Quadratic discriminant analysis (QDA)

LDA: $f_1(x), \cdots f_k(x)$ ← multivariate normal means $\mu_1, \cdots, \mu_k$ and covariance C

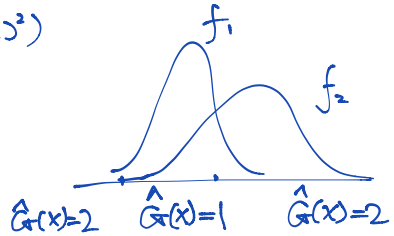
sub-populations

- optimal classifications rule → linear boundaries
- now allow different covariance matrices
  - get boundaries determined by paraboloids

Example: $p=1$, $k=2$ $(\lambda_1=\lambda_2=\frac{1}{2})$

$$f_1(x)=\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)$$

$$f_2(x)=\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right)$$

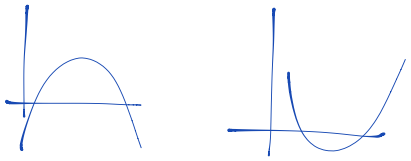


$\hat{G}(x)=1$ if $f_1(x) > f_2(x)$

$$\ln\left(\frac{f_1(x)}{f_2(x)}\right) > 0$$

$$\ln\left(\frac{f_1(x)}{f_2(x)}\right)=\frac{1}{2\sigma_2^2}(x-\mu_2)^2-\frac{1}{2\sigma_1^2}(x-\mu_1)^2+\ln(\sigma_2/\sigma_1)$$

$$=\left[\frac{1}{2\sigma_2^2}-\frac{1}{2\sigma_1^2}\right]x^2-\left(\frac{\mu_2}{\sigma_2^2}-\frac{\mu_1}{\sigma_1^2}\right)x+\frac{\mu_2^2}{\sigma_2^2}-\frac{\mu_1^2}{\sigma_1^2}+\ln(\sigma_2/\sigma_1)$$



QDA in practice: need to estimate $\mu_1, \ldots \mu_k$ and $C_1, \cdots, C_k$

$$\hat{\mu}_j=\frac{\sum_{i=1}^{n} x_i I(g_i=j)}{\sum_{i=1}^{n} I(g_i=j)} \qquad (j=1,\cdots,k)$$

$$\hat{C}_j=\sum_{i=1}^{n}(x_i-\hat{\mu}_j)(x_i-\hat{\mu}_j)^T I(g_i=j)\Big/\left(\sum_{i=1}^{n} I(g_i=j)-1\right)$$

Example (on Blackboard) Iris data

3 species of iris

4 variables: length of sepal
width
length of petal
width

- estimate misclassification rate using leave-one-out cross-validation: error rate $=\frac{4}{150}$

(continued)

- generate 15000 observations from 3 sub-populations using the estimated mean vectors $\hat{\mu_1}, \hat{\mu_2}, \hat{\mu_3}$ and covariance matrices C_1, C_2, C_3
- classification rate every similar: 1.2% of simulated observations are classified differently.

Alternative mehtods for classification

So far: have data $(g_1, x_1), \ldots, (g_n, x_n)$
Model:$(G, X) \rightarrow P(G = j) = \lambda_j$ and conditional on $G = j$, X has density $f_j(X)$.

$$P(G = j|X = x) = \frac{\lambda_j f_j(x)}{\sum_{l=1}^{k} \lambda_l f_l(x)}$$

quantity of interest

LDA, QDA: Assume $f_1(x), \ldots, f_k(x)$ multivariates normal densities -- use data to estimate unknowns.

Key point: Explicitly model distributions of X.
But in prctice, this is difficult to do

- discrete variables
- P may be very large

Alternative approach: model $P(G = j|X = x)$ directly.

- analogous to regression modeling
  - G is the response
  - X is the predictors
- we implicityly assume that the distribution of X (i.e. not condtional on $G = j$) is not particularly informative.

Multiple regression: $y_i = x_i^T \beta + \epsilon_i$ --> look at conditional distributions of response given predictors.

Special case: $k = 2$

$$P(G = 1|X = x) = 1 - P(G = 2|X = x)$$

where the LHS is $\theta(x) = g(x, \beta)$, and β is unknown parameters.

Logistic regression model

$$\theta(x) = \frac{\exp(\beta_0 + x^T \beta)}{1 + \exp(\beta_0 + x^T \beta)}$$

$$1-\theta(\boldsymbol{x}) = P(G=2|X=\boldsymbol{x}) = \frac{1}{1+\exp(\beta_0+\boldsymbol{x}^T\boldsymbol{\beta})}$$

Note that $0 < \theta(\boldsymbol{x}) < 1$ for any $\beta_0, \boldsymbol{\beta}$

- **logit transform**:

$$\text{logit}(\theta(\underset{\sim}{x})) = \ln\left(\frac{\theta(\underset{\sim}{x})}{1-\theta(\underset{\sim}{x})}\right) = \beta_0 + \underset{\sim}{x}^T\underset{\sim}{\beta} \quad \begin{cases} >0 \Rightarrow \theta(\underset{\sim}{x}) > \frac{1}{2} & (\hat{G}=1) \\ <0 \Rightarrow \theta(\underset{\sim}{x}) < \frac{1}{2} & (\hat{G}=2) \end{cases}$$

Estimation: Given $(g_1, \underset{\sim}{x}_1), \cdots, (g_n, \underset{\sim}{x}_n)$ the likelihood function is

$$L(\beta_0, \underset{\sim}{\beta}) = \prod_{i=1}^n \{\theta(\underset{\sim}{x}_i)^{I(g_i=1)}\}^{y_i} (1-\theta(\underset{\sim}{x}_i))^{1-I(g_i=1)} \quad 1-y_i$$

$$= \prod_{i=1}^n \left\{\left[\frac{\theta(\underset{\sim}{x}_i)}{1-\theta(\underset{\sim}{x}_i)}\right]^{y_i} (1-\theta(\underset{\sim}{x}_i))\right\}$$

$$\ln L(\beta_0, \underset{\sim}{\beta}) = \sum_{i=1}^n \{y_i(\beta_0 + \underset{\sim}{x}_i^T\underset{\sim}{\beta}) - \ln(1+\exp(\beta_0 + \underset{\sim}{x}_i^T\underset{\sim}{\beta}))\}$$

$$\frac{\partial}{\partial\beta_0} \ln L(\beta_0, \underset{\sim}{\beta}) = \sum_{i=1}^n \left\{y_i - \frac{\exp(\beta_0 + \underset{\sim}{x}_i^T\underset{\sim}{\beta})}{1+\exp(\beta_0 + \underset{\sim}{x}_i^T\underset{\sim}{\beta})}\right\}$$

$$\frac{\partial}{\partial\underset{\sim}{\beta}} \ln L(\beta_0, \underset{\sim}{\beta}) = \sum_{i=1}^n \left\{y_i - \frac{\exp(\beta_0 + \underset{\sim}{x}_i^T\underset{\sim}{\beta})}{1+\exp(\beta_0 + \underset{\sim}{x}_i^T\underset{\sim}{\beta})}\right\} \underset{\sim}{x}_i$$

Set derivatives equal to 0 an solve for MLEs
R function: glm(group ~ x1+x2+..., family=binomial)

Example: Generate data from 2 bivariate normal distributions

$$f(\underset{\sim}{x}) = \frac{1}{4} f_1(\underset{\sim}{x}) + \frac{3}{4} f_2(\underset{\sim}{x}) \leftarrow N_2\left(\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$$

$$N_2\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right) \qquad n=200$$

① Use LDA (optimal for those data)
② Use logistic regression
Error rate using LOO CV:
LDA: 17/200
Logistic reg: 18/200

LDA

observed	Classified 1	2	
1	37	8	45
2	9	146	155

Logistic Reg

observed	Classified 1	2	
1	35	10	45
2	8	147	155

Question: Is this similarity a coincidence?

Not really. For multivariate normal model we have:

$$\theta(\underset{\sim}{x}) = P(G=1 \mid \underset{\sim}{X}=\underset{\sim}{x}) = \frac{\lambda_1 f_1(\underset{\sim}{x})}{\lambda_1 f_1(\underset{\sim}{x}) + \lambda_2 f_2(\underset{\sim}{x})} = \frac{\exp(\beta_0 + \underset{\sim}{x}^T \underset{\sim}{\beta})}{1+\exp(\beta_0 + \underset{\sim}{x}^T \underset{\sim}{\beta})}$$

→ depend on $\lambda_1, \lambda_2, \underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2$ and C

Also MLEs in logistic regression satisfy the equation

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\underset{\sim}{\beta}} \end{pmatrix} = (X^T W X)^{-1} X^T W \underset{\sim}{z}$$

where $X = \begin{pmatrix} 1 & \underset{\sim}{x}_1^T \\ \vdots & \vdots \\ 1 & \underset{\sim}{x}_n^T \end{pmatrix}$ $\quad W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$ with $w_i = [\hat{\theta}(\underset{\sim}{x}_i)(1-\hat{\theta}(\underset{\sim}{x}_i))]^{-1}$

$\underset{\sim}{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$ with $z_i = \ln\left(\frac{\hat{\theta}(\underset{\sim}{x}_i)}{1-\hat{\theta}(\underset{\sim}{x})}\right) + \frac{y_i - \hat{\theta}(\underset{\sim}{x}_i)}{\hat{\theta}(\underset{\sim}{x})(1-\hat{\theta}(\underset{\sim}{x}_i))}$

Boundary: $\hat{\beta}_0 + \underset{\sim}{x}^T \underset{\sim}{\beta} = 0$

$$\| $$

$$(1-\underset{\sim}{x}^T)(X^T W X)^{-1} X^T W \underset{\sim}{z}$$

If $\hat{\theta}(\underset{\sim}{x}_i)(1-\hat{\theta}(\underset{\sim}{x}_i))$ is approximately constant then boundaries defined by logistic regression are close to those for LDA.

Multinomial logit

Need to find a model for

$$\theta_j(\underset{\sim}{x}) = P(G=j \mid \underset{\sim}{X}=\underset{\sim}{x}) \quad \text{for } j=1,\dots,k>2$$

Note: $\theta_1(\underset{\sim}{x}) + \theta_2(\underset{\sim}{x}) + \cdots + \theta_k(\underset{\sim}{x}) = 1$ for each $\underset{\sim}{x}$

Assume that for each $j=1,\dots,k-1$, we have $\ln\left(\frac{\theta_j(\underset{\sim}{x})}{\theta_k(\underset{\sim}{x})}\right) = \beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j$

$$\frac{\theta_j(\underset{\sim}{x})}{\theta_k(\underset{\sim}{x})} = \exp(\beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j)$$

$$\theta_j(\underset{\sim}{x}) = \theta_k(\underset{\sim}{x}) \exp(\beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j)$$

Now sum RHS, LHS over j from 1 to $k-1$

$$\sum_{j=1}^{n-1} \theta_j(\underset{\sim}{x}) = 1 - \theta_k(\underset{\sim}{x}) = \theta_k(\underset{\sim}{x}) \sum_{j=1}^{n-1} \exp(\beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j)$$

$$\Rightarrow \theta_k(\underset{\sim}{x}) = \frac{1}{1+\sum_{j=1}^{k-1} \exp(\beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j)}$$

$$\Rightarrow \theta_j(\underset{\sim}{x}) = \frac{\exp(\beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j)}{1 + \sum_{j=1}^{n-1} \exp(\beta_{0j} + \underset{\sim}{x}^T \underset{\sim}{\beta}_j)}$$

In R: ① package "mlogit" → allows you to fit very complicated multi-logits models

② Package "nnet": function multinorm is a simpler option for this course.

Comments: Model $(G, \underset{\sim}{X})$ vs model $G|\underset{\sim}{X}=\underset{\sim}{x}$.

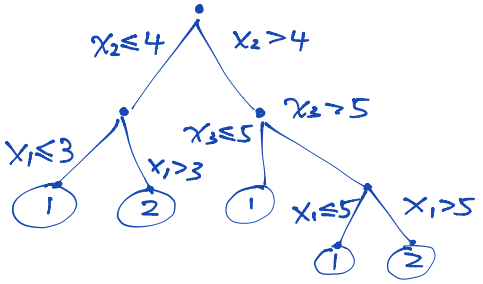
① Which is better?
- Modeling $(G, \underset{\sim}{X})$ is better if you have a good model for $(G, \underset{\sim}{X})$
- Otherwise modeling $G|\underset{\sim}{X}=\underset{\sim}{x}$ gives more flexibility

② Modeling $P(G=j \mid \underset{\sim}{X}=\underset{\sim}{x})$ allows to apply many regression modeling techniques
- variables selection (subject selection)
- also use shrinkage / Bayesian methodology

Classification trees

Idea: Use data to construct a binary tree to classify observations.

e.g. $k=2$, $p=3$ (x_1, x_2, x_3)



Advantages: very easy to interpret
- adapt to different types of variables (cont., discrete, categorical, ordinal)

Disadvantages: • Computationally complex
- When to stop growing tree?
- No clear model