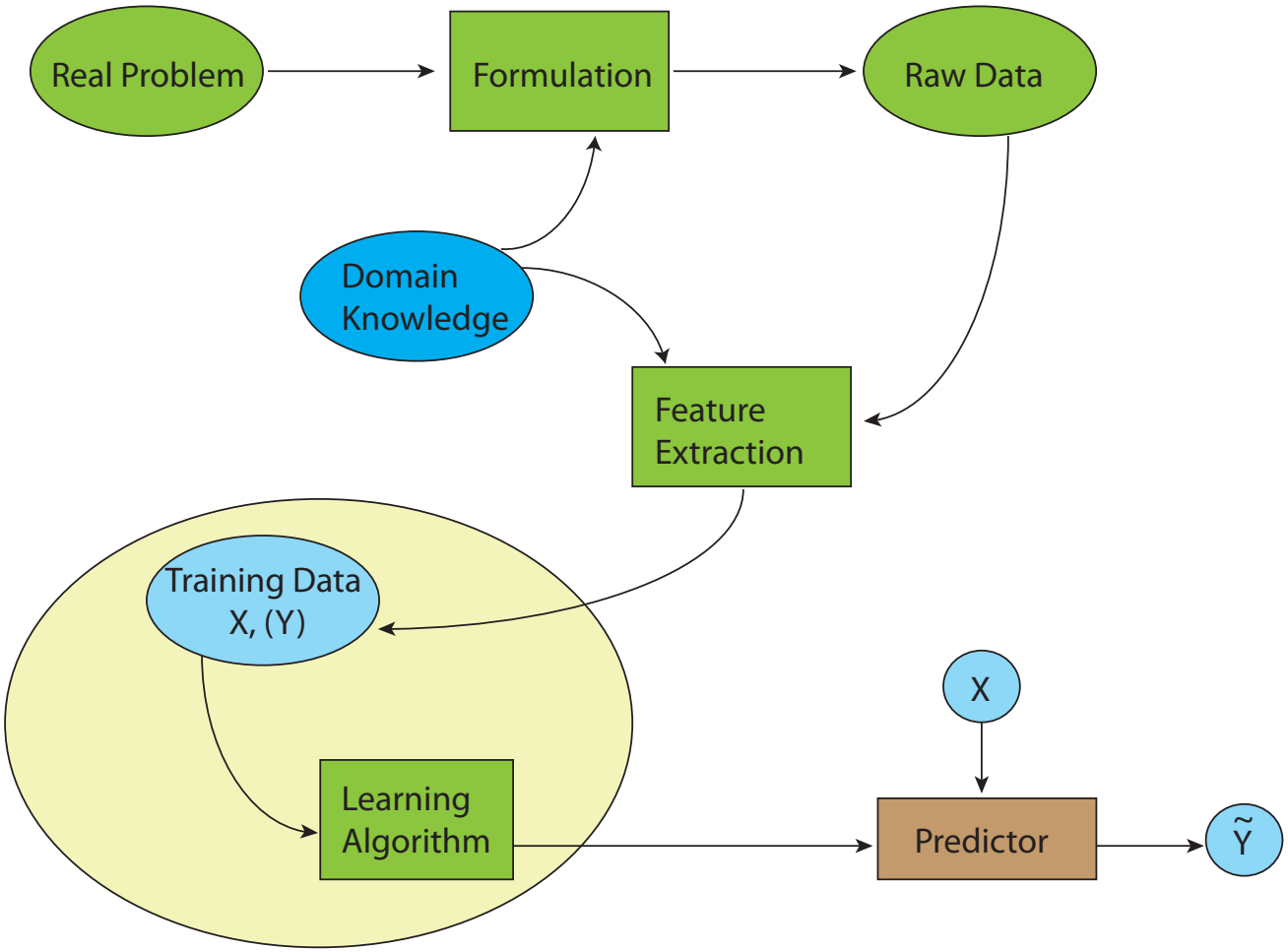
Introduction to Classification

Statistical Data Mining I

Rachael Hageman Blair

Birds Eye View: Prediction



Why Classification

- Categorical data is not meant to carry quantitative values.
- The quantitative scheme is arbitrary, but implies a distance between “classes”.

$$Y = \begin{cases} 1 & heroin \\ 2 & alcohol \\ 3 & cigarettes \\ 4 & cocaine \end{cases}$$

- Less of an issue for a 2-class problem.

Classification

- The data:
  - Training data: $\{(x_1, g_1), (x_2, g_2), \ldots, (x_N, g_N)\}$.
  - The feature vector: $X = \{X_1, X_2, \ldots, X_p\}$ where each variable is quantitative.
  - The response variable G is categorical, $G \in \{1, 2, \ldots, K\}$.

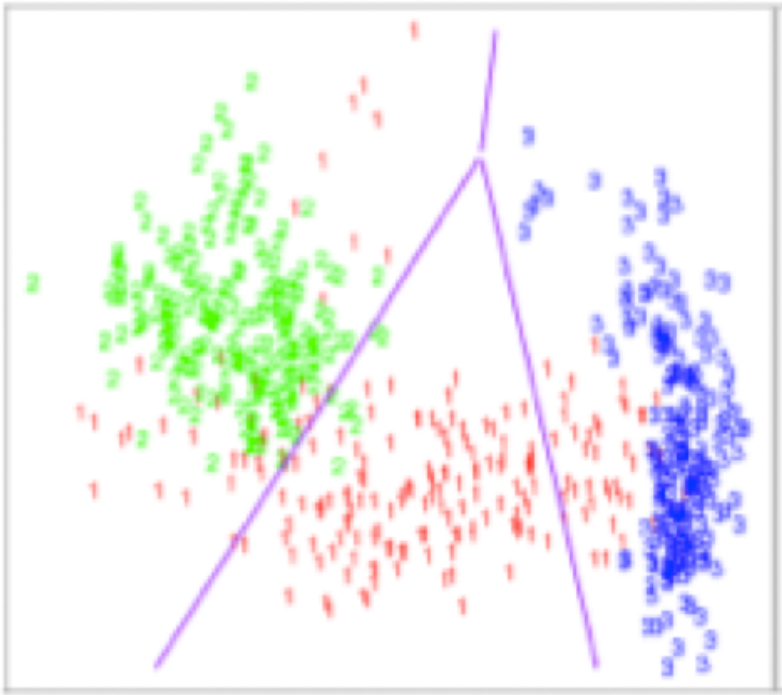
- The goal:
  - Form a predictor $G(x)$ to predict G based on X.
  - Example: G has two values: 1 denoting a useful email and 2 denoting a junk email. X is a 57-dimensional vector, each element being the relative frequency of a word or punctuation mark.
  - Goal: divides the input space (feature vector space) into a collection of $G(x)$ regions, each labeled by one class.

Two main goals:

Discrimination: Use the information in a learning set of labeled observations to construct a *classifier* (or *classification rule*) that will separate the predefined classes as much as possible.

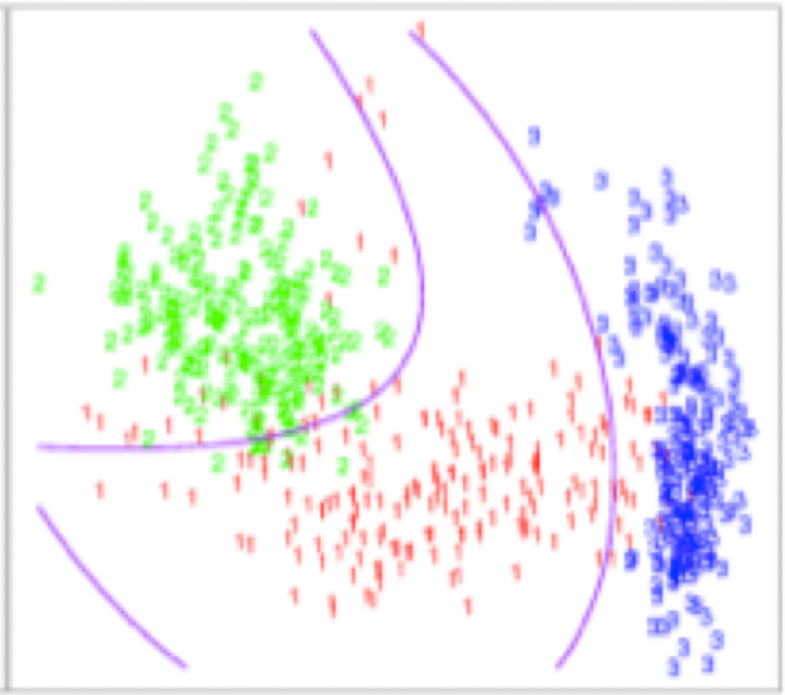
Classification: Given a set of measurements on a new *unlabeled* observation, use the classifier to predict the class of that observation.

Linear Decision Boundaries



Finding linear boundaries in the two dimensional space: X_1, X_2.

Quadratic Decision Boundaries



Finding linear boundaries in the Augmented five-dimensional space: $X_1, X_2, X_1X_2, X_1^2, X_2^2$. Linear inequalities In this space are quadratic inequalities in the original space.

Linear Methods

- In order for the decision boundaries to be linear: there must be a monotone transformation of the discriminant function δ_k for class k, OR, $\Pr\left(G = k \mid X = x\right)$ must be linear in its arguments.

Linear Methods

An Example: two-class problem (k=2):

$$\Pr(G=1 \mid X=x) = \frac{\exp\left(\beta_0 + \beta^T x\right)}{1+\exp\left(\beta_0 + \beta^T x\right)}$$

$$\Pr(G=2 \mid X=x) = \frac{1}{1+\exp\left(\beta_0 + \beta^T x\right)}.$$

Linear Methods

The decision boundary is the set of points for which the log-odds ratio:

$$\log \frac{\Pr(G=1 \mid X=x)}{\Pr(G=2 \mid X=x)} = \beta_0 + \beta^T x$$

is equal to zero. This is the hyperplane defined by:

$$\{x \mid \beta_0 + \beta^T x\}.$$

Linear Methods

The decision boundary is the set of points for which the log-odds ratio:

$$\log \frac{\Pr(G=1 \mid X=x)}{\Pr(G=2 \mid X=x)} = \beta_0 + \beta^T x$$

is equal to zero. This is the hyperplane defined by:

$$\{x \mid \beta_0 + \beta^T x\}.$$

Two popular methods:

Linear discriminant analysis

Linear logistic regression

Linear Methods

- The decision boundary between the two classes is a hyperplane in the feature vector space.
- A hyperplane in the p-dimensional input space is the set:

$$x : \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j = 0.$$

The two regions separated by the hyperplane:

$$\left\{ x : \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j > 0 \right\} \text{ and } \left\{ x : \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j < 0 \right\}.$$

Bayes Classification Rule

- For a 0-1 loss, i.e.,

$$L(g,g') = \begin{cases} 1 & g \neq g' \\ 0 & g = g \end{cases}$$

  and $E_{G|X=x} L(g,G) = 1 - \Pr(G = g \mid X = x).$

- The Bayes rule becomes the rule of maximum a posteriori probability:

$$G(x) = \arg\min_g E_{G|X=x} L(g,G)$$
$$= \arg\min_g \Pr(G = g \mid X = x).$$

- Many classification algorithms attempt to estimate

  $\Pr(G = g \mid X = x)$.

Linear Regression of an Indicator Matrix

- If G has K classes, there will be K class indicators $y_K, k = 1 \ldots K$.

g	y1	y2	y3	y4
3	0	0	1	0
1	1	0	0	0
2	0	1	0	0
4	0	0	0	1
1	1	0	0	0

Indicator matrix

- The idea is to fit a regression model for each $y_K, k = 1 \ldots K$,

$$\hat{Y} = X(X^T X)^{-1} X^T Y.$$

- We need to estimate a coefficient vector for each response column (class) $Y(:,k)$ this yields a $(p+1) \times K$ coefficient matrix.

Classification Procedure

- Define: $\hat{B} = (X^T X)^{-1} X^T Y \in \Re^{(p+1)\times K}$.
- For a new observation with input x, compute the fitted output:

$$\begin{aligned}\hat{f}(x) &= \left[(1, x)\hat{B}\right]^T \\ &= \left[(1, x_1, x_2, \ldots, x_p)\hat{B}\right]^T \\ &= \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_K(x) \end{pmatrix}.\end{aligned}$$

- Find the largest component of $\hat{f}(x)$ and classify:

$$\hat{G}(x) = \arg\max_{k \in G} \hat{f}_k(x).$$

Rationale:

- The linear regression Y_k on X is a linear approximation to $E(Y_k \mid X = x)$. Note that: $E(Y_k \mid X = x) = \Pr(G = k \mid X = x)$

- According to Bayes Rule, the optimal classifier is given as:

$$G^*(x) = \arg\max_{k \in G} \Pr(G = k \mid X = x).$$

- Linear regression of an indicator matrix:
  - Approximate $\Pr(G = k \mid X = x)$ by a linear function of x using linear regression.

The question is, how well do this do?

Rationale:

- The question: Are the $\hat{f}_k(x)$ functions good approximations of the posterior probabilities?

- Note: $\sum_{k \in G} \hat{f}_k(x) = 1$ for any x when there is an intercept in the model.

- However, $\hat{f}_k(x)$ can be negative or greater than one. This is especially true if we make predictions outside of the training set.

- These observations do not make the model invalid, or suggest it is not working.

The Phenomenon of Masking

- When the number of classes $K \geq 3$, especially when *K* is large, a class may be masked by others, that is, there is no region in the feature space that is labeled as this class.
- The linear regression model is too rigid.

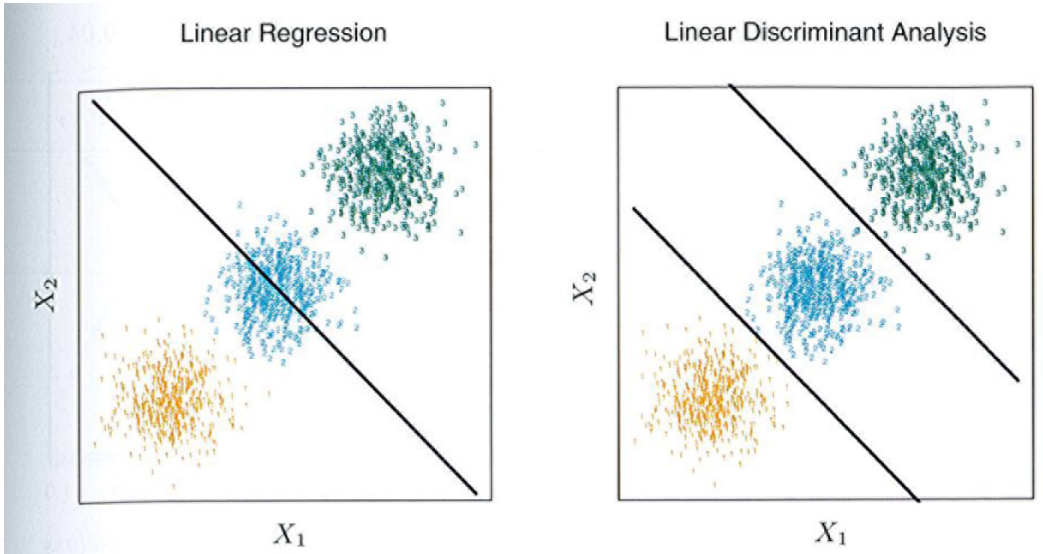
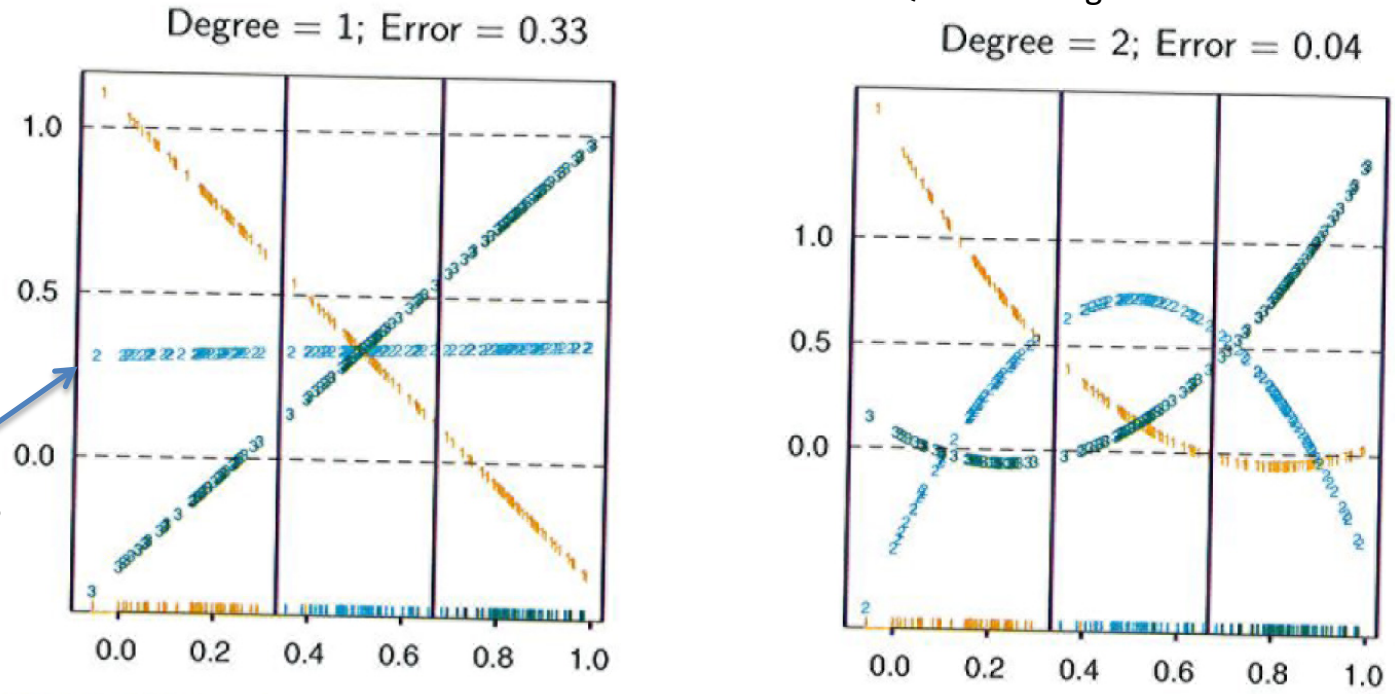


FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle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

The Phenomenon of Masking

Linear Regression

Quadratic Regression



The line corresponding to the middle class never dominates.

FIGURE 4.3. *The effects of masking on linear regression in* $\mathbb{R}$ *for a three-class problem. The* rug plot *at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, for the blue class,* y_{blue} *is* 1 *for the blue observations, and* 0 *for the green and orange. The fits are linear and quadratic polynomials. Above each plot is the training error rate. The Bayes error rate is* 0.025 *for this problem, as is the LDA error rate.*

The Phenomenon of Masking

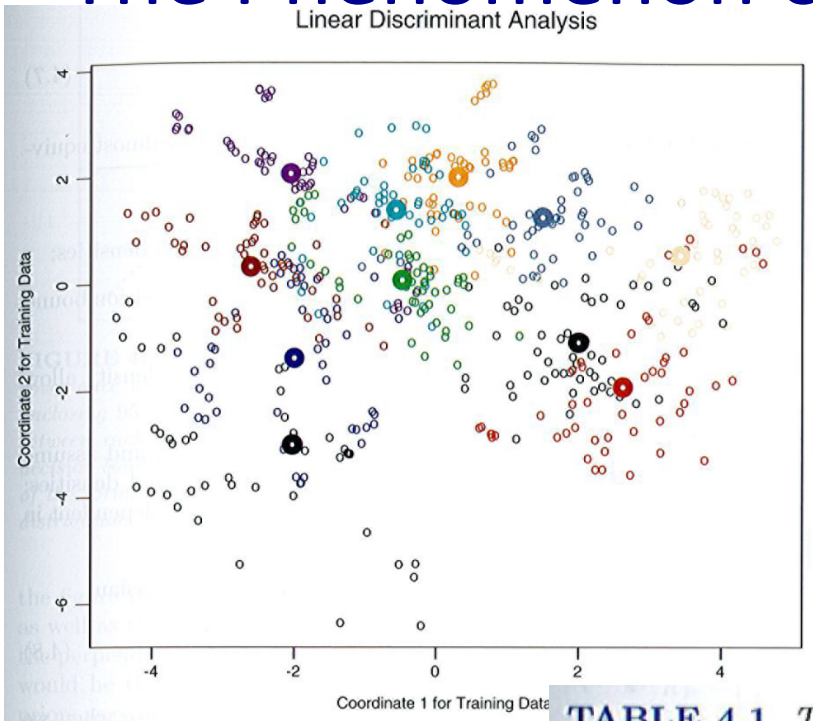


FIGURE 4.4. *A two-dimensional plot of the vowel eleven classes with $X \in \mathbb{R}^{10}$, and this is the best view (Section 4.3.3). The heavy circles are the projected me The class overlap is considerable.*

Vowel Recognition Data
11 Classes, projected into a 2-D space

TABLE 4.1. *Training and test error rates using a variety of linear techniques on the vowel data. There are eleven classes in ten dimensions, of which three account for 90% of the variance (via a principal components analysis). We see that linear regression is hurt by masking, increasing the test and training error by over 10%.*

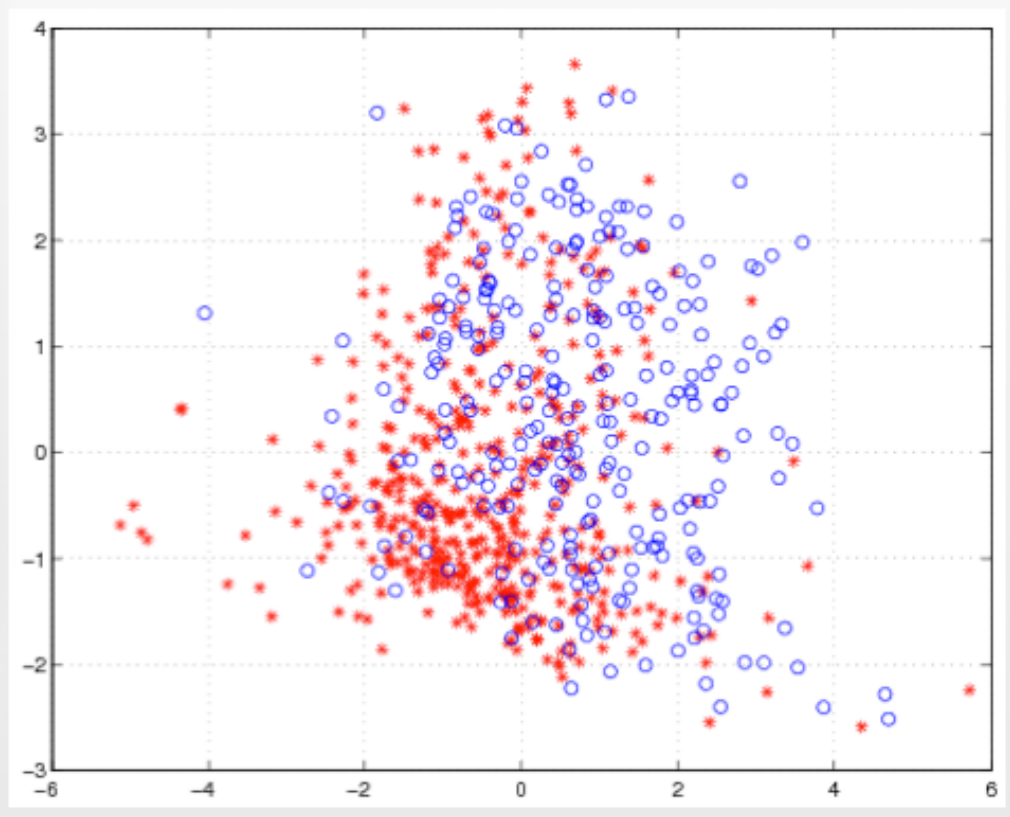
Technique	Error Rates	
	Training	Test
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

Example: Diabetes

- There are 768 cases in the dataset, of which, 268 show signs of diabetes according to the World Health Organization criteria. Each case contains 8 quantitative variables, including diastolic blood pressure, triceps skin fold thickness, body mass index, etc.
  - Two classes: with or without signs of diabetes.
  - Denote the 8 original variables by $\tilde{X}_1, \tilde{X}_2, \ldots \tilde{X}_8$,
  - Remove the mean of $\tilde{X}_j$ and normalize it to unit variance.

Example: Diabetes

- The scatterplot without diabetes (class 1: stars) and with diabetes (class 2: circles).



Example: Diabetes

- Two principal components X_1 and X_2 are used in the classification:

$$\hat{B} = (X^T X)^{-1} X^T Y$$

$$= \begin{pmatrix} 0.6510 & 0.3490 \\ -0.1256 & 0.1256 \\ -0.0729 & 0.0729 \end{pmatrix}$$

$$\hat{Y}_1 = 0.6510 - 0.1256 X_1 - 0.0729 X_2$$

$$\hat{Y}_2 = 0.3490 + 0.1256 X_1 + 0.0729 X_2.$$

Example: Diabetes

- The classification rule:

$$\hat{G}(x) = \begin{cases} 1 & \hat{Y}_1 \geq \hat{Y}_2 \\ 2 & \hat{Y}_1 < \hat{Y}_2 \end{cases}.$$

- Within training data classification error rate: 28.52%
- Sensitivity (probability of claiming positive with the truth is positive): 44.03%.
- Specificity (probability of claiming a negative with the truth is negative): 86.20%.

Example: Diabetes

- The scatterplot without diabetes (class 1: stars) and with diabetes (class 2: circles).