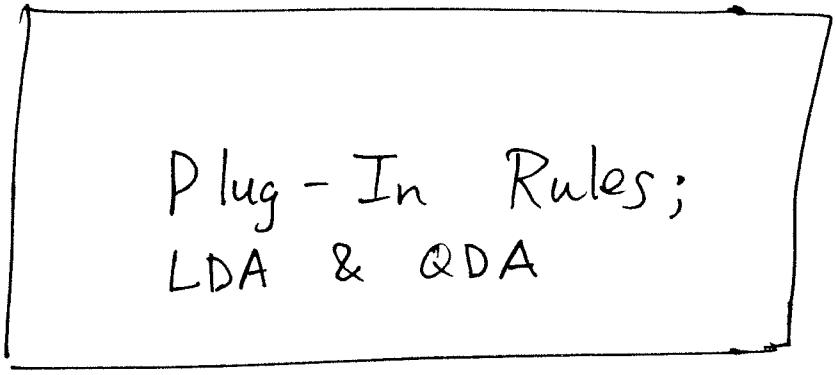

Plug-In Rules;
LDA & QDA

Plug-In Rules

Recall the Bayes classifier:

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ 0 & \text{if } \eta(x) < \frac{1}{2} \end{cases}$$

$$= \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} \geq \frac{\pi_0}{\pi_1} \\ 0 & \text{otherwise} \end{cases}$$ discrete

$$= \begin{cases} 1 & \text{if } \frac{g_1(x)}{g_0(x)} \geq \frac{\pi_0}{\pi_1} \\ 0 & \text{otherwise} \end{cases}$$ continuous

A <u>plug-in classifier</u> estimates $\eta(x) = P\{Y=1 \mid X=x\}$ or the pmfs / pdfs of $X \mid Y=y$, and "plugs" the estimate in to the formulas above.

<u>Recall</u>: Such estimates are based on the training data

$$(x_i, y_i) \in \mathbb{R}^d \times \{0,1\}, \quad i = 1, \ldots, n.$$

Discrete data

Suppose X is discrete, and let

$$z_1, \ldots, z_m$$

be the distinct values taken on by $x_1, \ldots, x_n$.

There are two approaches: estimate $\eta(x)$ or estimate $p_0(x)$ and $p_1(x)$.

Let's consider the first approach:

$$\hat{\eta}(z_k) =$$

What's the picture?

is the simplest estimate.

What about the second approach?

$$\hat{p}_0(z_k) = \frac{|\{i : y_i = 0, x_i = z_k\}|}{n_0}$$

$$\hat{p}_1(z_k) = \frac{|\{i : y_i = 1, x_i = z_k\}|}{n_1}$$

Claim: These two plug in estimates are equivalent.

Notation:

$$n(k) = |\{i : x_i = z_k\}|$$

$$n_0(k) = |\{i : x_i = z_k, y_i = 0\}|$$

$$n_1(k) = |\{i : x_i = z_k, y_i = 1\}|$$

Then

$$\hat{\eta}(z_k) = \frac{n_1(k)}{n(k)} \geq \frac{1}{2}$$

$$\Updownarrow$$

$$n_1(k) \geq n_0(k)$$

and

$$\frac{\hat{P}_1(z_k)}{\hat{P}_0(z_k)} \geq \frac{\hat{\pi}_0}{\hat{\pi}_1} = \frac{\frac{n_0}{n}}{\frac{n_1}{n}}$$

$$\Updownarrow$$

$$\frac{\frac{n_1(k)}{n_1}}{\frac{n_0(k)}{n_0}} \geq \frac{n_0}{n_1}$$

pretty simple!

$$\Updownarrow$$

$$n_1(k) \geq n_0(k).$$

What if we observe

$$x \notin \{z_1, \ldots, z_m\}?$$

We just have to guess the label?

What if $n(k)$ is so small that our estimates are unreliable (high variance)?

One can assume a <u>parametric</u> model for $\eta(x)$ or for $p_1(x)$, $p_0(x)$. The method we just discussed is <u>nonparametric</u>.

<u>Example</u>

x_i = # of cars on Plymouth road in a day

$$y_i = \begin{cases} 1 & \text{if it's a weekday} \\ 0 & \text{if its the weekend} \end{cases}$$

Given x, predict y.

What if we only have 20 days worth of data? Then it is highly likely that a future x is not one of the observed x_i.

A solution: Assume

$$X|Y=0 \sim \text{Poisson}(\lambda_0)$$

$$X|Y=1 \sim \text{Poisson}(\lambda_1).$$

$$\text{MLE} \Longrightarrow \quad \hat{\lambda}_0 = \frac{1}{n_0} \sum_{i:y_i=0} x_i = \bar{x}_0$$

$$\hat{\lambda}_1 = \frac{1}{n_1} \sum_{i:y_i=1} x_i = \bar{x}_1$$

So the plug in classifier is

$$\hat{f}_n(x)=1 \iff \frac{\hat{p}_1(x)}{\hat{p}_0(x)} \geq \frac{\hat{\pi}_0}{\hat{\pi}_1} \iff \frac{\frac{e^{-\hat{\lambda}_1}\hat{\lambda}_1^x}{x!}}{\frac{e^{-\hat{\lambda}_0}\hat{\lambda}_0^x}{x!}} \geq \frac{n_0}{n_1}$$

$$\iff \quad x \geq \frac{\log\left(\frac{n_0}{n_1}\right) + (\bar{x}_1 - \bar{x}_0)}{\log\left(\bar{x}_1/\bar{x}_0\right)}$$

(assuming $\bar{x}_1 > \bar{x}_0$)

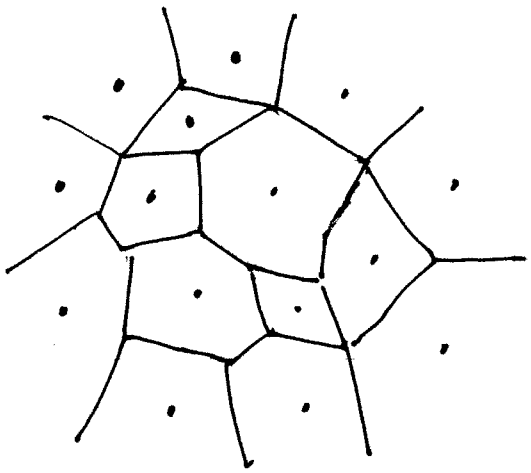
Continuous data

Classification of continuous data is (or can be) more challenging for a couple of reasons.

- There is no obvious nonparametric plug in method
- Parametric models are often not as obvious as they are for discrete data

Two broad approaches:

- Quantize your data to make it discrete
- Density estimation



histogram

Density Estimation

The goal is to estimate $g_0(x)$, $g_1(x)$

from $(x_1, y_1), \ldots, (x_n, y_n)$

Three approaches

- parametric : assumes specific form for $g_0(x)$, $g_1(x)$, e.g. Gaussian. Works well if this assumption is correct, works poorly otherwise
- nonparametric : makes no assumption about specific form of densities
- semiparametric : a compromise between parametric and nonparametric

We will focus on the following incarnations of these approaches:

parametric

- LDA
- QDA

nonparametric

- kernel density estimation

semiparametric

- Gaussian mixture model

LDA and QDA

Assume

$$g_y(x) = g_y(x|\mu_y, \Sigma_y) = (2\pi)^{-\frac{d}{2}} |\Sigma_y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y)\right\}$$

$y = 0, 1.$

Exercise] Assuming these models are correct, and that $\Sigma_0 = \Sigma_1 = \Sigma$, show that the Bayes classifier can be written

$$f^*(x) = \begin{cases} 1 & \text{if } a^T x + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Solution

$$\frac{g_1(x)}{g_0(x)} = \exp\left\{-\frac{1}{2}\left[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\left(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1\right) - \left(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}\mu_0\right)\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[-2x^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)\right]\right\}$$

Take logarithm:

$$\frac{g_1(x)}{g_0(x)} \gtrless \frac{\pi_0}{\pi_1} \iff$$

$$\iff x^T\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) \gtrless \log\left(\frac{\pi_0}{\pi_1}\right)$$

$$\iff a^Tx + b \gtrless 0$$

where

$$a = \Sigma^{-1}(\mu_1-\mu_0)$$

$$b = \frac{-\mu_1^T\Sigma^{-1}\mu_1 + \mu_0^T\Sigma^{-1}\mu_0}{2} - \log\left(\frac{\pi_0}{\pi_1}\right)$$

Linear discriminant analysis

- estimate μ_0, μ_1, Σ
- plug in to Bayes classifier

$\Rightarrow$ linear classifier

How do we estimate μ_0, μ_1, Σ? The most common method is maximum likelihood estimation:

$$\hat{\mu}_0 = \frac{1}{n_0} \sum_{i: y_i = 0} x_i$$

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i: y_i = 1} x_i$$

$$\hat{\Sigma} = \frac{1}{n} \left(\sum_{i: y_i = 0} (x_i - \hat{\mu}_0)(x_i - \hat{\mu}_0)^T + \sum_{i: y_i = 1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T \right)$$

In addition,

$$\hat{\pi}_1 = \frac{n_1}{n}$$

$$\hat{\pi}_0 = \frac{n_0}{n}$$

Interpretation of LDA

We may write the classifier as

$$(x-\hat{\mu}_1)^T \hat{\Sigma}^{-1}(x-\hat{\mu}_1) - 2\log\hat{\pi}_1 \underset{1}{\overset{0}{\gtrless}} (x-\hat{\mu}_0)^T \hat{\Sigma}^{-1}(x-\hat{\mu}_0) - 2\log\hat{\pi}_0$$

Assume for simplicity that $n_0 = n_1$.

Then the LDA classifier is

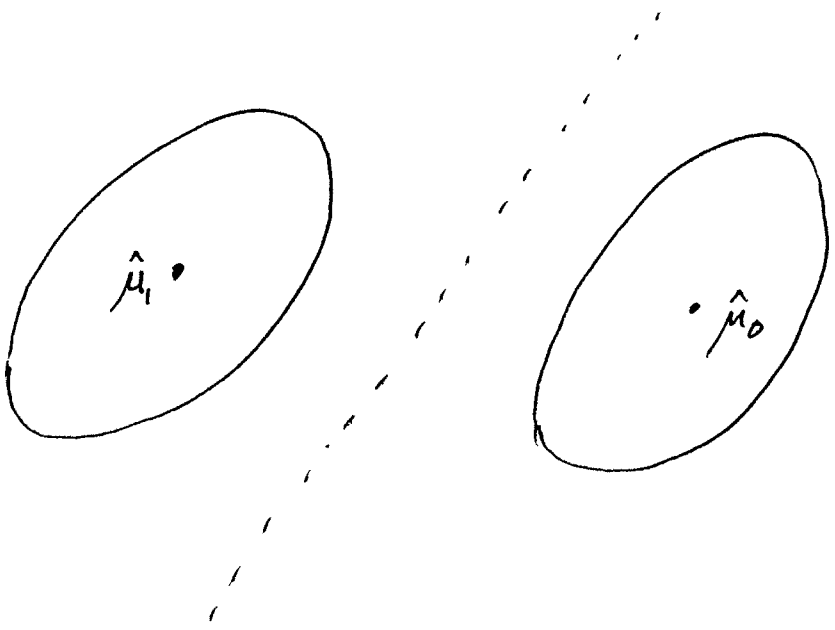
$$d_M(x \mid \hat{\mu}_1, \hat{\Sigma}) \underset{1}{\overset{0}{\gtrless}} d_M(x \mid \hat{\mu}_0, \hat{\Sigma}),$$

where

$$d_M(x \mid \mu, \Sigma) = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

is the Mahalanobis distance from x to μ.

Claim: $\{ x : d_M(x \mid \mu, \Sigma) = r^2 \}$ is an ellipse



Proof of claim: Write $\Sigma = \sum_{i=1}^{d} \lambda_i v_i v_i^T = V \Lambda V^T$.

Introduce the change of coordinates

$$x' = V^T (x - \mu)$$

orthogonal

Then

$$r^2 = \sum_{i=1}^{d} \frac{1}{\lambda_i} (x-\mu)^T v_i v_i^T (x-\mu)$$

$$= \sum \frac{1}{\lambda_i} (x_i')^2$$

Is the equation of an ellipse.

When $n_1 \neq n_0$, the hyperplane is shifted toward the less populous class.

Other special cases

- $\Sigma = \sigma^2 I$.

  Then Mahalanobis distance $\propto$ Euclidean distance. If $n_0 = n_1$, then we have

$$\frac{1}{\hat{\sigma}^2}\| x - \hat{\mu}_1 \|^2 \underset{1}{\overset{0}{\gtrless}} \frac{1}{\hat{\sigma}^2}\| x - \hat{\mu}_0 \|^2$$

$$\Updownarrow$$

$$\| x - \hat{\mu}_1 \|^2 \underset{1}{\overset{0}{\gtrless}} \| x - \hat{\mu}_0 \|^2$$

$$\Rightarrow \boxed{\text{nearest centroid classifier}}$$

- $\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_d^2 \end{bmatrix}$

$$\Rightarrow \sum_{i=1}^{d} \frac{1}{\hat{\sigma}_i^2}(x_i - \hat{\mu}_{1,i})^2 \underset{1}{\overset{0}{\gtrless}} \sum_{i=1}^{d} \frac{1}{\hat{\sigma}_i^2}(x_i - \hat{\mu}_{0,i})^2$$

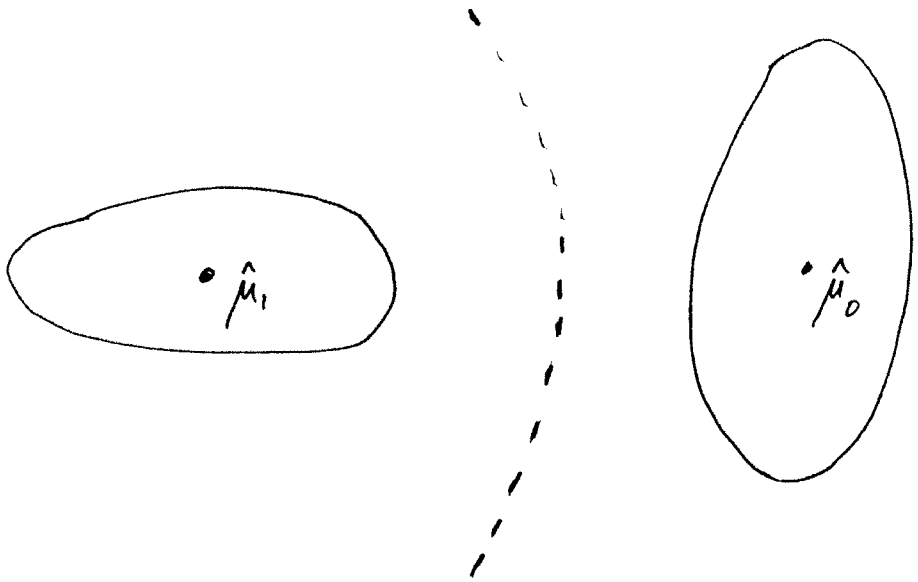
$\Rightarrow$ **weighted nearest centroid classifier**

If $n_0 \neq n_1$, hyperplane shifts accordingly.

Quadratic discriminant analysis (QDA)

$$\Sigma_0 \neq \Sigma_1 \Rightarrow \hat{\Sigma}_0 \neq \hat{\Sigma}_1$$

$\Rightarrow$ quadratic decision boundary



Mahalanobis distance interpretation still holds:

$$d_M(x \mid \hat{\mu}_1, \hat{\Sigma}_1) - 2 \log \hat{\pi}_1 \underset{1}{\overset{0}{\gtrless}} d_M(x \mid \hat{\mu}_0, \hat{\Sigma}_1) - 2 \log \hat{\pi}_0$$

where now

$$\hat{\Sigma}_0 = \frac{1}{n_0} \sum_{i: y_i = 0} (x_i - \hat{\mu}_0)(x_i - \hat{\mu}_0)^T$$

$$\hat{\Sigma}_1 = \frac{1}{n_1} \sum_{i: y_i = 1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T$$

The name quadratic discriminant analysis comes from the fact that now

$$f^*(x) = \begin{cases} 1 & \text{if } x^T C x + a^T x + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

for some C, a, b.

Multiple classes

The Bayes classifier is

$$f^*(x) = \arg\max_y \pi_y \cdot g_y(x)$$

So a plug-in classifier is

$$\hat{f}_n(x) = \arg\max_y \hat{\pi}_y \cdot g_y(x \mid \hat{\mu}_y, \hat{\Sigma}_y)$$

$$= \arg\min_y -2\log \hat{\pi}_y + d_M(x \mid \hat{\mu}_y, \hat{\Sigma}_y).$$

So the LDA / QDA classifier for multiple classes is

> "given a test point x, assign the label of the class to which x is closest in terms of Mahalanobis distance (adjusted by $-2\log \hat{\pi}_y$)."

LDA picture ($\Sigma_y = \Sigma$)

Four classes