LDA and QDA

Eric Roemmele

Feb 15, 2018

LDA

- Let $\widehat{G} : \chi \to \mathcal{G}$ be a classifier, where χ is the feature space and $\mathcal{G}$ is $\mathcal{G} = \{1, 2, \ldots, K\}$ are the set of numeric class labels.
- Under zero-one loss, we can show (ESL 2.4) the optimal classifier is

  $$\widehat{G}(X) = k \text{ if } \mathrm{P}(k|X = x) = \max_g P(g|X = x).$$

  In other words, classify the observation to the largest posterior probability, given the features $X = x$.
- This is called the Bayes classifier, since we are looking at posterior membership probabilities.
- Most classifiers, seek to model the posterior memberships $\mathrm{P}(G = g|X)$. (Why we can take X as "fixed")

LDA Setup

- Suppose the class-conditioned density $X|G = k \sim f_k(x)$, and let $\pi_k = \mathrm{P}(G = k)$ be the prior probability of class k.
- By Bayes Theorem,

$$\mathrm{P}(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{1}$$

- Now further suppose that $f_k(x)$ is the Gaussian density

$$f_k(x) = \left((2\pi)^p |\mathbf{\Sigma}_k|\right)^{-1/2} e^{-\frac{1}{2}(x-\mu_k)^{\mathrm{T}}\mathbf{\Sigma}_k^{-1}(x-\mu_k)}$$

- **LDA Assumption** We further suppose all classes have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}$ for $\forall k$.

Classification Boundary

- Then, classify observation x to k if it has the largest posterior probability $\mathrm{P}(G = k|X = x)$.
- What does our classification rule look like?
- Examine the log-ratio of two classes k and l :

$$\log \frac{\mathrm{P}(G = k|X = x)}{\mathrm{P}(G = l|X = x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^{\mathrm{T}} \mathbf{\Sigma}^{-1}(\mu_k - \mu_l) + x^{\mathrm{T}} \Sigma^{-1}(\mu_k - \mu_l)$$

  which is linear in x.
- Therefore, our decision boundaries are (linear , technically affine) p-dimensional hyperplanes.

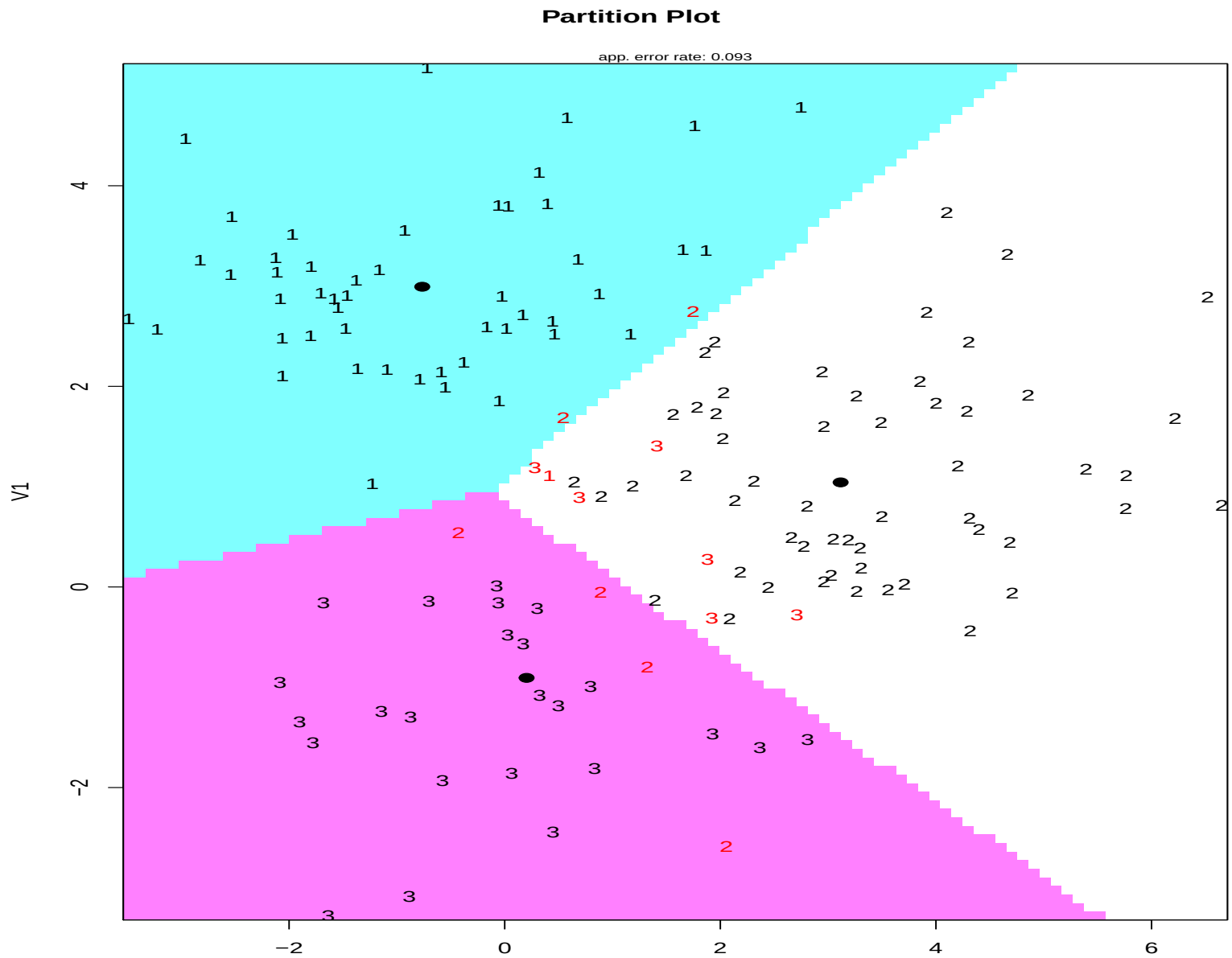
Classification

- Now, write

$$\log \frac{\mathrm{P}(G=k|X=x)}{\mathrm{P}(G=l|X=x)} > 0 \iff x^{\mathrm{T}}\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^{\mathrm{T}}\Sigma^{-1}\mu_k + \log \pi_k$$
$$> x^{\mathrm{T}}\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^{\mathrm{T}}\Sigma^{-1}\mu_l + \log \pi_l$$
$$\iff \delta_k(x) > \delta_l(x)$$

- Equivalently, we can say our decision rule is $\widehat{G}(x) = \underset{k}{\operatorname{argmax}}\ \delta_k(x)$
- We call the $\delta_k(x)$ the linear discriminant functions.
- In essence, we assume we have a mixture of Gaussian distributions, and we construct the optimal linear hyperplane that separates those mixing populations.
- In practice, we need to estimate the parameters (mixing proportions, class means, and covariance matrix) by in-sample estimates.

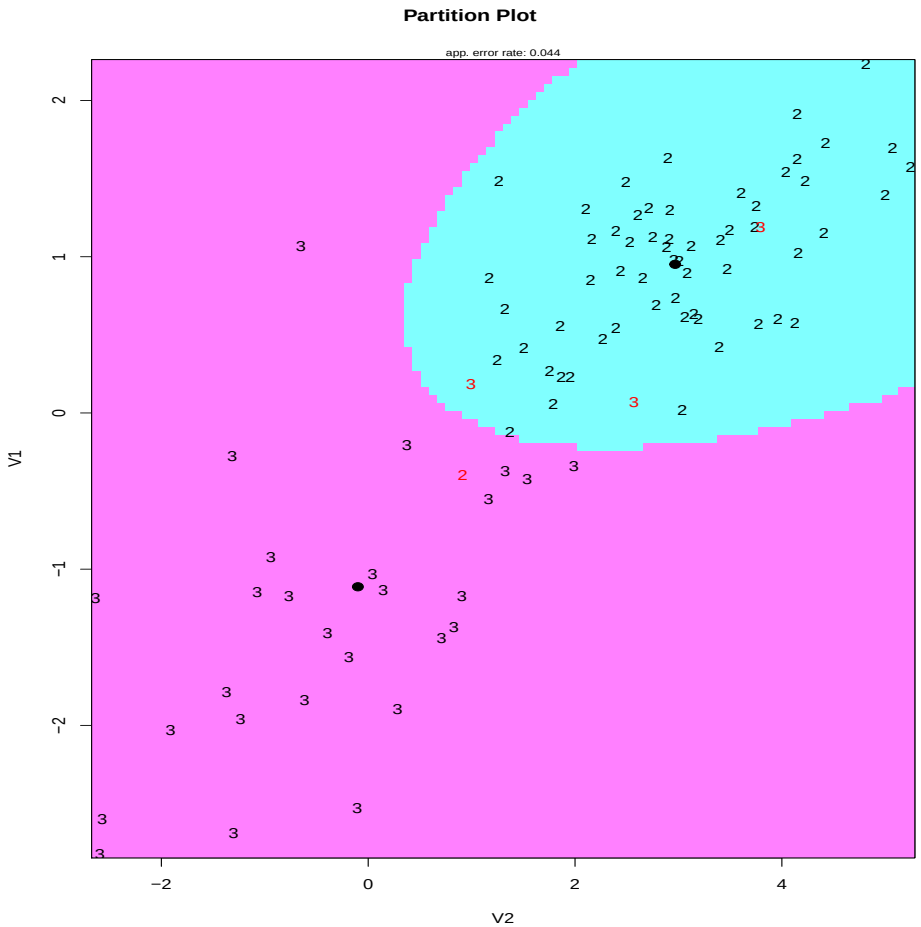
Graph



QDA

- Same process for QDA, excpet each group has their own covariance matrix.
- The discriminant functions are now

$$\delta_k(x) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^{\mathrm{T}}\mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k$$



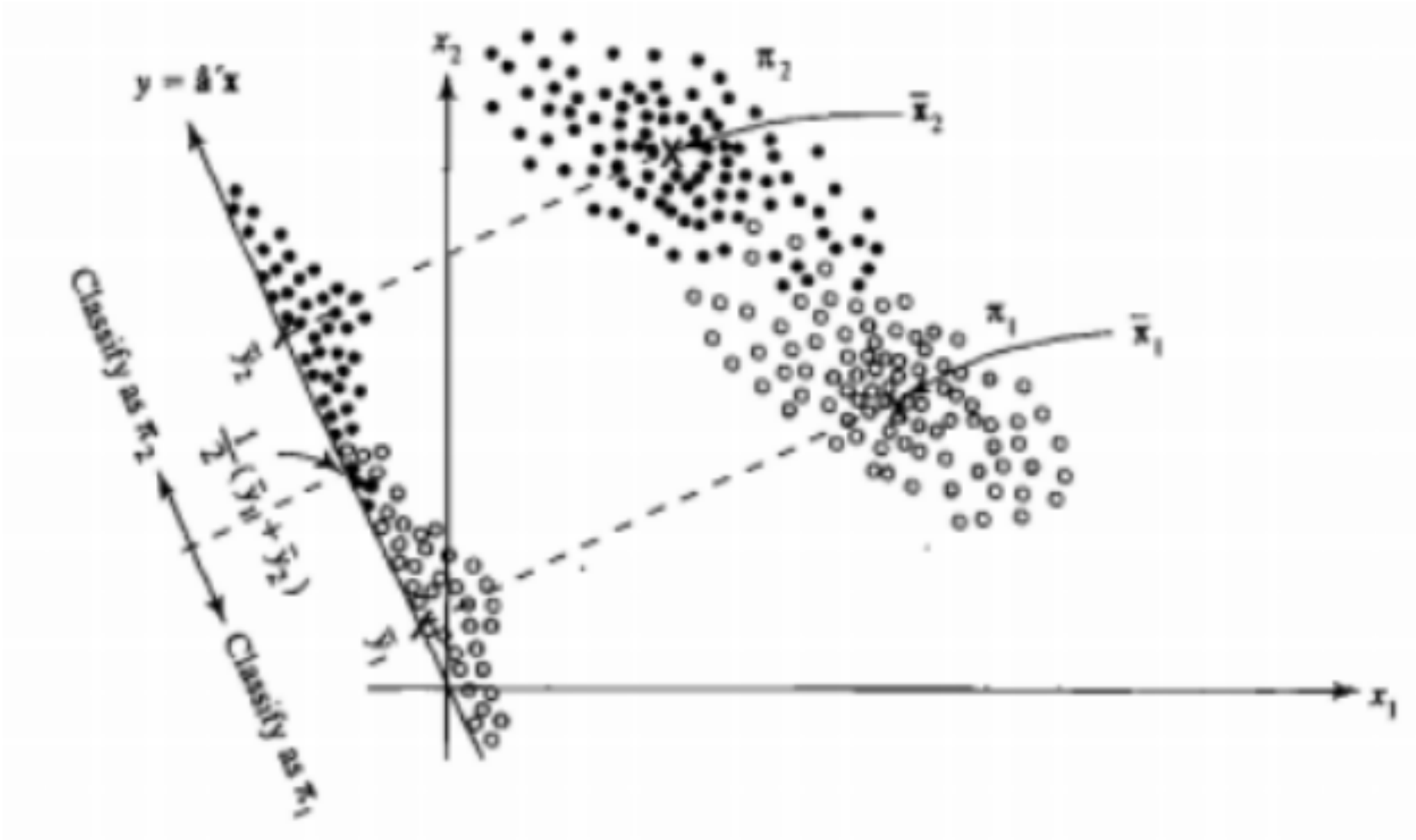
Why LDA?

- LDA is surprisingly hard to beat (in terms of accuracy)
- Why does it work?
  - Our data probably aren't Gauassian
  - In the case of LDA, the covariance matrices probably aren't equal.
  - "linear discriminant analysis frequently achieves good performances in the tasks of face and object recognition, even though the assumptions of common covariance matrix among groups and normality are often violated" (Duda, et al., 2001) (Tao Li, et al., 2006)
- The reason is likely to be stability and simplicity of linear boundary.
  - More bias, but much lower variance than exotic alternatives.

Dimension Reduction in LDA

- Another reason for LDA popularity is it contains a rank-reducing element.
- R.A. Fisher originally derived the linear classification statistic using an entirely different argument, without reference to multivariate Gaussian distributions.
- His argument was to examine linear combinations of multivariate observations x (i.e. $a^T x$) to univariate observations y (i.e. $y = a^T x$) such that the derived y from two populations were separated as much as possible.
- In other words, project the observations onto the axis where they are most separated, and then classify.

Visual


$y = \hat{a}'x$
x_2
π_2
$\bar{x}_2$
π_1
$\bar{x}_1$
x_1
$\bar{y}_2$
$\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$
$\bar{y}_1$
Classify as π_2
Classify as π_1

Formulation

- How do we find this projection?
- Define the following :

$$\text{Between Class Variance } \boldsymbol{B} = \sum_{i=1}^{K} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^{\mathrm{T}}$$

$$\text{Within Class Variance } \boldsymbol{W} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^{\mathrm{T}}$$

- Note that $\boldsymbol{S} = \boldsymbol{B} + \boldsymbol{W}$, where S is typical covariance matrix (without the (n-1) in the denominator).

Objective

- Goal : Maximize the following function :

$$J(a) = \frac{a^{\mathrm{T}} B a}{a^{\mathrm{T}} W a} \tag{2}$$

- In words -*Find the linear combination* $Y = a^T x$ *such that the between class variance is maximized relative to the within-class variance.*
- This objective function makes sense because we want the class means well separated (i.e. high variability) relative to the within group variability.
- Thus we have the means of the groups well-separated, as well as low variability within-group. This makes classification easier, along with low dimensional representation of the data.
- The solution to (2) above is given by $a = e_1$, where e_1 is the eigenvector with the largest eigenvalue λ_1 of the matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$. The max to (2) is λ_1. This is not hard to prove, just use Lagrange multipliers.

Fisher's Dicriminate Continued

- We can continue with the setup in (2) and ask, what is the vector a_2 that maximizes J(a) such that it is projected in the orthogonal direction to a_1?
- It turns out the solution is e_2, or eigenvector of $\boldsymbol{W}^{-1}\boldsymbol{B}$ with the second largest eigenvalue.
- Usually we look for reduction of the data to two dimensions.
- Suppose we stopped in r dimensional space.

Fisher's Classification Rule

$$\widehat{G}(x) = \underset{k}{\text{argmin}} \sum_{j=1}^{r} \left[e_j^{\mathrm{T}}(x - \mu_k) \right]^2$$

- To develop the above will takes a bit more work. Please refer to Johnson and Wichern Chapter 11, or any multivariate statistics book such as Rencher and Christensen.

Further LDA Related Methods to Investigate

- Friedman (1989) proposed a compromise between LDA and QDA where $\boldsymbol{\Sigma}_k(\alpha) = \alpha\boldsymbol{\Sigma}_k + (1-\alpha)\Sigma$. Shrinking towards a common covariance matrix.
- Naive or "Idiot" Bayes - Instead of f_k the Guassian density, replace it by a non-parametric estimate of the density assuming conditional independence between components (i.e. $f_k(x) = \prod_{i=1}^{p} f_{ki}(x_i)$).
- Kernel LDA - Map data to higher dimensional space, and then classify it.