especially, because the names FDA and LDA are often used interchangeably in the literature. Actually, other names of FDA are Fisher LDA (FLDA) and even LDA.

We know that if we project (transform) the data of a class using a projection vector $\boldsymbol{u} \in \mathbb{R}^p$ to a p dimensional subspace ($p \leq d$), i.e.:

$$\boldsymbol{x} \mapsto \boldsymbol{u}^\top \boldsymbol{x}, \tag{46}$$

for all data instances of the class, the mean and the covariance matrix of the class are transformed as:

$$\boldsymbol{\mu} \mapsto \boldsymbol{u}^\top \boldsymbol{\mu}, \tag{47}$$

$$\boldsymbol{\Sigma} \mapsto \boldsymbol{u}^\top \boldsymbol{\Sigma}\, \boldsymbol{u}, \tag{48}$$

because of characteristics of mean and variance.

The Fisher criterion (Xu & Lu, 2006) is the ratio of the between-class variance, σ_b^2, and within-class variance, σ_w^2:

$$f := \frac{\sigma_b^2}{\sigma_w^2} = \frac{(\boldsymbol{u}^\top \boldsymbol{\mu}_2 - \boldsymbol{u}^\top \boldsymbol{\mu}_1)^2}{\boldsymbol{u}^\top \boldsymbol{\Sigma}_2\, \boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{\Sigma}_1\, \boldsymbol{u}} = \frac{\big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2}{\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u}}. \tag{49}$$

The FDA maximizes the Fisher criterion:

$$\underset{\boldsymbol{u}}{\text{maximize}} \quad \frac{\big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2}{\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u}}, \tag{50}$$

which can be restated as:

$$\begin{aligned} &\underset{\boldsymbol{u}}{\text{maximize}} && \big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2, \\ &\text{subject to} && \boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u} = 1, \end{aligned} \tag{51}$$

according to Rayleigh-Ritz quotient method (Croot, 2005). The Lagrangian (Boyd & Vandenberghe, 2004) is:

$$\mathcal{L} = \big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2 - \lambda\big(\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u} - 1\big),$$

where λ is the Lagrange multiplier. Equating the derivative of $\mathcal{L}$ to zero gives:

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial \boldsymbol{u}} = 2\, (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2\, \boldsymbol{u} - 2\, \lambda\, (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u} \overset{\text{set}}{=} \boldsymbol{0} \\ &\implies (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2\, \boldsymbol{u} = \lambda\, (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u}, \end{aligned}$$

which is a generalized eigenvalue problem $\big((\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2, (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\big)$ according to (Ghojogh et al., 2019b). The projection vector is the eigenvector of $(\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2$; therefore, we can say:

$$\boldsymbol{u} \propto (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2.$$

In LDA, the equality of covariance matrices is assumed. Thus, according to Eq. (18), we can say:

$$\boldsymbol{u} \propto (2\, \boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2 \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2. \tag{52}$$

According to Eq. (46), we have:

$$\boldsymbol{u}^\top \boldsymbol{x} \propto \big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^2\big)^\top \boldsymbol{x}. \tag{53}$$

Comparing Eq. (53) with Eq. (23) shows that LDA and FDA are equivalent up to a scaling factor $\big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\big)$ (note that this term is multiplied as an exponential factor before taking logarithm to obtain Eq. (23), so this term a scaling factor). Hence, we can say:

$$\text{LDA} \equiv \text{FDA}. \tag{54}$$

In other words, FDA projects into a subspace. On the other hand, according to Section 7, LDA can be seen as a metric learning with a subspace where the Euclidean distance is used after projecting onto that subspace. *The two subspaces of FDA and LDA are the same subspace.* It should be noted that in manifold (subspace) learning, the scale does not matter because all the distances scale similarly.

Note that LDA assumes *one* (and not several) Gaussian for every class and so does the FDA. That is why FDA faces problem for multi-modal data (Sugiyama, 2007).

9. Relation to Logistic Regression

According to Eqs. (16) and (32), Gaussian and Bernoulli distributions are used for likelihood (class conditional) and prior, respectively, in LDA and QDA. Thus, we are making assumptions for the likelihood and prior, although we finally work with posterior in LDA and QDA according to Eq. (15). *Logistic regression* (Kleinbaum et al., 2002) says why do we make assumptions on the likelihood and prior when we want to work on posterior finally. Let us make assumption directly for the posterior.

In logistic regression, first a linear function is applied to the data to have $\boldsymbol{\beta}^\top \boldsymbol{x}'$ where $\mathbb{R}^{d+1} \ni \boldsymbol{x}' = [\boldsymbol{x}^\top, 1]^\top$ and $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ include the intercept. Then, logistic function is used in order to have a value in range $(0, 1)$ to simulate probability. Therefore, in logistic regression, the posterior is assumed to be:

$$\begin{aligned} &\mathbb{P}(\mathcal{C}(\boldsymbol{x}) \,|\, X = \boldsymbol{x}) \\ &= \Big(\frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}')}{1 + \exp(\boldsymbol{\beta}^\top \boldsymbol{x}')}\Big)^{\mathcal{C}(\boldsymbol{x})} \Big(\frac{1}{1 + \exp(\boldsymbol{\beta}^\top \boldsymbol{x}')}\Big)^{1 - \mathcal{C}(\boldsymbol{x})}, \end{aligned} \tag{55}$$

where $\mathcal{C}(\boldsymbol{x}) \in \{-1, +1\}$ for the two classes. Logistic regression considers the coefficient $\boldsymbol{\beta}$ as the parameter to be optimized and uses Newton's method (Boyd & Vandenberghe, 2004) for the optimization. Therefore, in summary, logistic regression makes assumption on the posterior while LDA and QDA make assumption on likelihood and prior.

10. Relation to Bayes Optimal Classifier and Gaussian Naive Bayes

The Bayes classifier maximizes the posteriors of the classes (Murphy, 2012):

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_{k}\, \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_k \,|\, X = \boldsymbol{x}). \tag{56}$$

According to Eq. (14) and Bayes rule, we have:

$$\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_k \,|\, X = \boldsymbol{x}) \propto \mathbb{P}(X = \boldsymbol{x} \,|\, \boldsymbol{x} \in \mathcal{C}_k)\, \underbrace{\mathbb{P}(\boldsymbol{x} \in \mathcal{C}_k)}_{\pi_k}, \tag{57}$$

where the denominator of posterior (the marginal) which is:

$$\mathbb{P}(X = \boldsymbol{x}) = \sum_{r=1}^{|\mathcal{C}|} \mathbb{P}(X = \boldsymbol{x} \,|\, \boldsymbol{x} \in \mathcal{C}_r)\, \pi_r,$$

is ignored because it is not dependent on the classes $\mathcal{C}_1$ to $\mathcal{C}_{|\mathcal{C}|}$.

According to Eq. (57), the posterior can be written in terms of likelihood and prior; therefore, Eq. (56) can be restated as:

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_{k}\; \pi_k\, \mathbb{P}(X = \boldsymbol{x} \,|\, \boldsymbol{x} \in \mathcal{C}_k). \tag{58}$$

Note that the Bayes classifier does not make any assumption on the posterior, prior, and likelihood, unlike LDA and QDA which assume the uni-modal Gaussian distribution for the likelihood (and we *may* assume Bernoulli distribution for the prior in LDA and QDA according to Eq. (32)). Therefore, we can say the difference of Bayes and QDA is in assumption of *uni-modal* Gaussian distribution for the likelihood (class conditional); hence, if the likelihoods are already uni-modal Gaussian, the Bayes classifier reduces to QDA. Likewise, the difference of Bayes and LDA is in assumption of Gaussian distribution for the likelihood (class conditional) and equality of covariance matrices of classes; thus, if the likelihoods are already Gaussian and the covariance matrices are already equal, the Bayes classifier reduces to LDA.

It is noteworthy that the Bayes classifier is an optimal classifier because it can be seen as an ensemble of hypotheses (models) in the hypothesis (model) space and no other ensemble of hypotheses can outperform it (see Chapter 6, Page 175 in (Mitchell, 1997)). In the literature, it is referred to as *Bayes optimal classifier*. To better formulate the explained statements, the Bayes optimal classifier estimates the class as:

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_{\mathcal{C}_k \in \mathcal{C}} \sum_{h_j \in \mathcal{H}} \mathbb{P}(\mathcal{C}_k \,|\, h_j)\, \mathbb{P}(\mathcal{D} \,|\, h_j)\, \mathbb{P}(h_j), \tag{59}$$

where $\mathcal{C} := \{\mathcal{C}_1, \dots, \mathcal{C}_{|\mathcal{C}|}\}$, $\mathcal{D} := \{\boldsymbol{x}_i\}_{i=1}^n$ is the training set, h_j is a hypothesis for estimating the class of instances, and $\mathcal{H}$ is the hypothesis space including all possible hypotheses.

According to Bayes rule, similar to what we had for Eq. (57), we have:

$$\mathbb{P}(h_j \,|\, \mathcal{D}) \propto \mathbb{P}(\mathcal{D} \,|\, h_j)\, \mathbb{P}(h_j).$$

Therefore, Eq. (59) becomes (Mitchell, 1997):

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_{\mathcal{C}_k \in \mathcal{C}} \sum_{h_j \in \mathcal{H}} \mathbb{P}(\mathcal{C}_k \,|\, h_j)\, \mathbb{P}(h_j \,|\, \mathcal{D}), \tag{60}$$

In conclusion, the Bayes classifier is optimal. Therefore, if the likelihoods of classes are Gaussian, QDA is an optimal classifier and if the likelihoods are Gaussian and the covariance matrices are equal, the LDA is an optimal classifier. Often, the distributions in the natural life are Gaussian; especially, because of central limit theorem (Hazewinkel, 2001), the summation of independent and identically distributed (iid) variables is Gaussian and the signals usually add in the real world. This explains why LDA and QDA are very effective classifiers in machine learning. We also saw that FDA is equivalent to LDA. Thus, the reason of effectiveness of the powerful FDA classifier becomes clear. We have seen the very successful performance of FDA and LDA in different applications, such as face recognition (Belhumeur et al., 1997; Etemad & Chellappa, 1997; Zhao et al., 1999), action recognition (Ghojogh et al., 2017; Mokari et al., 2018), and EEG classification (Malekmohammadi et al., 2019).

Implementing Bayes classifier is difficult in practice so we approximate it by *naive Bayes* (Zhang, 2004). If x_j denotes the j-th dimension (feature) of $\boldsymbol{x} = [x_1, \dots, x_d]^\top$, Eq. (58) is restated as:

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_{k}\; \pi_k\, \mathbb{P}(x_1, x_2, \dots, x_d \,|\, \boldsymbol{x} \in \mathcal{C}_k). \tag{61}$$

The term $\mathbb{P}(x_1, x_2, \dots, x_d \,|\, \boldsymbol{x} \in \mathcal{C}_k)$ is very difficult to compute as the features are possibly correlated. Naive Bayes relaxes this possibility and naively assumes that the features are conditionally independent ($\perp\!\!\!\perp$) when they are conditioned on the class:

$$\begin{aligned}&\mathbb{P}(x_1, x_2, \dots, x_d \,|\, \boldsymbol{x} \in \mathcal{C}_k) \\ &\overset{\perp\!\!\!\perp}{\approx} \mathbb{P}(x_1 \,|\, \mathcal{C}_k)\, \mathbb{P}(x_2 \,|\, \mathcal{C}_k) \cdots \mathbb{P}(x_d \,|\, \mathcal{C}_k) = \prod_{j=1}^{d} \mathbb{P}(x_j \,|\, \mathcal{C}_k).\end{aligned}$$

Therefore, Eq. (61) becomes:

$$\widehat{\mathcal{C}}(\boldsymbol{x}) = \arg\max_{k}\; \pi_k \prod_{j=1}^{d} \mathbb{P}(x_j \,|\, \mathcal{C}_k). \tag{62}$$

In *Gaussian naive Bayes*, univariate Gaussian distribution is assumed for the likelihood (class conditional) of every feature:

$$\mathbb{P}(x_j \,|\, \mathcal{C}_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\Big(-\frac{(x_j - \mu_k)^2}{2\sigma_k^2}\Big), \tag{63}$$

where the mean and unbiased variance are estimated as:

$$\mathbb{R} \ni \widehat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} x_{i,j}\, \mathbb{I}\big(\mathcal{C}(\boldsymbol{x}_i) = k\big), \tag{64}$$

$$\mathbb{R} \ni \widehat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n} (x_{i,j} - \widehat{\mu}_k)^2\, \mathbb{I}\big(\mathcal{C}(\boldsymbol{x}_i) = k\big), \tag{65}$$

where $x_{i,j}$ denotes the j-th feature of the i-th training instance. The prior can again be estimated using Eq. (32).

According to Eqs. (62) and (63), Gaussian naive Bayes is equivalent to QDA where the covariance matrices are *diagonal*, 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. Moreover, it is obvious that Gaussian naive Bayes and QDA are equivalent for *one* dimensional data. Comparing to LDA, the Gaussian naive Bayes is equivalent to LDA if the covariance matrices are diagonal and they are all equal, i.e., $\sigma_1^2 = \dots = \sigma_{|\mathcal{C}|}^2$; therefore, LDA and Gaussian naive Bayes have their own assumptions, one on the off-diagonal of covariance matrices and the other one on equality of the covariance matrices. As Gaussian naive Bayes has some level of optimality (Zhang, 2004), it becomes clear why LDA and QDA are such effective classifiers.

11. Relation to Likelihood Ratio Test

Consider two hypotheses for estimating some parameter, a null hypothesis H_0 and an alternative hypothesis H_A. The probability $\mathbb{P}(\text{reject } H_0 \,|\, H_0)$ is called type 1 error, false positive error, or false alarm error. The probability $\mathbb{P}(\text{accept } H_0 \,|\, H_A)$ is called type 2 error or false negative error. The $\mathbb{P}(\text{reject } H_0 \,|\, H_0)$ is also called *significance level*, while $1 - \mathbb{P}(\text{accept } H_0 \,|\, H_A) = \mathbb{P}(\text{reject } H_0 \,|\, H_A)$ is called *power*.

If $L(\theta_A)$ and $L(\theta_0)$ are the likelihoods (probabilities) for the alternative and null hypotheses, the likelihood ratio is:

$$\Lambda = \frac{L(\theta_A)}{L(\theta_0)} = \frac{f(\boldsymbol{x};\theta_A)}{f(\boldsymbol{x};\theta_0)}. \tag{66}$$

The Likelihood Ratio Test (LRT) (Casella & Berger, 2002) rejects the H_0 in favor of H_A if the likelihood ratio is greater than a threshold, i.e., $\Lambda \geq t$. The LRT is a very effective statistical test because according to the Neyman-Pearson lemma (Neyman & Pearson, 1933), it has the largest power among all statistical tests with the same significance level.

If the sample size is large, $n \to \infty$, and the θ_A is estimated using MLE, the logarithm of the likelihood ratio asymptotically has the distribution of χ^2 under the null hypothesis (White, 1984; Casella & Berger, 2002):

$$2\ln(\Lambda) \overset{H_0}{\sim} \chi^2_{(df)}, \tag{67}$$

where the degree of freedom of χ^2 distribution is $df := \text{dim}(H_A) - \text{dim}(H_0)$ and dim(.) is the number of unspecified parameters in the hypothesis.

There is a connection between LDA or QDA and the LRT (Lachenbruch & Goldstein, 1979). Recall Eq. (12) or (15) which can be restated as:

$$\frac{f_2(\boldsymbol{x})\,\pi_2}{f_1(\boldsymbol{x})\,\pi_1} = 1, \tag{68}$$

which is for the decision boundary. The Eq. (22) dealt with the difference of $f_2(x)\,\pi_2$ and $f_1(x)\,\pi_1$; however, here we are dealing with their ratio. Recall Fig. 1 where if we move x^* to the right and left, the ratio $f_2(x^*)\,\pi_2 / f_1(x^*)\,\pi_1$ decreases and increases, respectively, because the probabilities of the first and second class happening change. In other words, moving the x^* changes the significance level and power. Therefore, Eq. (68) can be used to have a statistical test where the posteriors are used in the ratio, as we also used posteriors in LDA and QDA. The null/alternative hypothesis an be considered to be the mean and covariance of the first/second class. In other words, the two hypotheses say that the point belongs to a specific class. Hence, if the ratio is larger than a value t, the instance $\boldsymbol{x}$ is estimated to belong to the second class; otherwise, the first class is chosen. According to Eq. (16), the Eq. (68) becomes:

$$\frac{(|\boldsymbol{\Sigma}_2|)^{-1/2} \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\big)\,\pi_2}{(|\boldsymbol{\Sigma}_1|)^{-1/2} \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\big)\,\pi_1} \geq t, \tag{69}$$

for QDA. In LDA, the covariance matrices are equal, so:

$$\frac{\exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\big)\,\pi_2}{\exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\big)\,\pi_1} \geq t. \tag{70}$$

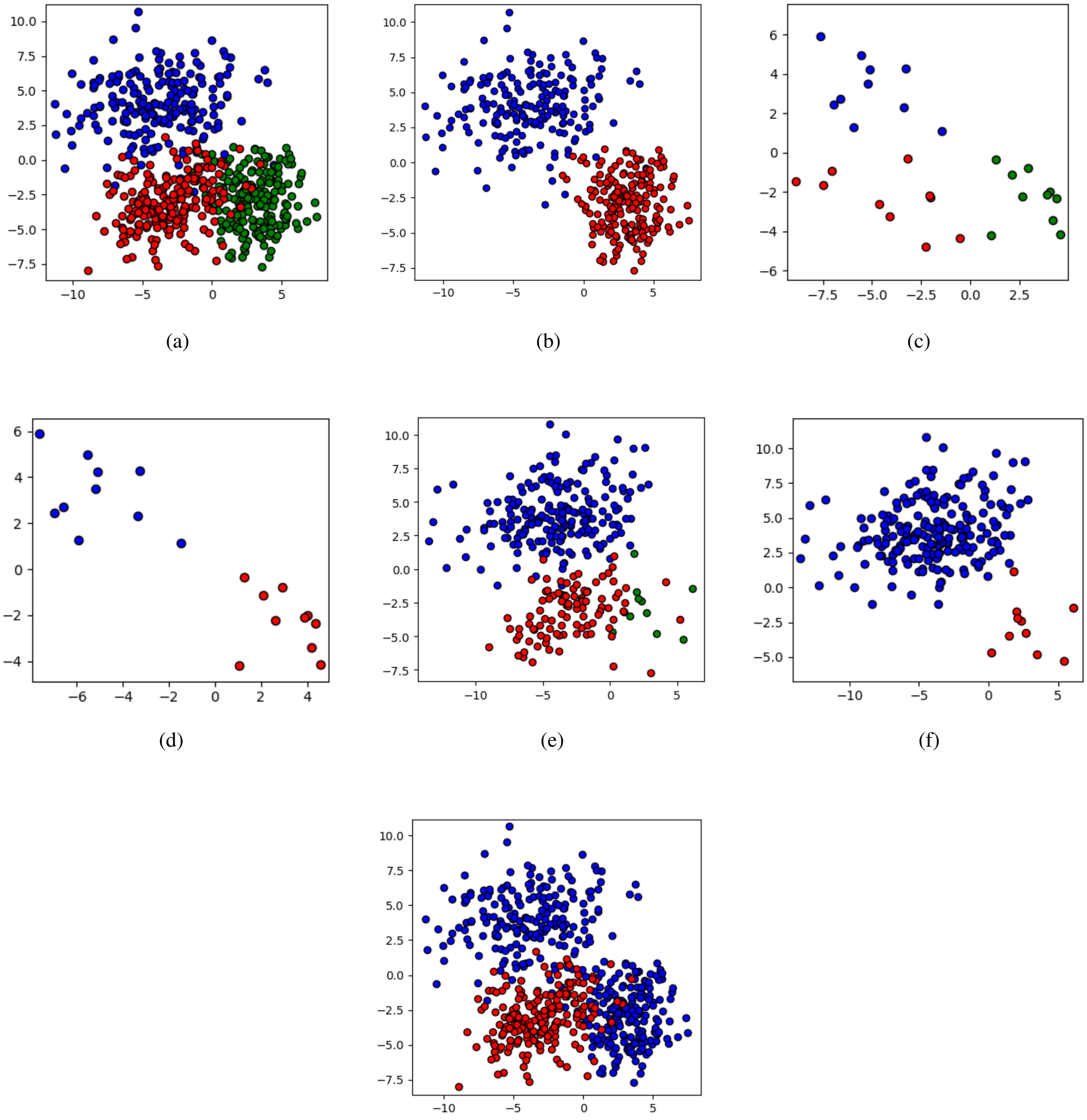
As can be seen, changing the priors change impacts the ratio as expected. Moreover, the value of t can be chosen according to the desired significance level in the χ^2 distribution using the χ^2 table. The Eqs. (69) and (70) show the relation of LDA and QDA with LRT. As the LRT has the largest power (Neyman & Pearson, 1933), the effectiveness of LDA and QDA in classification is explained from a hypothesis testing point of view.

12. Simulations

In this section, we report some simulations which make the concepts of tutorial clearer by illustration.

12.1. Experiments with Equal Class Sample Sizes

We created a synthetic dataset of three classes each of which is a two dimensional Gaussian distribution. The means and covariance matrices of the three Gaussians from



(a) (b) (c)

(d) (e) (f)

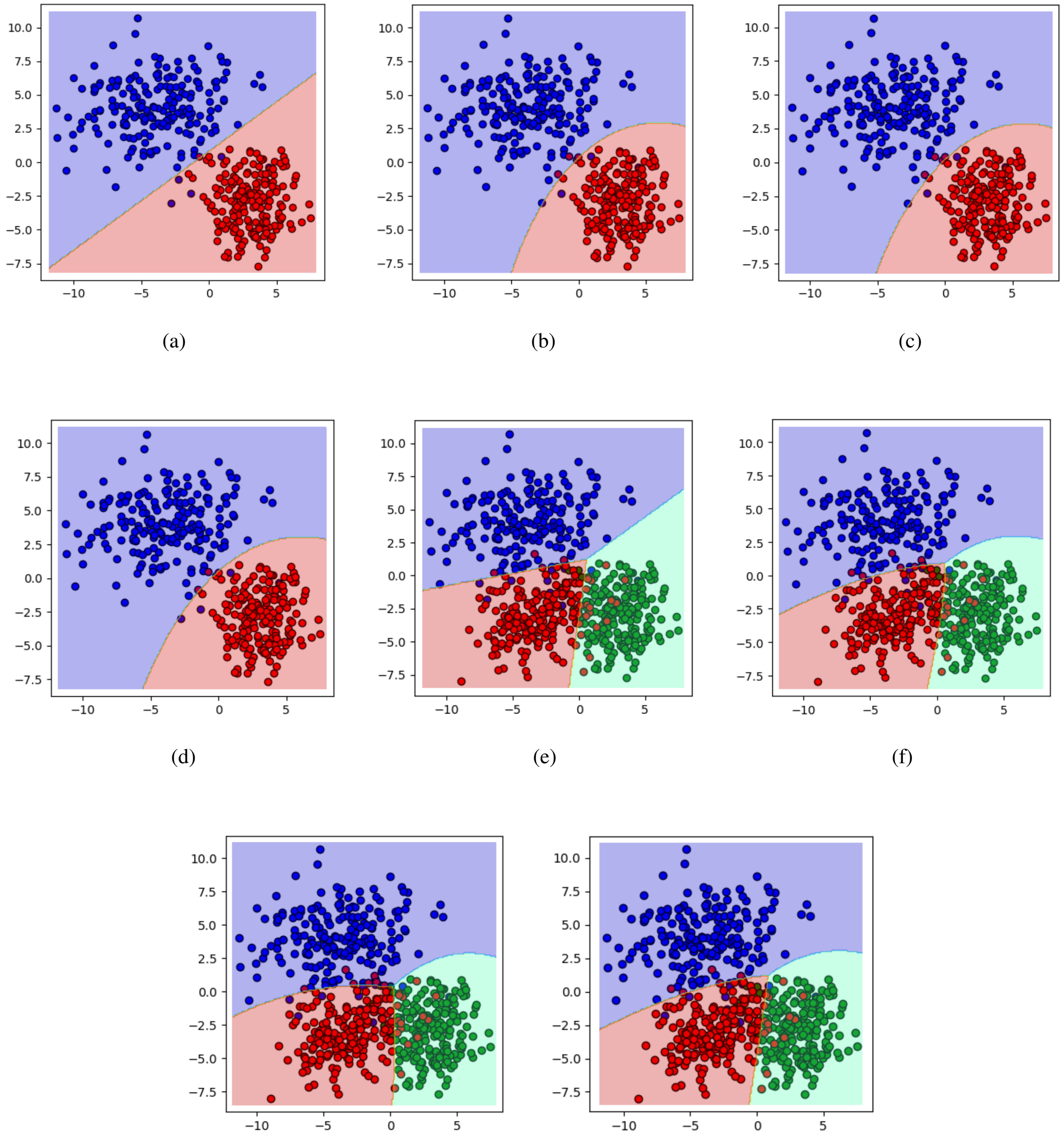
(g)

Figure 3. The synthetic dataset: (a) three classes each with size 200, (b) two classes each with size 200, (c) three classes each with size 10, (d) two classes each with size 10, (e) three classes with sizes 200, 100, and 10, (f) two classes with sizes 200 and 10, and (g) two classes with sizes 400 and 200 where the larger class has two modes.

which the class samples were randomly drawn are:

$$\boldsymbol{\mu}_1 = [-4, 4]^\top, \;\; \boldsymbol{\mu}_2 = [3, -3]^\top, \;\; \boldsymbol{\mu}_1 = [-3, 3]^\top,$$
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 10 & 1 \\ 1 & 5 \end{bmatrix}, \;\; \boldsymbol{\Sigma}_2 = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}, \;\; \boldsymbol{\Sigma}_3 = \begin{bmatrix} 6 & 1.5 \\ 1.5 & 4 \end{bmatrix}.$$

The three classes are shown in Fig. 3-a where each has sample size 200. Experiments were performed on the three classes. We also performed experiments on two of the three classes to test a binary classification. The two classes are shown in Fig. 3-b. The LDA, QDA, naive Bayes, and



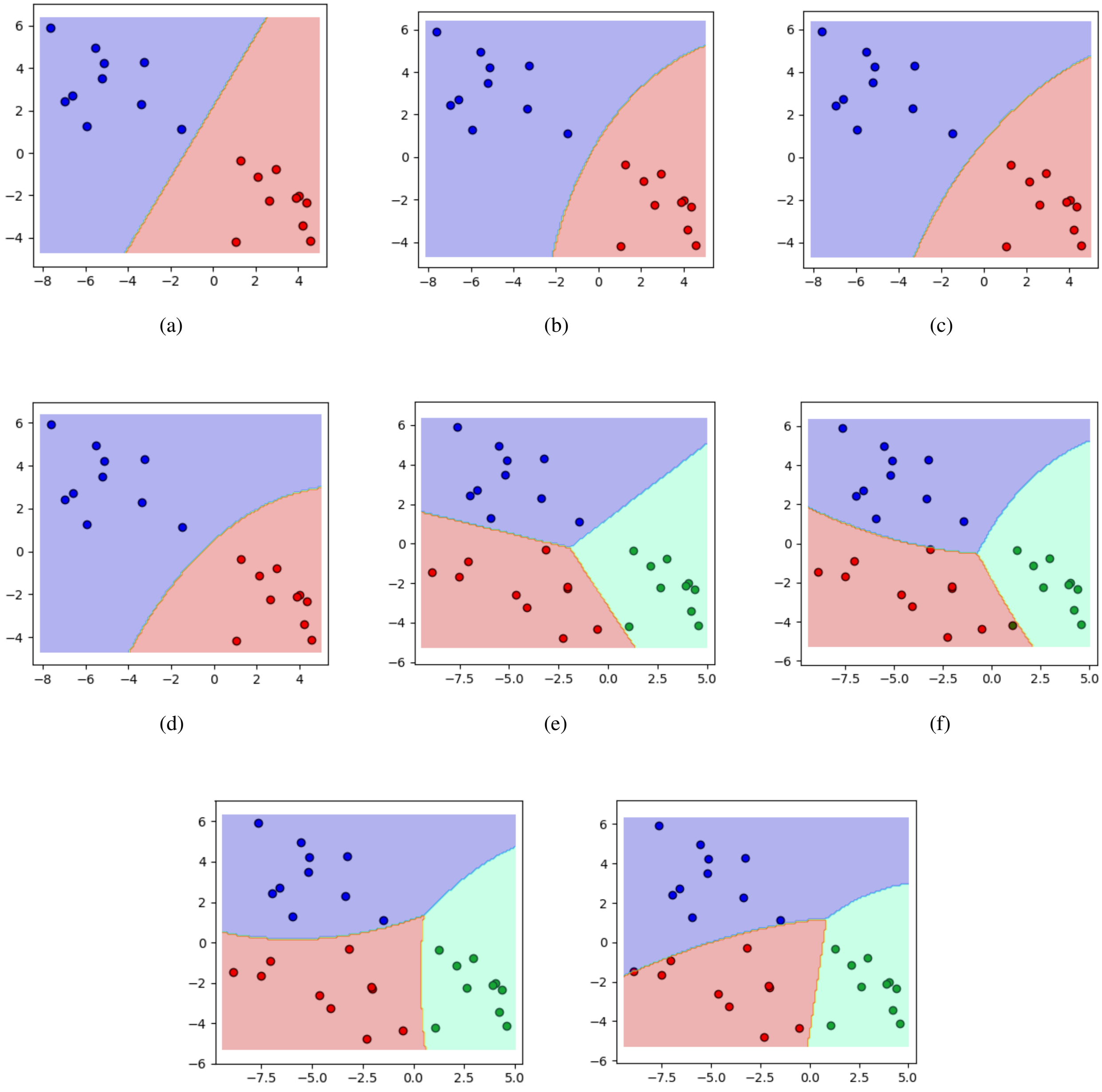
(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 4. Experiments with equal class sample sizes: (a) LDA for two classes, (b) QDA for two classes, (c) Gaussian naive Bayes for two classes, (d) Bayes for two classes, (e) LDA for three classes, (f) QDA for three classes, (g) Gaussian naive Bayes for three classes, and (h) Bayes for three classes.

Bayes classifications of the two and three classes are shown in Fig. 4. For both binary and ternary classification with LDA and QDA, we used Eqs. (31) and (28), respectively, with Eq. (29). We also estimated the mean and covariance using Eqs. (33), (35), and (36). For Gaussian naive Bayes, we used Eqs. (62) and (63) and estimated the parameters using Eqs. (64) and (65). For Bayes classifier, we used Eq. (58) with Eq. (63) but we do not estimate the mean and variance; except, in order to use the exact likelihoods in Eq. (58), we use the exact mean and covariance matrices



(a)

(b)

(c)

(d)

(e)

(f)

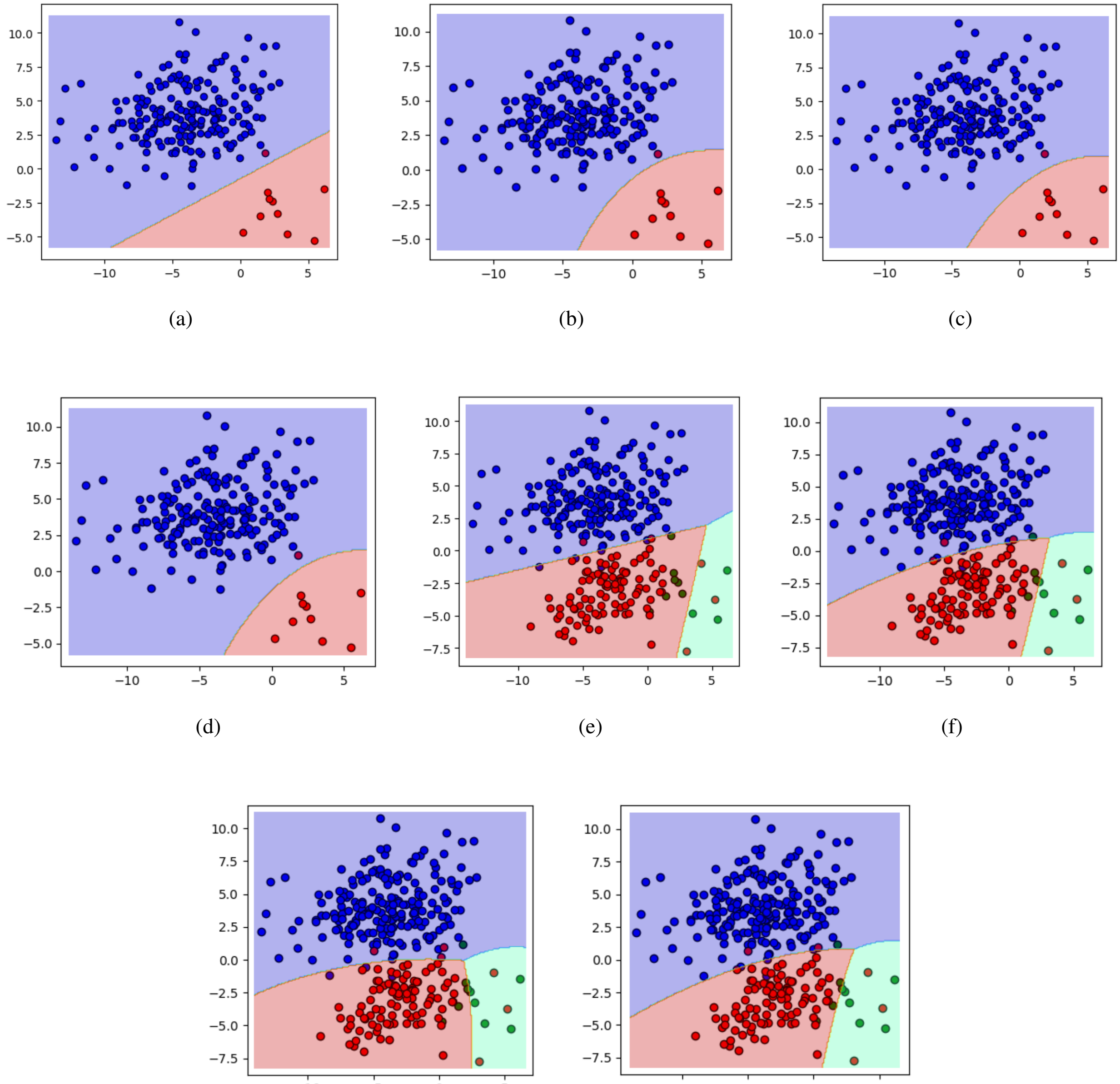
(g)

(h)

Figure 5. Experiments with small class sample sizes: (a) LDA for two classes, (b) QDA for two classes, (c) Gaussian naive Bayes for two classes, (d) Bayes for two classes, (e) LDA for three classes, (f) QDA for three classes, (g) Gaussian naive Bayes for three classes, and (h) Bayes for three classes.

of the distributions which we sampled from. We, however, estimated the priors. The priors were estimated using Eq. (32) for all the classifiers.

As can be seen in Fig. 4, the space is partitioned into two/three parts and this validates the assertion that LDA and QDA can be considered as metric learning methods as discussed in Section 7. As expected, the boundaries of LDA and QDA are linear and curvy (quadratic), respectively. The results of QDA, Gaussian naive Bayes, and Bayes are very similar although they have slight differences. This is because the classes are already Gaussian so if the estimates of means and covariance matrices are accurate enough, QDA and Bayes are equivalent. The classes are Gaussians and the off-diagonal elements of covariance



(a) (b) (c)

(d) (e) (f)

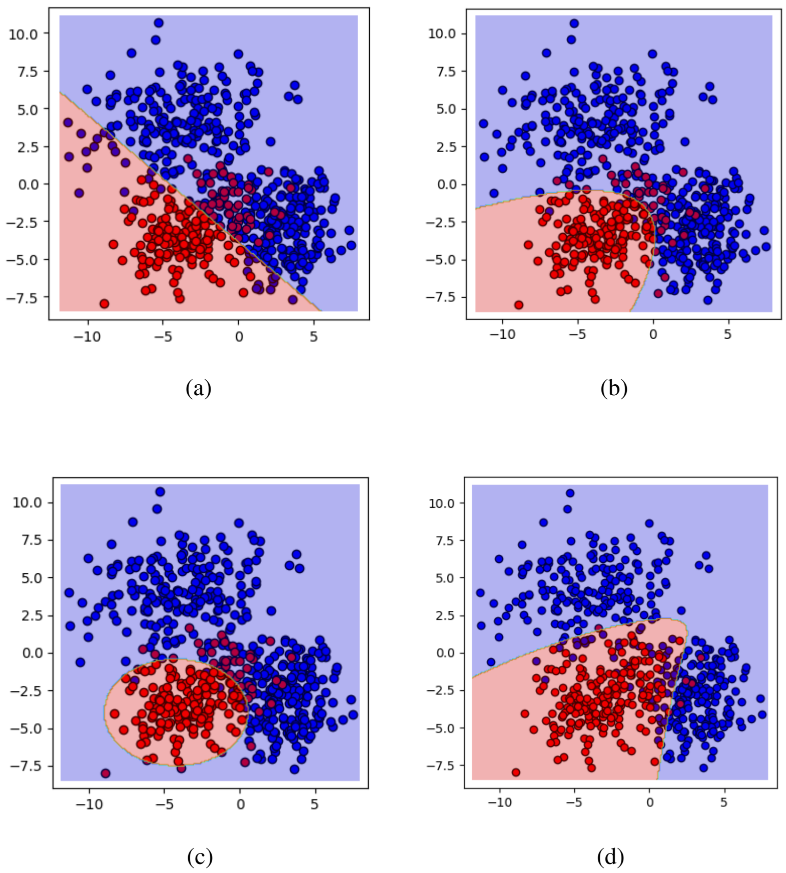
(g) (h)

Figure 6. Experiments with different class sample sizes: (a) LDA for two classes, (b) QDA for two classes, (c) Gaussian naive Bayes for two classes, (d) Bayes for two classes, (e) LDA for three classes, (f) QDA for three classes, (g) Gaussian naive Bayes for three classes, and (h) Bayes for three classes.

matrices are also small compared to the diagonal; therefore, naive Bayes is also behaving similarly.

12.2. Experiments with Small Class Sample Sizes

According to Monte-Carlo approximation (Robert & Casella, 2013), the estimates in Eqs. (33), (35), (64) and (65) are more accurate if the sample size goes to infinity, i.e., $n \to \infty$. Therefore, if the sample size is small, we expect mode difference between QDA and Bayes classifiers. We made a synthetic dataset with three or two classes with the same mentioned means and covariance matrices. The sample size of every class was 10. Figures 3-c and 3-d show these datasets. The results of LDA, QDA, Gaussian naive Bayes, and Bayes classifiers for this dataset are



(a)

(b)

(c)

(d)

Figure 7. Experiments with multi-modal data: (a) LDA, (b) QDA, (c) Gaussian naive Bayes, and (d) Bayes.

shown in Fig. 5. As can be seen, now, the results of QDA, Gaussian naive Bayes, and Bayes are different for the reason explained.

12.3. Experiments with Different Class Sample Sizes

According to Eq. (32) used in Eqs. (28), (31), (58), and (62), the prior of a class changes by the sample size of the class. In order to see the effect of sample size, we made a synthetic dataset with different class sizes, i.e., 200, 100, and 10, shown in Figs. 3-e, 3-f. We used the same mentioned means and covariance matrices. The results are shown in Fig. 6. As can be seen, the class with small sample size has covered a small portion of space in discrimination which is expected because its prior is small according to Eq. (32); therefore, its posterior is small. On the other hand, the class with large sample size has covered a larger portion because of a larger prior.

12.4. Experiments with Multi-Modal Data

As mentioned in Section 8, LDA and QDA assume unimodal Gaussian distribution for every class and thus FDA or LDA faces problem for multi-modal data (Sugiyama, 2007). For testing this, we made a synthetic dataset with two classes, one with sample size 400 having two modes of Gaussians and the other with sample size 200 having one mode. We again used the same mentioned means and covariance matrices. The dataset is shown in Fig. 3-g.

The results of the LDA, QDA, Gaussian naive Bayes, and Bayes classifiers for this dataset are shown in Fig. 7. The mean and covariance matrix of the larger class, although it has two modes, were estimated using Eqs. (33), (35), (64) and (65) in LDA, QDA, and Gaussian naive Bayes. However, for the likelihood used in Bayes classifier, i.e., in Eq. (58), we need to know the exact multi-modal distribution. Therefore, we fit a mixture of two Gaussians (Ghojogh et al., 2019a) to the data of the larger class:

$$\mathbb{P}(X = \boldsymbol{x} \,|\, \boldsymbol{x} \in \mathcal{C}_k) = \sum_{k=1}^{2} w_k \, f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad (71)$$

where $f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is Eq. (16) and we the fitted parameters were:

$$\boldsymbol{\mu}_1 = [-3.88, 4]^\top,\ \ \boldsymbol{\mu}_2 = [3.04, -2.92]^\top,$$
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 9.27 & 0.79 \\ 0.79 & 4.82 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.87 & 0.03 \\ 0.03 & 3.78 \end{bmatrix},$$
$$w_1 = 0.49,\ \ w_2 = 0.502.$$

As Fig. 7 shows, LDA has not performed well enough as expected. The performance of QDA is more acceptable than LDA but still not good enough because QDA also assumes a uni-modal Gaussian for every class. The result of Gaussian naive Bayes is very different from the Bayes here because the Gaussian naive Bayes assumes uni-modal Gaussian with diagonal covariance for every class. Finally, the Bayes has the best result as it takes into account the multi-modality of the data and it is optimum (Mitchell, 1997).

13. Conclusion and Future Work

This paper was a tutorial paper for LDA and QDA as two fundamental classification methods. We explained the relations of these two methods with some other methods in machine learning, manifold (subspace) learning, metric learning, statistics, and statistical testing. Some simulations were also provided for better clarification.

This paper focused on LDA and QDA which are discriminators with one and two polynomial degrees of freedom, respectively. As the future work, we will work on a tutorial paper for non-linear discriminant analysis using kernels (Baudat & Anouar, 2000; Li et al., 2003; Lu et al., 2003), which is called *kernel discriminant analysis*, to have discriminators with more than two degrees of freedom.

Acknowledgment

The authors hugely thank Prof. Ali Ghodsi (see his great online related courses (Ghodsi, 2015; 2017)), Prof. Mu Zhu, Prof. Hoda Mohammadzade, and other professors whose courses have partly covered the materials mentioned in this tutorial paper.

References

Baudat, Gaston and Anouar, Fatiha. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

Belhumeur, Peter N, Hespanha, João P, and Kriegman, David J. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):711–720, 1997.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Casella, George and Berger, Roger L. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Cox, Trevor F and Cox, Michael AA. *Multidimensional scaling*. Chapman and hall/CRC, 2000.

Croot, Ernie. The Rayleigh principle for finding eigenvalues. Technical report, Georgia Institute of Technology, School of Mathematics, 2005. Online: http://people.math.gatech.edu/∼ecroot/notes_linear.pdf, Accessed: March 2019.

De Maesschalck, Roy, Jouan-Rimbaud, Delphine, and Massart, Désiré L. The Mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

Etemad, Kamran and Chellappa, Rama. Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America A*, 14(8):1724–1733, 1997.

Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. *The elements of statistical learning*, volume 2. Springer series in statistics, New York, NY, USA, 2009.

Ghodsi, Ali. Classification course, department of statistics and actuarial science, university of Waterloo. Online Youtube Videos, 2015. Accessed: January 2019.

Ghodsi, Ali. Data visualization course, department of statistics and actuarial science, university of Waterloo. Online Youtube Videos, 2017. Accessed: January 2019.

Ghojogh, Benyamin, Mohammadzade, Hoda, and Mokari, Mozhgan. Fisherposes for human action recognition using Kinect sensor data. *IEEE Sensors Journal*, 18(4): 1612–1627, 2017.

Ghojogh, Benyamin, Ghojogh, Aydin, Crowley, Mark, and Karray, Fakhri. Fitting a mixture distribution to data: Tutorial. *arXiv preprint arXiv:1901.06708*, 2019a.

Ghojogh, Benyamin, Karray, Fakhri, and Crowley, Mark. Eigenvalue and generalized eigenvalue problems: Tutorial. *arXiv preprint arXiv:1903.11240*, 2019b.

Ham, Ji Hun, Lee, Daniel D, Mika, Sebastian, and Schölkopf, Bernhard. A kernel view of the dimensionality reduction of manifolds. In *International Conference on Machine Learning*, 2004.

Hazewinkel, Michiel. Central limit theorem. *Encyclopedia of Mathematics, Springer*, 2001.

Jolliffe, Ian. *Principal component analysis*. Springer, 2011.

Kleinbaum, David G, Dietz, K, Gail, M, Klein, Mitchel, and Klein, Mitchell. *Logistic regression*. Springer, 2002.

Kulis, Brian. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

Lachenbruch, Peter A and Goldstein, M. Discriminant analysis. *Biometrics*, pp. 69–85, 1979.

Li, Yongmin, Gong, Shaogang, and Liddell, Heather. Recognising trajectories of facial identities using kernel discriminant analysis. *Image and Vision Computing*, 21 (13-14):1077–1086, 2003.

Lu, Juwei, Plataniotis, Konstantinos N, and Venetsanopoulos, Anastasios N. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, 14(1):117–126, 2003.

Malekmohammadi, Alireza, Mohammadzade, Hoda, Chamanzar, Alireza, Shabany, Mahdi, and Ghojogh, Benyamin. An efficient hardware implementation for a motor imagery brain computer interface system. *Scientia Iranica*, 26:72–94, 2019.

McLachlan, Goeffrey J. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999.

Mitchell, Thomas. *Machine learning*. McGraw Hill Higher Education, 1997.

Mokari, Mozhgan, Mohammadzade, Hoda, and Ghojogh, Benyamin. Recognizing involuntary actions from 3d skeleton data using body states. *Scientia Iranica*, 2018.

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Neyman, Jerzy and Pearson, Egon Sharpe. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Robert, Christian and Casella, George. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Strange, Harry and Zwiggelaar, Reyer. *Open Problems in Spectral Dimensionality Reduction*. Springer, 2014.

Sugiyama, Masashi. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of machine learning research*, 8(May):1027–1061, 2007.

Tharwat, Alaa, Gaber, Tarek, Ibrahim, Abdelhameed, and Hassanien, Aboul Ella. Linear discriminant analysis: A detailed tutorial. *AI communications*, 30(2):169–190, 2017.

Welling, Max. Fisher linear discriminant analysis. Technical report, University of Toronto, Toronto, Ontario, Canada, 2005.

White, Halbert. *Asymptotic theory for econometricians*. Academic press, 1984.

Xu, Yong and Lu, Guangming. Analysis on fisher discriminant criterion and linear separability of feature space. In *2006 International Conference on Computational Intelligence and Security*, volume 2, pp. 1671–1676. IEEE, 2006.

Yang, Liu and Jin, Rong. Distance metric learning: A comprehensive survey. Technical report, Department of Computer Science and Engineering, Michigan State University, 2006.

Zhang, Harry. The optimality of naive Bayes. In *American Association for Artificial Intelligence (AAAI)*, 2004.

Zhao, Wenyi, Chellappa, Rama, and Phillips, P Jonathon. *Subspace linear discriminant analysis for face recognition*. Citeseer, 1999.