Lecture 10: Bayes Classifier, LDA and QDA

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Read: ELSII Ch. 4.3, ISLR Chs. 4.4–4.6, and SLS Ch. 8.4

Outline

1. Bayes Rule motivated LDA & QDA
2. Fisher's Linear Discriminant (optimization and geometry views)
3. Related Methods
4. Regularized & Sparse LDA

The next section would be ......

1. Bayes Rule motivated LDA & QDA
2. Fisher's Linear Discriminant (optimization and geometry views)
3. Related Methods
4. Regularized & Sparse LDA

Setup

Assume $K = 2$ for simplicity

$$\boldsymbol{X} \mid (Y = 1) \sim f_1(\boldsymbol{x}), \quad \boldsymbol{X} | (Y = 2) \sim f_2(\boldsymbol{x}).$$

Denote a random observation from this population (a mixture of two sub-populations) by $(\boldsymbol{X}, Y)$. We assume

$$P(Y = 1) = \pi_1, \quad P(Y = 2) = \pi_2, \quad \pi_1 + \pi_2 = 1$$

If the observed value of $\boldsymbol{X}$ is $\boldsymbol{x}$, then Bayes theorem yields the posterior probability that $\boldsymbol{X}$ was from Class 1 is

$$\eta(\boldsymbol{x}) := P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{f_1(\boldsymbol{x})\pi_1}{f_1(\boldsymbol{x})\pi_1 + f_2(\boldsymbol{x})\pi_2}.$$

0-1 Loss and Bayes (decision) rule

- Use a decision theory framework, if we care about the 0-1 loss: $\mathbb{1}\{Y \neq \delta(\boldsymbol{X})\}$, we would hope to choose a decision function to minimize the risk associated with the 0-1 loss.

$$\mathrm{E}[\mathbb{1}\{Y \neq \delta(\boldsymbol{X})\}] = P[Y \neq \delta(\boldsymbol{X})]$$

- Only need to choose the best decision $\delta(\boldsymbol{x})$ for each given $\boldsymbol{x}$, which minimizes $P[Y \neq \delta(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}]$
- Rewrite $P[Y \neq \delta(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}]$ as

$$\mathbb{1}\{\delta(\boldsymbol{x}) = 1\}P(Y \neq 1 \mid \boldsymbol{X} = \boldsymbol{x}) + \mathbb{1}\{\delta(\boldsymbol{x}) = 2\}P(Y \neq 2 \mid \boldsymbol{X} = \boldsymbol{x})$$

- The blue parts are either 0 or 1. Hence we only need to choose $\delta(\boldsymbol{x})$ to be j with a greater $P(Y = j \mid \boldsymbol{X} = \boldsymbol{x})$

Bayes Rule for Gaussian Data

Bayes Rule classifier assigns $\boldsymbol{x}$ to the class label (1 or 2) which gives the higher posterior probability:

$$\phi_{\text{Bayes}}(\boldsymbol{x}) = \operatorname*{argmax}_{k=1,2} P(Y = k \mid \boldsymbol{X} = \boldsymbol{x}).$$

Now assume Gaussian data as example:

$$\boldsymbol{X} \mid (Y = 1) \sim N_p(\mu_1, \Sigma), \quad \boldsymbol{X} | (Y = 2) \sim N_p(\mu_2, \Sigma), \quad \mu_1 \neq \mu_2.$$

Recall that

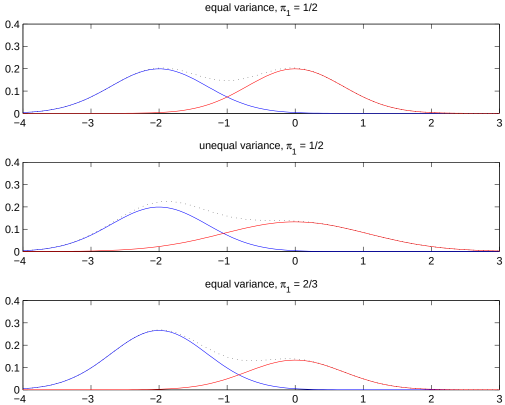
$$\eta(\boldsymbol{x}) := P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{f_1(\boldsymbol{x})\pi_1}{f_1(\boldsymbol{x})\pi_1 + f_2(\boldsymbol{x})\pi_2}.$$

Bayes Rule Classifier: Gaussian 1-D

As a special case, assume $p = 1$ and

$$X|(Y=1) \sim N_1(\mu_1, \sigma_1^2), \quad X|(Y=2) \sim N_1(\mu_2, \sigma_2^2), \quad \mu_1 \neq \mu_2.$$

Then $P(Y = i \mid X = x) \propto \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{(x-\mu_i)^2}{2\sigma_i^2}) \times \pi_i$



Bayes Rule Classifier: Gaussian 1-D

Bayes rule classifies x into class 1 (the blue class)

1. case I ($\sigma_1 = \sigma_2, \pi_1 = \pi_2$): if
$$|x - \mu_1| < |x - \mu_2|.$$
(Linear; bisector cutoff.)
2. case II ($\sigma_1 < \sigma_2, \pi_1 = \pi_2$): if
$$(\frac{x - \mu_1}{\sigma_1})^2 < (\frac{x - \mu_2}{\boxed{\sigma_2}})^2 + \log(\sigma_2^2/\sigma_1).$$
(Quadratic; biased cutoff.)
3. case III ($\sigma_1 = \sigma_2, \pi_1 > \pi_2$): if
$$(\frac{x - \mu_1}{\sigma_1})^2 < (\frac{x - \mu_2}{\sigma_1})^2 + 2\log(\pi_1/\pi_2).$$
$$(\frac{-2x\mu_1 + \mu_1^2}{\sigma_1^2}) < (\frac{-2x\mu_2 + \mu_2^2}{\sigma_1^2}) + 2\log(\pi_1/\pi_2).$$
(Linear; biased cutoff.)

Multiclass Bayes Rule

In general, for $k = 1, \dots, K > 2$, if observed value of $\boldsymbol{X} = \boldsymbol{x}$, then

$$\eta_k(\boldsymbol{x}) := P(Y = k \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\boxed{f_k(\boldsymbol{x})\pi_k}}{f(\boldsymbol{x})},$$

where $f_k(\boldsymbol{x})$ is density function of $\boldsymbol{X} \mid (Y = k)$ and $f(\boldsymbol{x})$ is the marginal density of $\boldsymbol{X}$. Bayes rule assigns $\boldsymbol{x}$ to class label with highest $\eta_k(\boldsymbol{x})$:

$$\phi_{\text{Bayes}}(\boldsymbol{x}) = \operatorname*{argmax}_{k=1,\dots,K} \frac{\boxed{f_k(\boldsymbol{x})\pi_k}}{f(\boldsymbol{x})} = \operatorname*{argmax}_{k=1,\dots,K} \boxed{f_k(\boldsymbol{x})\pi_k}.$$

(True) Quadratic Discriminant Analysis

Gaussian Example:

$$\boldsymbol{X} \mid (Y = k) \sim N_p(\boldsymbol{\mu}_k, \Sigma_k), \quad k = 1, \dots, K,$$

$$P(Y = k) = \pi_k, \quad \sum_{k=1}^{K} \pi_k = 1,$$

$$\boldsymbol{X} \sim f(\boldsymbol{x}), \text{ a mixture density for multivariate normals}$$

Then

$$\phi_{\text{Bayes}}(\boldsymbol{x}) = \operatorname*{argmax}_k \eta_k(\boldsymbol{x}) = \operatorname*{argmax}_k \delta_k(\boldsymbol{x})$$

where

$$\delta_k(\boldsymbol{x}) = \left[\log(\pi_k) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right].$$

(True) Quadratic Discriminant Analysis - Binary Case

For binary classification ($k = 1, 2$), it is equivalent to calculating the sign of

$$\delta_1(\boldsymbol{x}) - \delta_2(\boldsymbol{x}) = -\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)'\Sigma_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\right] - \frac{1}{2}\log(|\Sigma_1|/|\Sigma_2|) + \log(\pi_1/\pi_2)$$

$$= -\frac{1}{2}\boldsymbol{x}'\nabla\boldsymbol{x} + (\Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_2^{-1}\boldsymbol{\mu}_2)'\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}\log(|\Sigma_1|/|\Sigma_2|) + \log(\pi_1/\pi_2)$$

where $\nabla = \Sigma_1^{-1} - \Sigma_2^{-1}$

Quadratic Discriminant Analysis

- In practice, we do not know the parameters $(\boldsymbol{\mu}_k, \Sigma_k, \pi_k)$.
- Given n observations $(\boldsymbol{x}_{ij}, i)$, $(i = 1, \ldots, K), (j = 1, \ldots, n_k)$, $n = \sum_{k=1}^{K} n_k$, QDAs are obtained by substituting
  1. $\boldsymbol{\mu}_k$ with $\hat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{x}}_k$,
  2. Σ_k with $\widehat{\Sigma}_k = \mathbf{S}_k$,
  3. $\hat{\pi}_k = n_k/n$.

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis is essentially an estimate of the Bayes rule classifier for Gaussian data, which is

$$\phi(\boldsymbol{x}) = \operatorname*{argmax}_{k=1,\dots,K} \hat{\delta}_k(\boldsymbol{x}), \text{ where}$$

$$\hat{\delta}_k(\boldsymbol{x}) = \left[\log(n_k/n) - \frac{1}{2}\log(|\mathbf{S}_k|) - \frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)'\mathbf{S}_k^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)\right]$$

For binary classification, QDA is $\phi(\boldsymbol{x}) = 1$ if

$$-\frac{1}{2}\boldsymbol{x}'\widehat{\nabla}\boldsymbol{x} + (\mathbf{S}_1^{-1}\bar{\boldsymbol{x}}_1 - \mathbf{S}_2^{-1}\bar{\boldsymbol{x}}_2)'\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_1'\mathbf{S}_1^{-1}\bar{\boldsymbol{x}}_1 + \frac{1}{2}\bar{\boldsymbol{x}}_2'\mathbf{S}_2^{-1}\bar{\boldsymbol{x}}_2$$
$$-\frac{1}{2}\log(|\mathbf{S}_1|/|\mathbf{S}_2|) + \log(\hat{\pi}_1/\hat{\pi}_2) > 0 \text{ where } \widehat{\nabla} = \mathbf{S}_1^{-1} - \mathbf{S}_2^{-1}$$

or $= -1$ otherwise (assuming ± 1 coding)

Mahalanobis distance

In a special case where

$\Sigma_k \equiv \Sigma, \pi_k = 1/K$, for all k:

The Bayes rule classifier boils down to comparing the quantity $d_M^2(\boldsymbol{x}, \boldsymbol{\mu}_k) = (\boldsymbol{x} - \boldsymbol{\mu}_k)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$, which is called (squared) Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_k$.

- Mahalanobis distance $d_M(\boldsymbol{x}, \boldsymbol{\mu})$ measures how much $\boldsymbol{x}$ is away from the center of the distribution $N_p(\boldsymbol{\mu}, \Sigma)$.
- The set of points with the same Mahalanobis distance away from $\boldsymbol{\mu}$ is an ellipsoid.
- Replacing Σ with its estimator $\mathbf{S}$, and $\boldsymbol{x}$ with the sample mean $\bar{\boldsymbol{x}}$, the squared Mahalanobis distance is proportional to Hotelling's T^2 statistic.

Now consider an even more special case

Recall that the True QDA, which is the Bayes rule classifier for the Gaussian data, is

$$\phi_{\mathsf{QDA}}(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\, \eta_k(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\, \delta_k(\boldsymbol{x})$$

where

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= \left[\log(\pi_k) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right] \\ &= \left[\log(\pi_k) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}\boldsymbol{x}'\Sigma_k^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k'\Sigma_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k'\Sigma_k^{-1}\boldsymbol{x}\right]\end{aligned}$$

Now we assume Equal covariance: $\Sigma_k \equiv \Sigma$, then

$-\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}\boldsymbol{x}'\Sigma_k^{-1}\boldsymbol{x}$ will be same for all classes.

By removing $-\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}\boldsymbol{x}'\Sigma_k^{-1}\boldsymbol{x}$, $\delta_k(\boldsymbol{x})$ is simplified to

$$\begin{aligned}\tilde{\delta}_k(\boldsymbol{x}) &= \log(\pi_k) - \frac{1}{2}\boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{x} \\ &= b_{0k} + \boldsymbol{b}_k'\boldsymbol{x}, \text{ where} \\ b_{0k} &= \log(\pi_k) - \frac{1}{2}\boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{\mu}_k, \; \boldsymbol{b}_k = \Sigma^{-1}\boldsymbol{\mu}_k\end{aligned}$$

In this case, for binary classification $(K = 2)$: $\phi(\boldsymbol{x}) = 1$ when

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}) < \log(\pi_1/\pi_2),$$

that is

$$\boldsymbol{v}'\boldsymbol{x} - \boldsymbol{v}'\bar{\boldsymbol{\mu}} < \log(\pi_1/\pi_2), \quad \text{where } \boldsymbol{v} = \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

or $= -1$ otherwise.

Linear Discriminant Analysis

In practice, since we do not know the parameters $(\boldsymbol{\mu}_k, \Sigma, \pi_k)$, substitute

1. $\boldsymbol{\mu}_k$ with $\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k$,
2. Σ with $\widehat{\Sigma} = \mathbf{S}_P$ (pooled sample covariance matrix), and
3. $\hat{\pi}_k = n_k/n$.

Then we have the (sample) Linear Discriminant Analysis - estimated Bayes rule for Gaussian data with equal covariance assumption.

Linear Discriminant Analysis

Binary case,

$$\phi(\boldsymbol{x}) = 1 \text{ if } \boldsymbol{v}'(\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2}{2}) < \log(\frac{n_1}{n_2}), \quad \boldsymbol{v} = \mathbf{S}_P^{-1}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1).$$

In general, $\phi(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}(b_{0k} + \boldsymbol{b}_k'\boldsymbol{x})$, where $\boldsymbol{b}_k = \mathbf{S}_P^{-1}\bar{\boldsymbol{x}}_k$.

Quadratic Discriminant Analysis

Binary case, $\phi(\boldsymbol{x}) = 1$ if

$$-\frac{1}{2}\boldsymbol{x}'\widehat{\nabla}\boldsymbol{x} + (\mathbf{S}_1^{-1}\bar{\boldsymbol{x}}_1 - \mathbf{S}_2^{-1}\bar{\boldsymbol{x}}_2)'\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_1'\mathbf{S}_1^{-1}\bar{\boldsymbol{x}}_1 + \frac{1}{2}\bar{\boldsymbol{x}}_2'\mathbf{S}_2^{-1}\bar{\boldsymbol{x}}_2$$
$$-\frac{1}{2}\log(|\mathbf{S}_1|/|\mathbf{S}_2|) + \log(\hat{\pi}_1/\hat{\pi}_2) > 0 \text{ where } \widehat{\nabla} = \mathbf{S}_1^{-1} - \mathbf{S}_2^{-1}$$

Generally $\phi(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\left[\log(n_k/n) - \frac{1}{2}\log(|\mathbf{S}_k|) - \frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)'\mathbf{S}_k^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)\right]$

$$\text{Bayes rule} \xrightarrow{\text{estimate, Gaussian data}} \text{QDA} \xrightarrow{\text{equal covariance}} \text{LDA}$$

The next section would be ......

Fisher's LDA

- LDA is often referred to as `R.A. Fisher's` Linear Discriminant Analysis.
- His original work did not involve any distributional assumption, and developed LDA through a geometric understanding of PCA.
- The LDA direction $\boldsymbol{v}_0 \in \mathbb{R}^p$ is a direction vector orthogonal to the separating hyperplane, and is found by maximizing the between-group variance while minimizing the within-group variance of the projected scores.

$$\boldsymbol{v}_0 = \operatorname*{argmax}_{\boldsymbol{u} \in \mathbb{R}^p} \frac{(\boldsymbol{u}'\bar{\boldsymbol{x}}_1 - \boldsymbol{u}'\bar{\boldsymbol{x}}_2)^2}{\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u}} = \frac{\boldsymbol{u}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{u}}{\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u}}$$

$$\boldsymbol{v}_0 = \underset{\boldsymbol{u}\in\mathbb{R}^p}{\operatorname{argmax}} \frac{(\boldsymbol{u}'\bar{\boldsymbol{x}}_1 - \boldsymbol{u}'\bar{\boldsymbol{x}}_2)^2}{\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u}} = \frac{\boldsymbol{u}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{u}}{\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u}}$$

Because the objective is invariant with respect to the scaling of the vector $\boldsymbol{u}$, we can just assume that $\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u} = 1$. Then we can transform the problem into the following constrained optimization problem,

$$\begin{aligned} &\min\ \boldsymbol{u}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{u} \\ &\text{s.t. } \boldsymbol{u}'\mathbf{S}_P\boldsymbol{u} = 1 \end{aligned}$$

This is a generalized eigenvalue problem.

From Theorem 2.5 in *Applied Multivariate Statistical Analysis* by Härdle and Simar, we have

$$\boldsymbol{v}_0 = \text{first eigenvector}\{S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\}$$

Observe the following identity

$$\left[S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\right] S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = \lambda S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

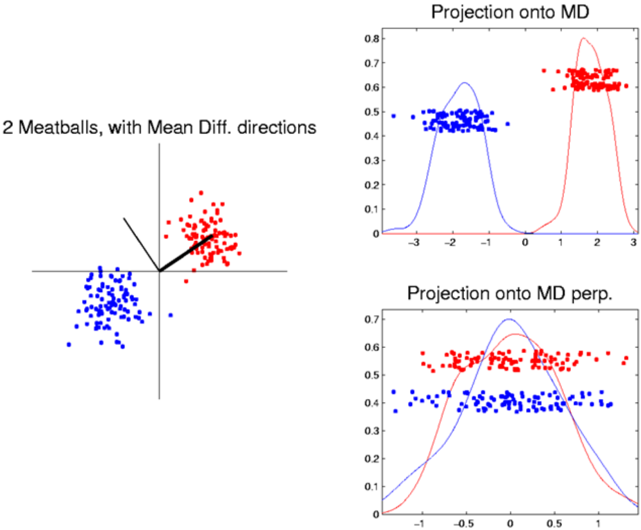
where $\lambda = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ which happens to be the greatest eigenvalue of $\left[S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\right]$. Hence the solution of $\boldsymbol{v}_0$ is actually

$$\boldsymbol{v}_0 = S_P^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

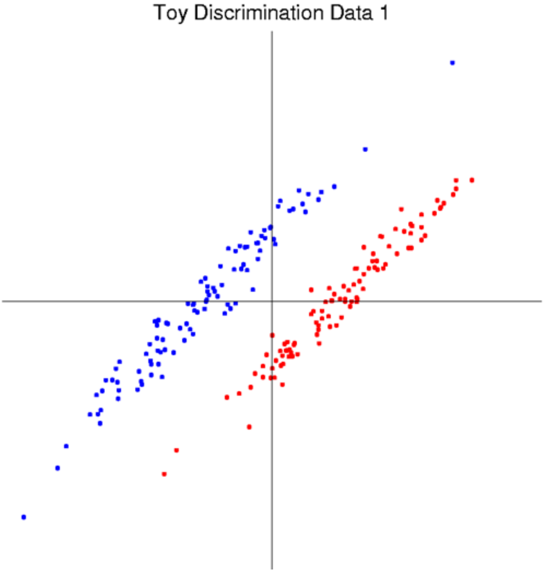
Fisher's LDA–Geometric understanding

Two data clouds, each with $\mathbf{S}_i = \mathbb{I}_2 = \mathbf{S}_P$.

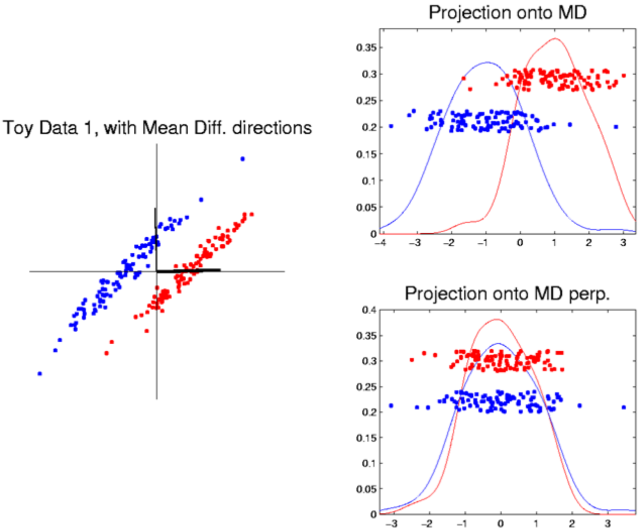
$$\boldsymbol{v}_0 \propto \mathbf{S}_P^{-1}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1) = (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1). \text{ (direction of mean difference)}$$



Slanted clouds. Assumed to have equal covariance.

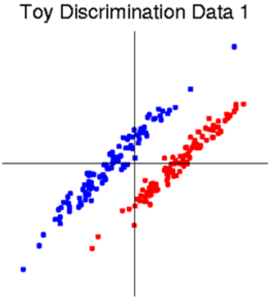


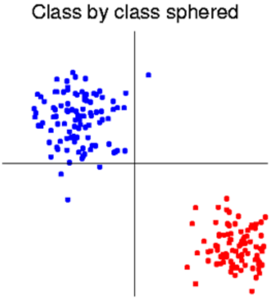
Mean difference direction not efficient, as $\mathbf{S}_P \neq c\mathbb{I}_2$.



Individually transform subpopulations so that both are '*spherical*' about their means.

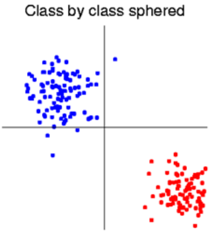
$$\tilde{\boldsymbol{x}}_{ij} = \mathbf{S}_P^{-1/2} \boldsymbol{x}_{ij}.$$





In transformed space, best separating hyperplane is the perpendicular bisector of line between means. Transformed mean diff. direction:

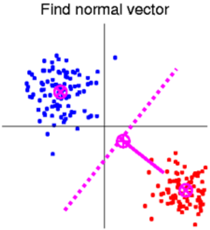
$$\tilde{\boldsymbol{v}} \propto \bar{\tilde{\boldsymbol{x}}}_2 - \bar{\tilde{\boldsymbol{x}}}_1 = \mathbf{S}_P^{-1/2}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1).$$



Transformed center:

$$\tilde{\boldsymbol{b}}_0 = (\bar{\tilde{\boldsymbol{x}}}_1 + \bar{\tilde{\boldsymbol{x}}}_2)/2 = \mathbf{S}_P^{-1/2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)/2.$$

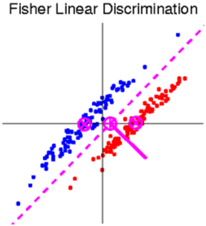
Transformed input $\tilde{\boldsymbol{x}} = \mathbf{S}_P^{-1/2}\boldsymbol{x}$ is classified to 1 if $\tilde{\boldsymbol{v}}'(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{b}}_0) < 0$

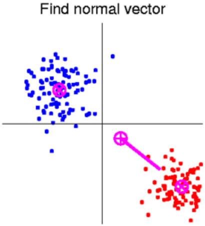


Original input $\boldsymbol{x} = \mathbf{S}_P^{1/2}\tilde{\boldsymbol{x}}$ is classified to 1 if

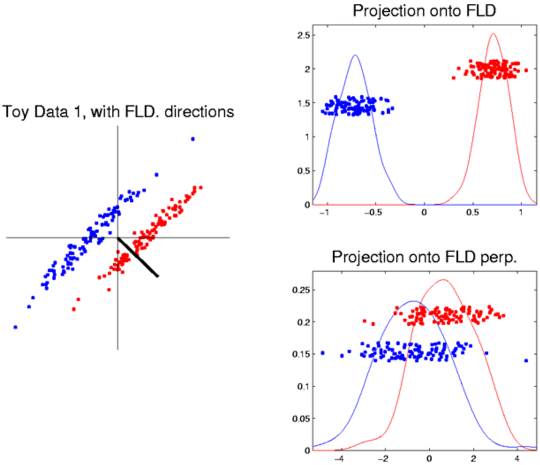
$$\begin{aligned} & \tilde{\boldsymbol{v}}'(\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{b}}_0) < 0 \\ \Leftrightarrow\ & \{\mathbf{S}_P^{-1/2}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)\}'(\mathbf{S}_P^{-1/2}\boldsymbol{x} - \mathbf{S}_P^{-1/2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)/2) < 0 \\ \Leftrightarrow\ & (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)'\mathbf{S}_P^{-1}(\boldsymbol{x} - (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)/2) < 0 \end{aligned}$$

THIS IS THE LDA RULE!!!





Leads to Fisher's LDA ($\boldsymbol{v}_0 \propto \mathbf{S}_P^{-1}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)$) by actively using covariance structure.



LDA vs QDA – Examples

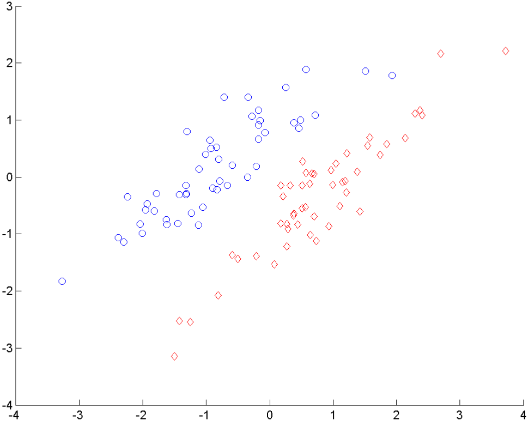
In the next four sets of examples,

- Blue and red points represent observations from two classes.
- Blue line is the *separating hyperplane* given by computing the sample LDA, and is a line perpendicular to *LDA direction* $\boldsymbol{b} = \mathbf{S}_P^{-1}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)$, and is the set

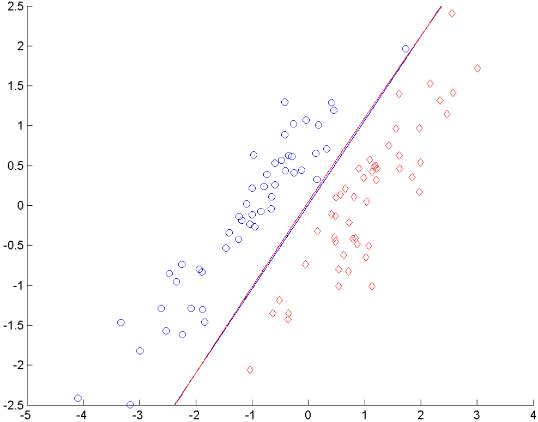
$$\{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{b}'(\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2}{2}) = \log(\frac{n_1}{n_2})\}.$$

- Red curve represents classification boundary of sample QDA $\{\boldsymbol{x} \in \mathbb{R}^2 : 0 = -\frac{1}{2}\boldsymbol{x}'\nabla\boldsymbol{x} + (\Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_2^{-1}\boldsymbol{\mu}_2)'\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}\log(|\Sigma_1|/|\Sigma_2|) + \log(\pi_1/\pi_2)\}$.

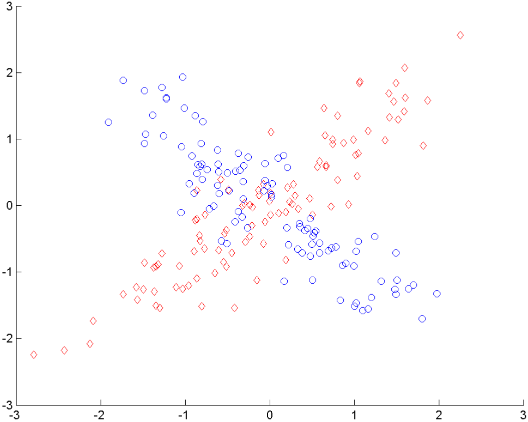
LDA vs QDA – Ex.1



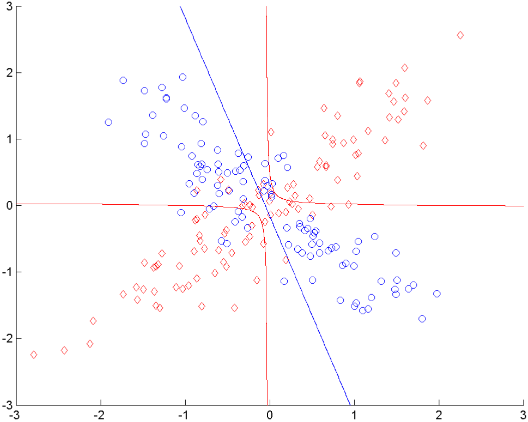
LDA vs QDA – Ex.1



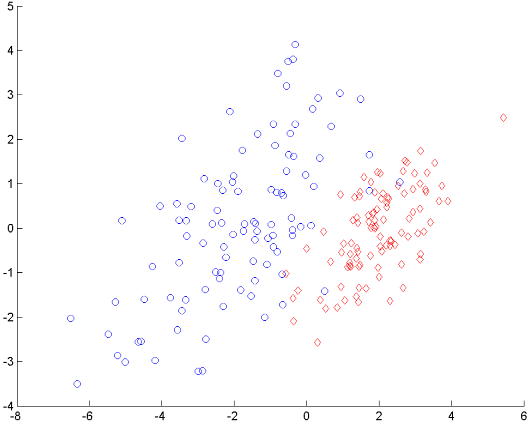
LDA vs QDA – Ex.2



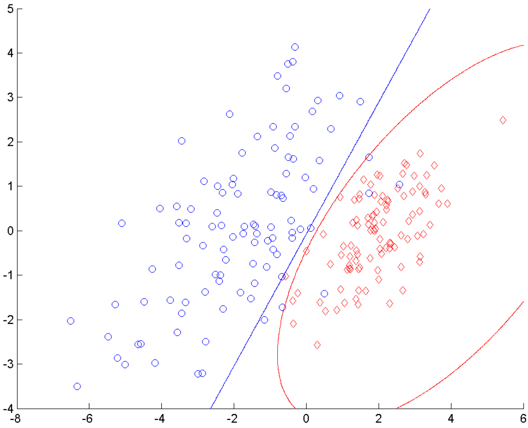
LDA vs QDA – Ex.2



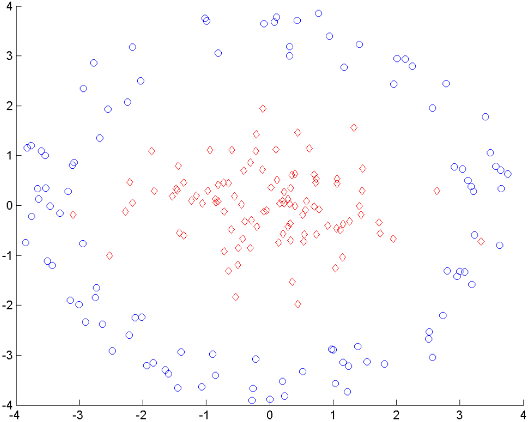
LDA vs QDA – Ex.3



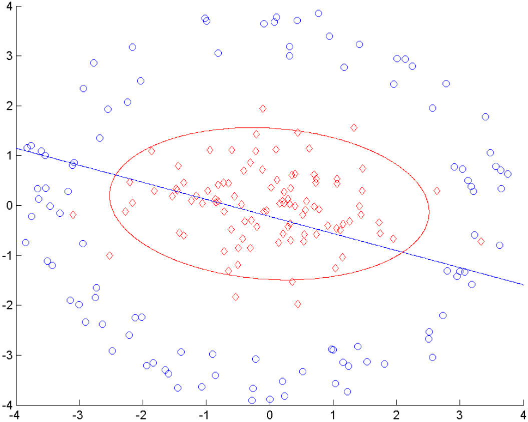
LDA vs QDA – Ex.3



LDA vs QDA – Ex.4



LDA vs QDA – Ex.4



The next section would be ......

Nearest Centroid (Mean Difference) rule

A simplification of LDA where $\mathbf{S}_P$ is replaced by $\mathbb{I}$

- the binary nearest centroid classifier is

$$\phi(\boldsymbol{x}) = 1 \text{ if } \boldsymbol{b}'(\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2}{2}) < 0, \quad \boldsymbol{b} = (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1),$$

  sometimes called mean difference classifier;

Naive Bayes classifier

A simplification of LDA where $\mathbf{S}_P$ is replaced by $\mathbf{D}_p = \mathrm{Diag}(\mathbf{S}_P)$ the diagonal matrix consisting of diagonal elements of $\mathbf{S}_P$

- the binary Naive Bayes classifier is

$$\phi(\boldsymbol{x}) = 1 \text{ if } \boldsymbol{b}'(\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2}{2}) < 0, \quad \boldsymbol{b} = \mathbf{D}_p^{-1}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1),$$

This is an (estimated) Bayes rule classifier which assumes that the (common) covariance matrix Σ is diagonal (which is quite a naive assumption).

LDA $\xrightarrow{\text{supress off-diagnol entries of covariance to 0}}$ naive Bayes $\xrightarrow{\text{force covariance to identity}}$ nearest centroid (mean difference) rule

Supervised dimension reduction

Recall that Fisher tried to maximize

$$\frac{(\boldsymbol{u}'\bar{\boldsymbol{x}}_1 - \boldsymbol{u}'\bar{\boldsymbol{x}}_2)^2}{\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u}} = \frac{\boldsymbol{u}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{u}}{\boldsymbol{u}'\mathbf{S}_P\boldsymbol{u}}$$

In general for $K \geq 3$, defined $\mathbf{B}$ as the covariance matrix of the class centroids and $\mathbf{W}$ the within-class covariance. $\mathbf{T} = \mathbf{W} + \mathbf{B}$ is the total covariance matrix of $\mathbf{X}$. Then we may iteratively solve

$$\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^T\mathbf{B}\boldsymbol{a}}{\boldsymbol{a}^T\mathbf{W}\boldsymbol{a}}$$

Find the optimal $\boldsymbol{a}_1$, then the next direction $\boldsymbol{a}_2$ orthogonal in $\mathbf{W}$ to $\boldsymbol{a}_1$ such that $\boldsymbol{a}_2^T\mathbf{B}\boldsymbol{a}_2/\boldsymbol{a}_2^T\mathbf{W}\boldsymbol{a}_2$ is maximized...

Similar to PCA, this is a way of dimension reduction. It finds directions that best separate the classes.

- These $\boldsymbol{a}_k$'s form a subspace $\text{span}\{\boldsymbol{a}_k\}_{k=1}^K$.
- It is equivalent to apply standard LDA to, instead of the original data, the projected data onto subspace $\text{span}\{\boldsymbol{a}_k\}_{k=1}^K$.
- Moreover, it can be shown that[1] this subspace is the same as

$$\mathbf{W}^{-1}\text{span}\{\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_\ell\}_{1\le k\ne\ell\le K}.$$

[1] Niu, Y.S., Hao, N. and Dong, B. (2018) "A New Reduced-Rank Linear Discriminant Analysis Method and Its Applications." Statistica Sinica, 28, 189-202.

Reduced-rank LDA

- Therefore, the original LDA can be viewed as first projecting the data to the subspace $\text{span}\{\boldsymbol{a}_k\}_{k=1}^{K}$ (as a form of dimension reduction), then applying standard LDA to the projected data.

- We may reduce the dimension more aggressively. The idea of the **reduced-rank LDA** is to find and project the data to a smaller subspace ($\text{span}\{\boldsymbol{a}_k\}_{k=1}^{L}$) (with $L < K$) before applying standard LDA.

- Treat L as a measure of model complexity and think about the trade-off between model fit and model complexity.

Strengths & Weaknesses of LDA.

- Logistic regression is less sensitive to non-Gaussian data, since there is no Gaussian assumption. Beats LDA when non-Gaussian or the covariance is not equal
- LDA is subject to outliers.
- LDA exploits the full likelihood while logistic regression uses conditional likelihood.
- Logistic regression less efficient than LDA (needs a large sample to work well.) LDA can be quite flexible.
- Both have big problem when $p \gg n$.

The next section would be ......

Regularized Discriminant Analysis

- Friedman (1989) proposed a compromise between LDA and QDA: replace $\mathbf{S}_k$ by

$$\mathbf{S}_k(\alpha) = \alpha \mathbf{S}_k + (1-\alpha)\mathbf{S}_P$$

- Similar modifications allow $\mathbf{S}_P$ to be shrunk toward the scalar covariance,: replace $\mathbf{S}_P$ by

$$\mathbf{S}_P(\alpha) = \alpha \mathbf{S}_P + (1-\alpha)\hat{\sigma}^2 \mathbb{I}$$

- Regularization with sparsity effect (next page)

Sparse LDA

FSDA (Wu et al., 2008):

$$\underset{\beta}{\operatorname{argmin}}\, \beta^T \widehat{\Sigma} \beta + \lambda \|\beta\|_1, \text{ subject to } (\hat{\mu}_2 - \hat{\mu}_1)^T \beta = 1$$

DSDA (Mai et al. 2012):

$$\beta \propto \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (\tilde{y}_i - \beta^T x_i)^2 + \lambda \|\beta\|_1$$

PDA (Witten and Tibshirani 2011):

$$\underset{\beta}{\operatorname{argmax}}\, \beta^T \mathbf{B} \beta - \lambda \sum_{j=1}^{p} \hat{\sigma}_j |\beta_j|, \text{ subject to } \beta^t \mathbf{W} \beta \leq 1$$

SOS (Clemmensen et al., 2011), ROAD (Fan et al. 2012), and many many others.