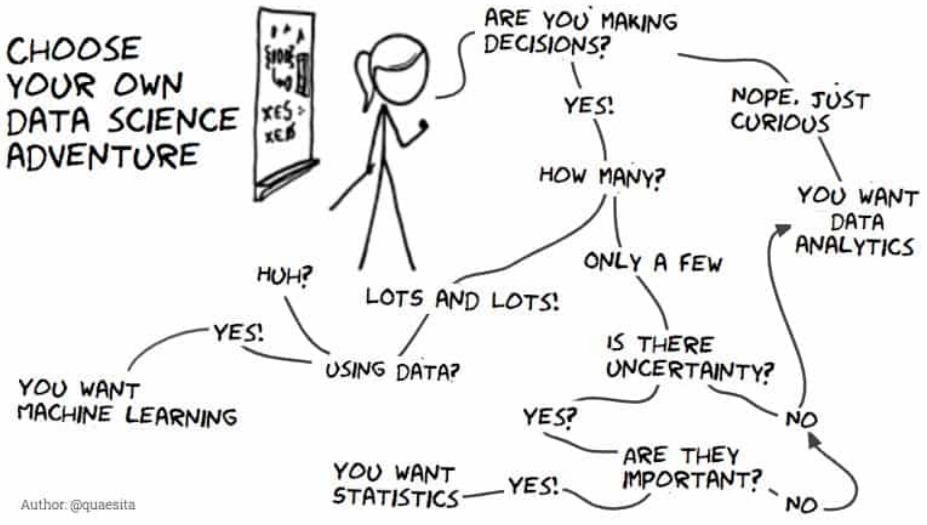
BIOS 835: Linear Classification Methods

Alexander McLain

September 21, 2023


CHOOSE YOUR OWN DATA SCIENCE ADVENTURE
ARE YOU MAKING DECISIONS?
YES!
NOPE. JUST CURIOUS
HOW MANY?
YOU WANT DATA ANALYTICS
ONLY A FEW
HUH?
LOTS AND LOTS!
YES!
USING DATA?
IS THERE UNCERTAINTY?
YOU WANT MACHINE LEARNING
YES?
NO
ARE THEY IMPORTANT?
YOU WANT STATISTICS
YES!
NO
Author: @quaesita

Introduction and notation

- So far we have focused on continuous methods that can be modeled via linear regression.
- We'll now turn from the problem of regression to that of classification.
- Assume that all data are a member of "class" k for some $k = 1, 2, \ldots, K$.
- For the moment, we'll focus on linear classifiers (this doesn't mean they have to be linear in x though).
  - What doesn't this include?

Linear Regression

- A simple classification procedure is to perform linear regression on an indicator matrix.
- Let $G = k$ indicate membership in group k where $G \in \mathcal{G}$.
- Suppose $\mathcal{G}$ has K classes, and $Y = (Y_1, Y_2, \ldots, Y_K)$ where

$$Y_k = \begin{cases} 1 & \text{if } G = k \\ 0 & \text{Otherwise} \end{cases}$$

- Then if $\boldsymbol{Y}$ is an $N \times K$ matrix of group indicators we can use

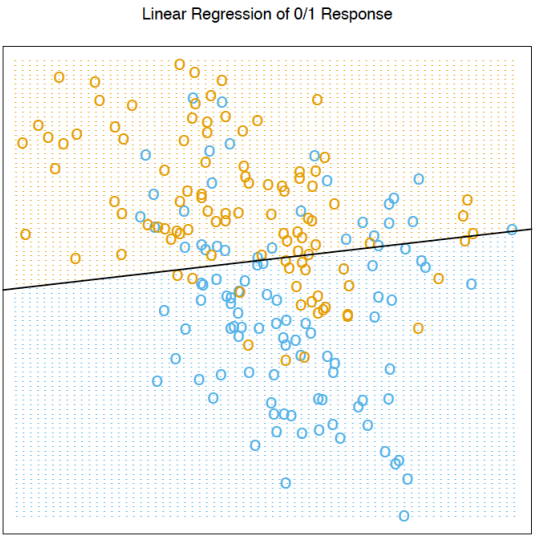
$$\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

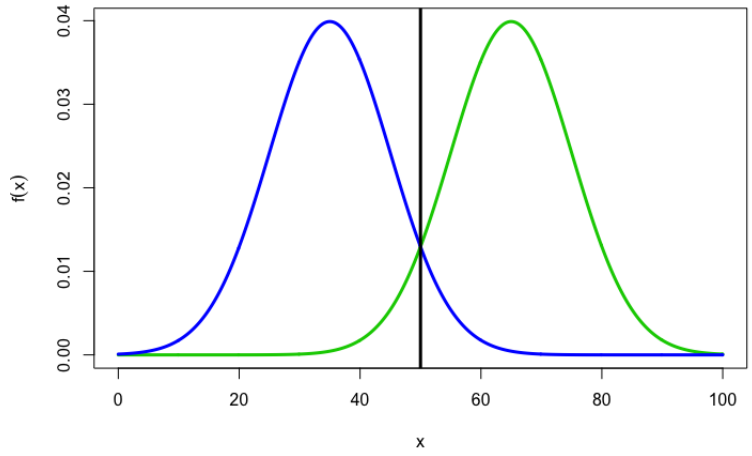
to predict group membership.

Linear Regression

- This yields the $(p+1) \times K$ matrix of coefficients $\hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.
- Classification can then be performed as follows
  1. compute $\hat{f}(x)' = (1, x')\hat{\boldsymbol{B}}$, a K vector
  2. find the largest value
$$\hat{G}(x) = \text{argmax}_{k \in \mathcal{G}} \hat{f}_k(x)$$
- Note that $E(Y_k|X = x) = \Pr(G = k|X = x)$.
- Problems can occur due to *masking* when $K \geq 3$.

Example




f(x)
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
100
x

Classification and Misclassification

- For the moment assume we have $K = 2$ classes with

$$X \sim \delta_1 f_1(x) + (1 - \delta_1) f_2(x)$$

  where $\delta_k = I(G = k)$ and $\pi_k = \Pr(\delta_k)$.
- This is the $2-$component mixture model, but can be generalized to a $K-$component mixture model

$$X \sim \sum_{k=1}^{K} \delta_k f_k(x)$$

Classification and Misclassification

- Let $\pi_k = \Pr(\delta_k)$ and

$$\Pr(\hat{G} = 2|G = 1) = \Pr(2|1) = \int_{R_2} f_1(x)dx,$$

  where R_2 is the region of x where we would classify an individual to group 2.
- Similarly let $\Pr(1|2) = \Pr(\hat{G} = 1|G = 2)$.
- All errors may not be created equal. Let C_{12} be the cost of classifying a "2" as a "1" (similar for C_{21}).

- Then the Expected cost of misclassification (ECM) is

$$\begin{aligned} ECM = E(C_{ij}) &= \sum_{j=1}^{2} \pi_j \sum_{i=1}^{2} E(C_{ij}|G=j) \\ &= \sum_{j=1}^{2} \pi_j \sum_{i=1}^{2} P(x \in R_i|G=j) C_{ij} \\ &= C_{21} \Pr(2|1)\pi_1 + C_{12} \Pr(1|2)\pi_2 \\ &= C_{21}\pi_1 \int_{R_2} f_1(x)dx + C_{12}\pi_2 \int_{R_1} f_2(x)dx \end{aligned}$$

- To minimize the ECM R_1 is chosen to be

$$R_1 = \left\{ x; \frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} \geq \frac{C_{12}}{C_{21}} \right\}$$

- Note: these regions depend on the ratios of:
  - the densities;
  - the costs of misclassification;
  - the prior probabilities.

- If $\frac{C_{12}}{C_{21}} = 1$ we can write this as

$$R_1 = \left\{ x; \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} \geq 1 \right\}$$

where

$$\begin{aligned} \Pr(G = k|X = x) &= \frac{\Pr(G = k)\Pr(X = x|G = k)}{\Pr(X = x)} \\ &= \frac{\pi_k f_k(x)}{\Pr(X = x)} \\ &= \frac{\pi_k f_k(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)} \propto \pi_k f_k(x) \end{aligned}$$

- Posterior mode: allocate to the population with the higher posterior probability

R_1 for MVN data

- Assume for group k that $X \sim MVN(\mu_k, \Sigma_k)$ for $k = 1, 2, \ldots, K$
- If $\Sigma_k = \Sigma$ for all $k = 1, 2, \ldots, K$ we have the following simplification

$$\begin{aligned}\log \frac{\Pr(G = k|X = x)}{\Pr(G = \ell|X = x)} = \log \frac{\pi_k f_k(x)}{\pi_\ell f_\ell(x)} &= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)'\Sigma^{-1}(\mu_k - \mu_\ell) \\ &\quad + x'\Sigma^{-1}(\mu_k - \mu_\ell) \\ &= b_0 + \boldsymbol{b}_1' x\end{aligned}$$

where $b_0 = \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)'\Sigma^{-1}(\mu_k - \mu_\ell)$ and $\boldsymbol{b}_1 = \Sigma^{-1}(\mu_k - \mu_\ell)$.

Linear Discriminant Analysis

- Linear Discriminant Analysis (LDA) is a procedure that uses the prior results to estimate the *linear discriminant function*

$$\delta_k(x) \quad = \quad x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log \pi_k$$

  with the classifier $G(x) = \text{argmax}_k \delta_k(x)$.
- Note that $\frac{\delta_k(x)}{\delta_\ell(x)} = b_0 + \boldsymbol{b}_1' x$ from the previous slide.

Estimating the parameters

- Estimating the parameters for LDA can be accomplished by using:
  1. $\hat{\pi}_k = \frac{N_k}{N}$ where N_k are the number of observations in group k and $N = \sum_{k=1}^{K} N_k$,
  2. $\hat{\mu}_k = \frac{1}{N_k} \sum_{i:g_i=k} x_i$, and
  3. $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i:g_i=k} \frac{(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)'}{N-K}$.
- An equivalence between LDA and linear regression can be shown (in some cases), thus LDA doesn't depend on the normality assumption as much as it seems.
- The intercept, i.e., b_0 above which gives the "cutpoint", does depend on normality. So we may want to choose it based on CV.

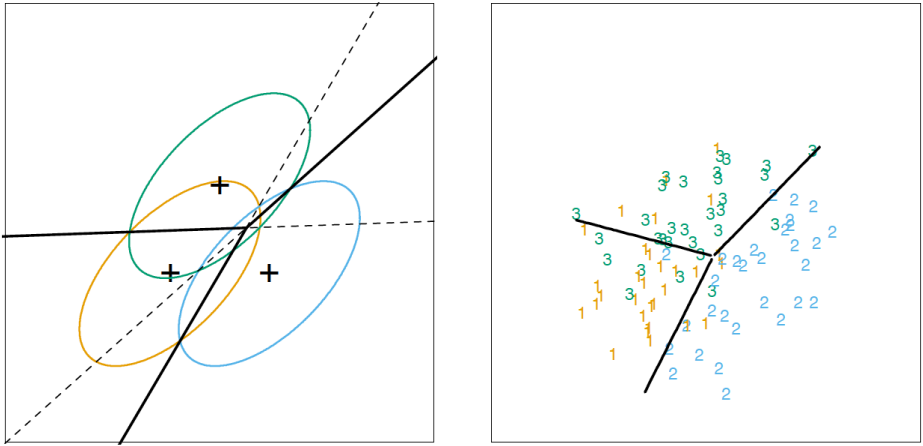


Figure: From ESL (online version).

Quadratic Discriminant Analysis (QDA)

- Not making the assumption that $\Sigma_k = \Sigma$ for all $k = 1, 2, \ldots, K$ leads to the following **quadratic discriminant function**

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k) + \log\pi_k$$

with the classifier $G(x) = \text{argmax}_k \delta_k(x)$.

Quadratic Discriminant Analysis (QDA)

- The decision boundary between classes k and ℓ is $\left\{x : \frac{\delta_k(x)}{\delta_\ell(x)} = 1\right\}$ where

$$\frac{\delta_k(x)}{\delta_\ell(x)} = \log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_k'\Sigma_1^{-1} - \mu_\ell'\Sigma_2^{-1})x - K$$

  and $K = \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2}(\mu_k'\Sigma_1^{-1}\mu_1 - \mu_\ell'\Sigma_2^{-1}\mu_2)$

- Note that

$$\frac{\delta_k(x)}{\delta_\ell(x)} = b_{k\ell 0} + \boldsymbol{b}_{k\ell 1}'x + x'\Omega_{k\ell}x$$

Quadratic Discriminant Analysis (QDA)

- QDA can be more sensitive to non-normality (figure on board).
- LDA estimates $(K-1) \times (p+1)$ parameters, QDA estimates $(K-1) \times p(p+3)/2 + 1$.
- A compromise between LDA and QDA was proposed by Friedman (1989) which used

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

other have advocated

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2\boldsymbol{I}.$$

These modifications can be put combined to result in $\hat{\Sigma}_k(\alpha, \gamma)$.

Computations for LDA

- LDA maximizes (over k) $\delta_k(x)$, which is equivalent to minimizing (over k)

$$\frac{1}{2}(x-\hat{\mu}_k)'\hat{\Sigma}^{-1}(x-\hat{\mu}_k) - \log\hat{\pi}_k$$

- Computationally it is helpful to use the eigendecomposition

$$\hat{\Sigma} = UDU'$$

where U is $p \times p$ and orthonormal and D is diagonal. Then

$$(x-\hat{\mu}_k)'\hat{\Sigma}^{-1}(x-\hat{\mu}_k) = ||D^{-1/2}U'x - D^{-1/2}U'\hat{\mu}_k||_2^2$$

which is the squared distance between $\tilde{x} = D^{-1/2}U'x$ and $\tilde{\mu}_k = D^{-1/2}U'\hat{\mu}_k$

Computations for LDA

A complete algorithm for LDA is then

1. Estimate $\hat{\Sigma}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\pi}}$.
2. Compute $\hat{\Sigma} = UDU'$.
3. Transform the class means (centroids) $\tilde{\mu}_k = D^{-1/2}U'\hat{\mu}_k$
4. Classify a point x by finding the $\tilde{\mu}_k$ closest to $\tilde{x} = D^{-1/2}U'x$.

Note that the $\tilde{x}$ have been "standardized" or put through "sphering"

$$Cov(\tilde{x}) = Cov(D^{-1/2}U'x) = D^{-1/2}U'\hat{\Sigma}UD^{-1/2} = \boldsymbol{I}.$$

This can be done with QDA as well.

Dimension reduction for LDA

- The subspace spanned by the K centroids is at most of rank $K-1$, denoted by H_{K-1}.
- x can be projected onto H_{K-1} **without affecting the classification result.**
- This yields a fundamental dimension reduction in LDA, namely that we only need to consider the data in a subspace of dimension at most $K-1$.
- If $K=3$, this could allow us to view the data in the two-dimensional plot
- If $K>3$ we might want to find a subspace H_L, where $L<K-1$ and optimal for LDA in some senses.

Fisher's problem

- Let $\boldsymbol{W} = \hat{\Sigma}$ from the LDA, $\hat{\mu} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_k$ and

$$\boldsymbol{B} = \sum_{k=1}^{K} \hat{\pi}_k (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})'.$$

- Fisher method (without Gaussian distribution assumption): Find the linear combinaton $Z = a'X$ such that between class variance ($\boldsymbol{B}$) is maximized relative to the within-class variance. That is,

$$\max_a \frac{a'\boldsymbol{B}a}{a'\boldsymbol{W}a}$$

or equivalently

$$\max_a a'\boldsymbol{B}a \text{ subject to } a'\boldsymbol{W}a = 1$$

Fisher's solution

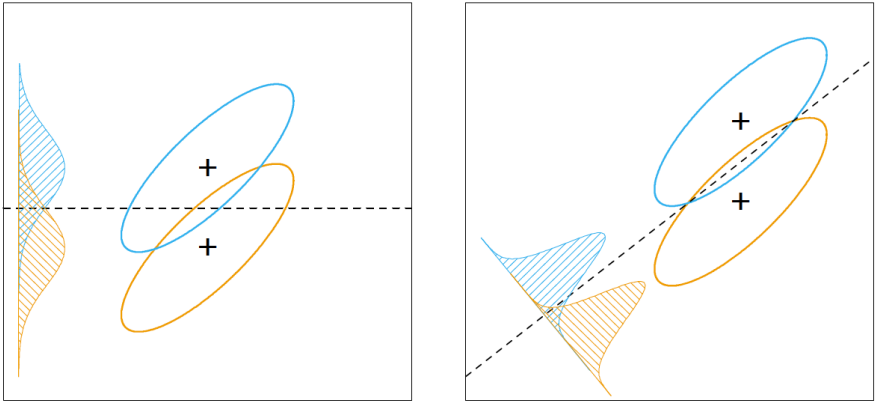


Figure: From ESL (online version).

Reduced Rank LDA

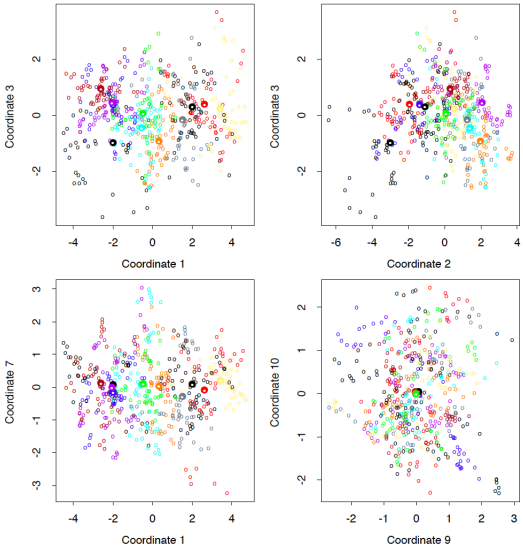
- Let $\boldsymbol{M}$ be the $K \times p$ matrix of centroids.
- compute $\boldsymbol{M}^* = \boldsymbol{M}\boldsymbol{W}^{-1/2}$ using the eigen-decomposition of $\boldsymbol{W}$.
- compute $\boldsymbol{B}^*$ the covariance matrix of $\boldsymbol{M}^*$ and it's eigen decomposition $\boldsymbol{B}^* = \boldsymbol{V}^* \boldsymbol{D}_{\boldsymbol{B}} \boldsymbol{V}^{*\prime}$.
- Let v_ℓ^* be the ℓth column of $\boldsymbol{V}^*$ and $v_\ell = \boldsymbol{W}^{-1/2} v_\ell^*$.

Then the solution to Fisher's problem is $a_1 = v_1$ and the ℓth *discriminant variable* is given by $Z_\ell = v_\ell' X$ for $\ell = 1, 2, \ldots, L < K - 1$. Smaller L means more regularization.

Vowel Data

- Example: this experiment recorded $n = 528$ instances of spoken-words. The words fall into $K = 11$ classes ("vowels"), and thereare $p = 10$ features measured on each instance.
- Hence there are $10 = K - 1$ possible dimensions for the classifier.

LDA canonical variates



RR LDA

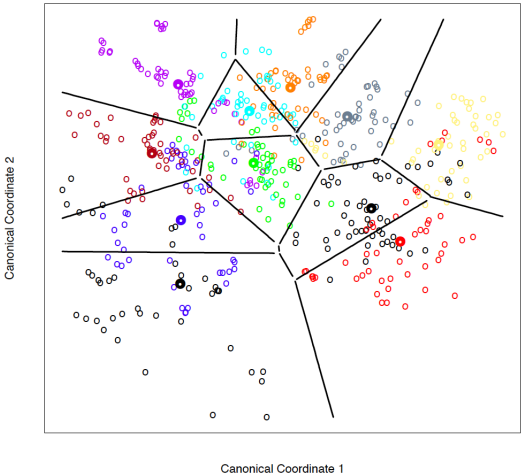


Figure: From ESL (online version).

RR LDA

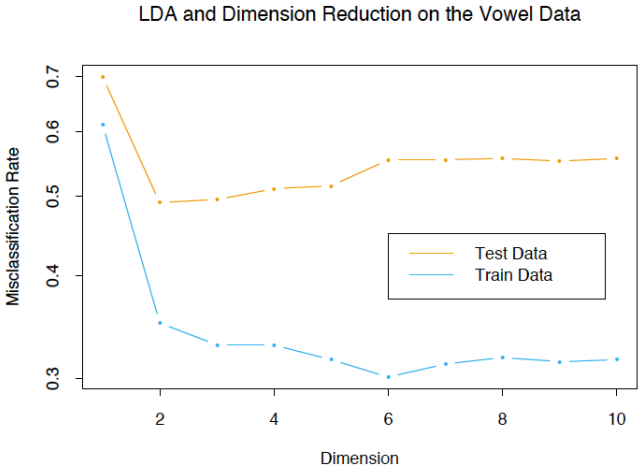


Figure: From ESL (online version).