STAT 5630

Discriminant Analysis and Logistic Regression

Xiwei Tang, Ph.D. <xt4yj@virginia.edu>

Department of Statistics, University of Virginia

October 1, 2019

Outline

- Classification problems and Bayes rule
- LDA and QDA
- Logistic Regression

Classification Problems

- We have n observations, p dimentional covariates X, and a binary outcome Y.
- Training data $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$:
  - $x_i \in \mathbb{R}^p$, and $y_i \in \{0, 1\}$ (sometimes we use $\{-1, +1\}$).
- The goal is to find a classifier

$$f : \mathbb{R}^p \longrightarrow \{0, 1\}$$

- At any target point x with outcome y, the performance of a classifier is usually measured by 0–1 loss

$$L(f(x), y) = \begin{cases} 0 & \text{if} \quad y = f(x) \\ 1 & \text{if} \quad \text{o.w.} \end{cases}$$

Bayes Rule

- The optimal classifier is the one that minimizes the risk

$$\mathsf{R}(f) = \mathsf{E}_{X,Y}\left[L\big(Y, f(X)\big)\right]$$

- If we define the probability of $Y = 1$ at each target point x as

$$\eta(x) = \mathsf{P}(Y = 1 | X = x)$$

then for 0–1 loss, the best rule we can get is

$$\arg\min_f \mathsf{R}(f) = f_B(x) = \begin{cases} 1 & \text{if} \quad \eta(x) > 1/2 \\ 0 & \text{if} \quad \eta(x) < 1/2. \end{cases}$$

- This rule f_B is called the Bayes rule, and the corresponding risk $\mathsf{R}(f_B)$ is referred to as the Bayes risk or Bayes error.

Bayes Rule

- The name of “Bayes rule” comes from understanding the optimal rule from the Bayes prospective:

$$\mathsf{P}(Y=1|X=x) = \frac{\mathsf{P}(X=x|Y=1)\mathsf{P}(Y=1)}{\mathsf{P}(X=x)}$$
$$\mathsf{P}(Y=0|X=x) = \frac{\mathsf{P}(X=x|Y=0)\mathsf{P}(Y=0)}{\mathsf{P}(X=x)}$$

- Hence, treating $\pi = \mathsf{P}(Y=1)$ and $(1-\pi) = \mathsf{P}(Y=0)$ as prior probabilities, and define the conditional densities of X as

$$f_1 = \mathsf{P}(X=x|Y=1) \text{ and } f_0 = \mathsf{P}(X=x|Y=0).$$

- The Bayes rule can also be written as

$$\arg\min_f \mathsf{R}(f) = f_B(x) = \begin{cases} 1 & \text{if} \quad \pi f_1(x) > (1-\pi) f_0(x) \\ 0 & \text{if} \quad \pi f_1(x) < (1-\pi) f_0(x). \end{cases}$$

Bayes Rule

- The decision boundary can be used to describe the optimal rule:

$$\{x : \pi f_1(x) = (1 - \pi) f_0(x)\}$$

- Linear methods for classification: the classification rules with $f_B(x)$ being linear in x, or equivalently, classification rules with linear decision boundaries.

Multi-Class Problems

- In multi-class problems, $y \in \{1, \dots K\}$. We want to construct classifier

$$f : \mathbb{R}^p \longrightarrow \{1, \dots, K\}$$

- The optimal rule is

$$f_B(x) = \arg\max_k \mathsf{P}(Y = k|X = x) = \arg\max_k \pi_k f_k(x)$$

  where π_k is prior probability and $f_k(x)$ is the conditional density for class k.
- Classify x to the most probable class by comparing $\mathsf{P}(Y|X = x)$.

Binary vs. Multi-Class

- We will focus on binary classifiers. Some binary classifiers can also handel multi-classes, such as discriminate analysis (LDA, QDA, NB), logistic regression, kNN and random forests. But for some others, the extension is non-trivial (SVM).

- There are some naive (although may not be optimal) ways to apply a binary classifier on a classification problem with $K > 2$ categories.
  - Train K one-vs-other classifiers
  - Train $K(K-1)/2$ pairwise classifiers

  Then we can combine the results to get a consensus prediction.

Discriminant Analysis

A Motivation

- A straightforward way to perform classification would be treating the y_i's as continuous , and fit a regression model to each class. However, this may suffer from some serious problems
- For the outcome Y, which may fall into categories $1, \ldots, K$, define a vector of indicators $(Y_1, \ldots, Y_K)$

$$Y_k = 1 \quad \text{if} \quad Y = k$$

  - Each vector $(Y_1, \ldots, Y_K)$ has a single 1.
  - The n training samples form an $n \times K$ indicator response matrix $\mathbf{Y}$, where each row is such an indicator vector.

Linear Regression and Its Problems

- Fit a linear regression model to each column of $\mathbf{Y}$ simultaneously

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \equiv \mathbf{X}\widehat{\mathbf{B}}$$

- The k'th column of the parameter matrix $\mathbf{B}$ represents the coefficients for modeling the probably of being category k
- Suppose we have a new input x, we can compute $\widehat{f}_k(x) = x^\mathsf{T}\mathbf{B}_{[\,,k]}$, and compare them to predict the class using

$$Y_{\text{pred}} = \widehat{f}(x) = \underset{k=1,\dots K}{\arg\max}\ \widehat{f}_k(x)$$

- Each $\widehat{f}_k(x)$ is thought of as an estimate of the conditional probability

$$\mathsf{E}(Y_k|X = x) = \mathsf{P}(Y = k|X = x)$$

- However, this suffered from some problems:
  - No guarantee that each $\widehat{f}_k(x) \in [0, 1]$
  - Serious masking problem for $K \geq 3$

Masking Problems

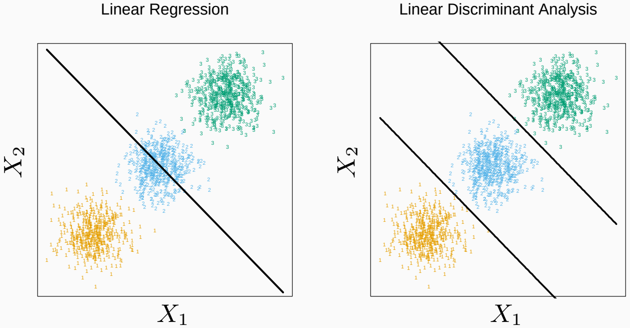


FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).*

Discriminant Analysis

- The idea is to model the distribution of X in each of the classes separately, and then use Bayes theorem to flip things around and obtain P$(Y|X = x)$.

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (LDA)
- Naive Bayes (NB)

Bayes Theorem for Classification

- As we demonstrated earlier (Bayes rules), the conditional probability can be formulated using Bayes Theorem:

$$\begin{aligned} \mathsf{P}(Y=k|X=x) &= \frac{\mathsf{P}(X=x|Y=k)\mathsf{P}(Y=k)}{\mathsf{P}(X=x)} \\ &= \frac{\mathsf{P}(X=x|Y=k)\mathsf{P}(Y=k)}{\sum_{l=1}^{K}\mathsf{P}(X=x|Y=l)\mathsf{P}(Y=l)} \\ &= \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)} \end{aligned}$$

  where $f_k(x)$ is the conditional density function of $X|Y=k$, and $\pi_k = \mathsf{P}(Y=k)$ is the prior probability.

- The best prediction is picking the one that maximizing the posterior

$$\arg\max_k \pi_k f_k(x)$$

- LDA and QDA model $f_k(x)$ as a normal distribution

Bayes Theorem for Classification

- Suppose we model each class density as multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, and assume that the covariance matrices are the same across all k, i.e., $\Sigma_k = \Sigma$. Then the

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k)\right]$$

- The log-likelihood function for the conditional distribution is

$$\begin{aligned}\log f_k(x) =& -\log((2\pi)^{p/2}|\Sigma|^{1/2}) - \frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k)\\ =& -\frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k) + \text{constant}\end{aligned}$$

- Hence we just need to select the category that attains the highest posterior density (MAP: maximum a posteriori):

$$\begin{aligned}Y_{pred} = \widehat{f}(x) &= \arg\max_k \ \log\left(\pi_k f_k(x)\right)\\ &= \arg\max_k \ -\frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k) + \log(\pi_k)\end{aligned}$$

Interpretations of LDA

- The term $(x-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma^{-1}(x-\boldsymbol{\mu}_k)$ is simply the Mahalanobis distance between x and the centroid $\boldsymbol{\mu}_k$ for class k
- Classify x to the class with the closest centroid (also adjusting the prior)
- Special case: $\Sigma = \mathbf{I}$ (only Euclidean distance is needed)

$$\arg\max_k \; -\frac{1}{2}\|x-\boldsymbol{\mu}_k\|^2 + \log(\pi_k)$$

Decision Boundary

- Noticing that that quadratic term can be simplified to

$$
\begin{aligned}
&-\frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k)\\
&= x^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_k + \text{irrelevant things}
\end{aligned}
$$

- Then the discriminant function is defined as

$$
\begin{aligned}
\delta_k(x) &= x^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_k + \log \pi_k\\
&= \mathbf{w}_k^\mathsf{T} x + b_k,
\end{aligned}
$$

- We can calculate $\mathbf{w}_k$'s and b_k's for each class k from the data.

Decision Boundary

- The decision boundary function between class k and l is

$$\begin{aligned} & \mathbf{w}_k^\mathsf{T} x + b_k = \mathbf{w}_l^\mathsf{T} x + b_l \\ \Leftrightarrow \quad & (\mathbf{w}_k - \mathbf{w}_l)^\mathsf{T} x + (b_k - b_l) = 0 \\ \Leftrightarrow \quad & \widetilde{\mathbf{w}}^\mathsf{T} x + \widetilde{b} = 0 \end{aligned}$$

- Since $\mathbf{w}_k = \Sigma^{-1} \boldsymbol{\mu}_k$ and $\mathbf{w}_l = \Sigma^{-1} \boldsymbol{\mu}_l$, the decision boundary has the directional vector

$$\widetilde{\mathbf{w}} = \Sigma^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)$$

Interpretations of LDA

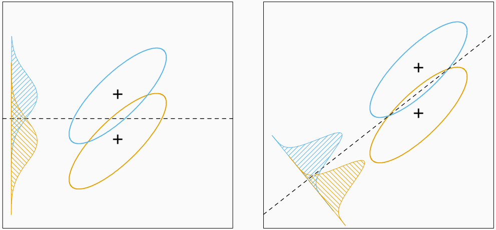


FIGURE 4.9. *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

Parameter Estimations in LDA

- We estimate the LDA parameters from the training data
  - Prior probabilities: $\widehat{\pi}_k = n_k/n = n^{-1}\sum_k \mathbf{1}\{y_i = k\}$, where n_k is the number of observations in class k.
  - Centroid: $\widehat{\boldsymbol{\mu}}_k = n_k^{-1}\sum_{i:\, y_i=k} x_i$
  - Pooled covariance:

$$\widehat{\Sigma} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:\, y_i=k}(x_i - \widehat{\boldsymbol{\mu}}_k)(x_i - \widehat{\boldsymbol{\mu}}_k)^{\mathsf{T}}$$

Quadratic Discriminant Analysis

- Quadratic Discriminant Analysis simply abandons the assumption of the common covariance matrix, i.e., the Σ_k's are not equal.
- In this case, the determinant $|\Sigma_k|$ of each covariance matrix will be different. The MAP decision becomes

$$\begin{aligned}
&\max_k \ \log\big(\pi_k f_k(x)\big) \\
=&\max_k \ -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma_k^{-1}(x-\boldsymbol{\mu}_k) + \log(\pi_k) \\
=&x^\mathsf{T}\mathbf{W}_k x + \mathbf{w}_k^\mathsf{T} x + b_k
\end{aligned}$$

- This leads to quadratic decision boundary between class k and l

$$\{x : x^\mathsf{T}(\mathbf{W}_k - \mathbf{W}_l)x + (\mathbf{w}_k^\mathsf{T} - \mathbf{w}_l^\mathsf{T})^\mathsf{T} x + (b_k - b_l) = 0\}$$

Estimations in QDA

- We estimate the QDA parameters from the training data
  - Prior probabilities: $\widehat{\pi}_k = n_k/n = n^{-1}\sum_k \mathbf{1}\{y_i = k\}$, where n_k is the number of observations in class k.
  - Centroid: $\widehat{\boldsymbol{\mu}}_k = n_k^{-1}\sum_{i:\, y_i = k} x_i$
  - Sample covariance matrix for each class:

$$\widehat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i:\, y_i = k}(x_i - \widehat{\boldsymbol{\mu}}_k)(x_i - \widehat{\boldsymbol{\mu}}_k)^{\mathsf{T}}$$

LDA vs. QDA

- More parameters in QDA than LDA, especially when p is large
- Both are extremely simple to implement
- Both LDA and QDA can perform well on real classification problems
- We can include selected quadratic terms of the covariates, such as $X_1 X_2$ or X_1^2, and still perform LDA

LDA vs. QDA

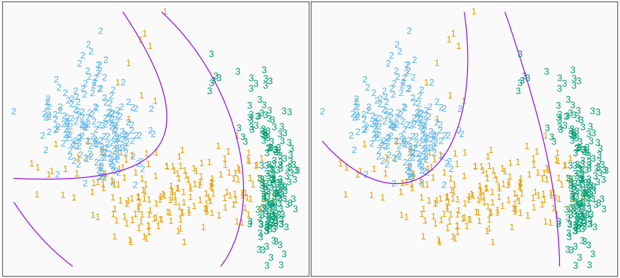


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$*). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Discriminant Analysis in Large p problems

- When p is large, QDA/LDA may not be applicable, because the inverse of $\widehat{\Sigma}$ does not exist
- Using generalized inverse matrix can easily overfit the data
- A warning sign: Classes are well-separated on the training data could be meaningless for high-dimensional data
- Regularization: sparse LDA, Naive Bayes, RDA

Sparse LDA

- Witten and Tibshirani (2011): penalized LDA

  $$\underset{a}{\text{maximize}} \; \{a^{\mathsf{T}}\mathbf{B}a + P(|a|)\} \quad \text{subject to} \quad a^{\mathsf{T}}(\mathbf{W} + \Omega)a = 1$$

  where Ω is some matrix that makes $(\mathbf{W} + \Omega)$ positive definite, and $P(|a|)$ is a penalty function over the vector $|a|$.
- Another approach Clemmensen et. al. (2011): similar idea with a different objective function that makes the optimization problem easier.

Regularized Discriminant Analysis (RDA)

- Friedman (1989): shrink the separate covariances of QDA toward a common covariance in LDA. Regularized covariance matrices are

$$\widehat{\Sigma}_k(\alpha) = \alpha\widehat{\Sigma}_k + (1-\alpha)\widehat{\Sigma}$$

- $\alpha \in [0,1]$, a continuum of models between LDA and QDA, if $\widehat{\Sigma}$ is the pooled covariance matrix used in LDA
- In practice, chose α using CV.
- We can further shrink Σ_k towards the diagonal covariance, with $\gamma \in [0,1]$

$$\widehat{\Sigma}_k(\alpha,\gamma) = \alpha\widehat{\Sigma}_k + (1-\alpha)\gamma\widehat{\Sigma} + (1-\alpha)(1-\gamma)\widehat{\sigma}^2\mathbf{I}$$

Naive Bayes

- Recall that the optimal decision rule is

$$\arg\max_k \mathsf{P}(Y = k | X = x) = \arg\max_k \pi_k f_k(x)$$

- We can approximate $f_k(x)$ by

$$f_k(x) \approx \prod_{j=1} f_{kj}(x_j),$$

  meaning that each dimension of x is approximately independently

- $f_{kj}(x_j)$ can be estimated using histograms (discrete), or kernel densities (continuous)

Logistic Regression

Motivation

- Instead of giving a classification rule that yields $\{0, 1\}$, we can directly model the probability

$$\eta(x) = \mathsf{P}(Y = 1|X = x)$$

- We have shown that the optimal rule (Bayes rule) depends on $\eta(x)$

$$f_B(x) = \begin{cases} 1 & \text{if} \quad \eta(x) > 1/2 \\ 0 & \text{if} \quad \eta(x) < 1/2. \end{cases}$$

- Directly model Y as a continues outcome does not make sense, so consider a link function g that transform $\eta(x)$ into $(-\infty, \infty)$, then

$$g(\eta(x)) = x^{\mathsf{T}}\beta$$

- Generalized linear model (GLM)

Motivation

- The response variable Y follows a Bernoulli distribution conditioning on x:

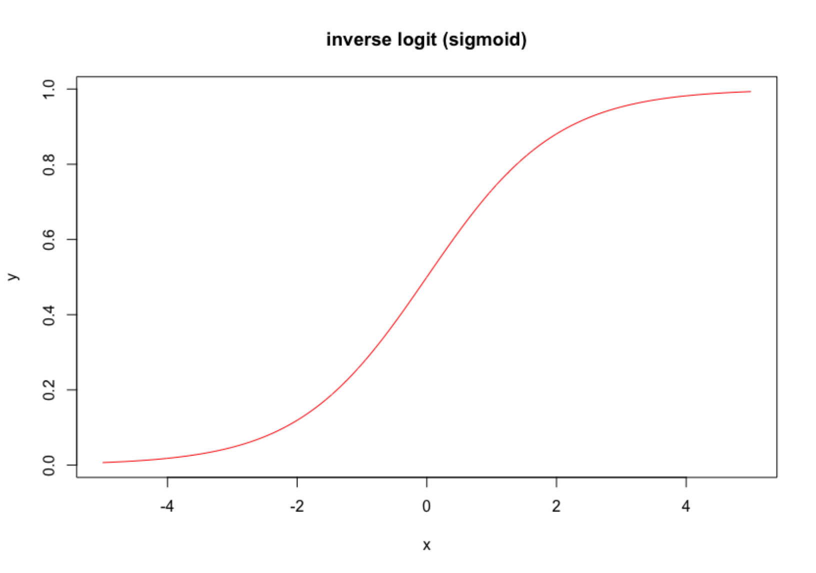
$$p(Y = y_i | X = x_i) = \eta(x_i)^{y_i}[1 - \eta(x_i)]^{1-y_i}$$

- For Logistic regression, we use the logit link function

$$\log \frac{\eta(x)}{1 - \eta(x)} = x^{\mathsf{T}}\beta, \quad \eta(x) = \frac{\exp(x^{\mathsf{T}}\beta)}{1 + \exp(x^{\mathsf{T}}\beta)}$$

  where $\log \eta(x)/(1 - \eta(x))$ is called log-odds, and we are modeling it as a linear function of x.

Inverse Logit Function



Fitting Logistic Models

- Maximize the log-likelihood function, using the conditional likelihood of Y given X:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log p(y_i | x_i, \boldsymbol{\beta})$$

- Consider $K = 2$,

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log \left\{ \eta(x_i)^{y_i} [1 - \eta(x_i)]^{1-y_i} \right\} \\ &= \sum_{i=1}^{n} y_i \log \frac{\eta(x_i)}{1 - \eta(x_i)} + \log[1 - \eta(x_i)] \\ &= \sum_{i=1}^{n} y_i x_i^{\mathsf{T}} \boldsymbol{\beta} - \log[1 + \exp(x_i^{\mathsf{T}} \boldsymbol{\beta})]\end{aligned}$$

NewtonC-Raphson

- To solve for $\widehat{\beta}$, we use Newton's method
- Choose an initial value $\boldsymbol{\beta}^0$
- Update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}}\right]^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

where

$$\text{(gradient)} \quad \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} y_i x_i^{\mathsf{T}} - \sum_{i=1}^{n} \frac{\exp(x_i^{\mathsf{T}} \boldsymbol{\beta}) x_i^{\mathsf{T}}}{1 + \exp(x_i^{\mathsf{T}} \boldsymbol{\beta})}$$

$$\text{(Hessian)} \quad \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}} = -\sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \eta(x_i)[1 - \eta(x_i)]$$

Logistic Regression vs. LDA

- For LDA, the log-posterior odds between class 1 and 0 are linear in x

$$\log \frac{\mathsf{P}(Y=1|X=x)}{\mathsf{P}(Y=0|X=x)} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}\boldsymbol{\mu}_1^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_0 + x^\mathsf{T}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \alpha_0 + x^\mathsf{T}\boldsymbol{\alpha}$$

- Logistic model has linear logics by construction

$$\log \frac{\mathsf{P}(Y=1|X=x)}{\mathsf{P}(Y=0|X=x)} = \beta_0 + x^\mathsf{T}\boldsymbol{\beta}$$

- Are they the same estimators?

Logistic Regression vs. LDA

- For LDA, the The linearity is a consequence of the Gaussian assumption for the class densities, and the assumption of a common covariance matrix.
- For logistic regression, the linearity comes by construction.
- The difference lies in how the coefficients are estimated.
- Which is more general?
  - LDA assumes Gaussian distribution of X; while logistic leaves the density of X arbitrary
- Logistic model is more general

R Functions

- LDA and QDA: R package MASS, functions lda, qda.
- Logistic: R function glm
- General optimization: R function optim