Linear Classification Problem

(Chapter 4 Hastie e.a.)

- Naïve linear Classifiers

Two more reasonable approaches:

- Fisher’s Linear Discriminant Analysis
- Logistic regression model
- Two class perceptron
- Optimal separation hyperplane

Linear Classifiers

Naïve Classification by linear regression

Each object X belongs to a group k out of p denoted by G=k and let Y=1
For G=k and Y=0 otherwise. Fit regression models to each group of (X,Y)'s

$$\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^{\,T} x$$

Decision boundary: $\{\hat{f}_k(x) = \hat{f}_l(x)\}$ => $\{\hat{\beta}_{k0} - \hat{\beta}_{l0} + (\hat{\beta}_k - \hat{\beta}_l)^T x = 0\}$

This is based on the idea that P(G=k)= E(Y|X=x)

Another model when we have only two groups:

$$\Pr(G=1|X=x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},$$

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

Decision Boundary $\beta_0 + \beta^T x = 0$ because $\log \frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)} = \beta_0 + \beta^T x.$

Linear classifiers can yield nonlinear separation by including nonlinear functions of the linear terms such as powers, exponentials, logs

Linear Classifiers

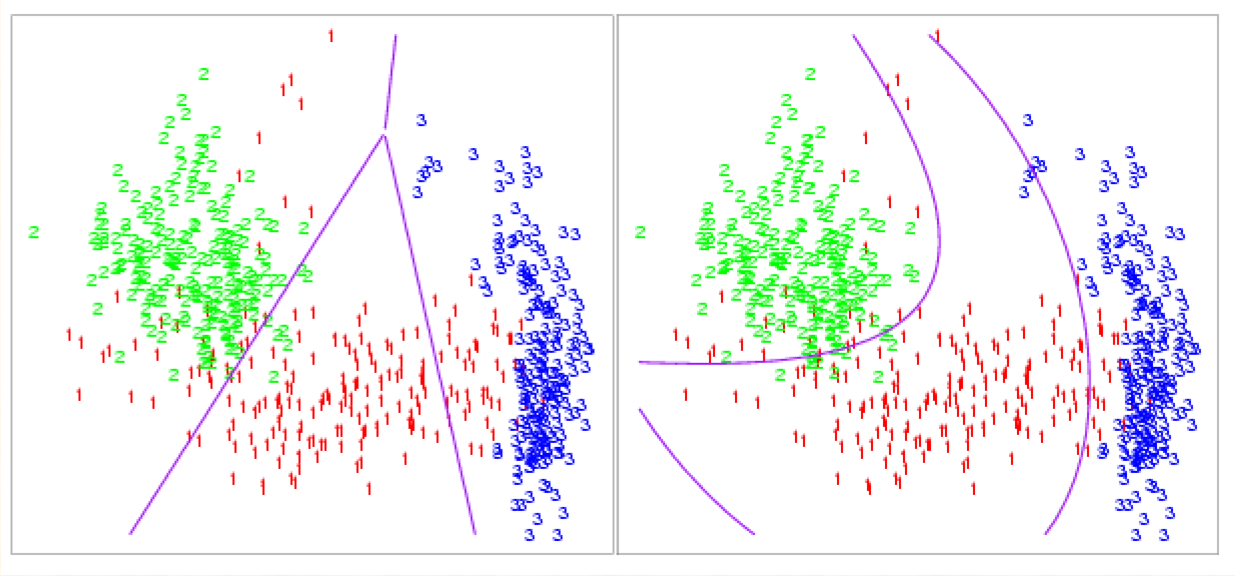


FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.*

Linear Classifiers

Fisher's discriminant function for several groups

A. All the Σ's are equal

Group 1: $Pop_1(\mu_1, \Sigma)$, … , Group k: $Pop_k(\mu_k, \Sigma)$ (notice that the Σ's are equal).

$$S_{pl} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i$$

Next we calculate the distance from y to the center of each group:

$$D_i^2(y) = (y - \bar{y}_i)' S_{pl}^{-1}(y - \bar{y}_i)$$

and assign **y** to the closet group center.

This is equivalent to using the linear discriminant function:

$$L_i(y) = \bar{y}_i' S_{pl}^{-1} y - \frac{1}{2}\bar{y}_i' S_{pl}^{-1}\bar{y}_i = a_i' y + a_{i0}$$

and assign **y** to the group with the largest $L_i(\mathbf{y})$.

Linear Classifiers

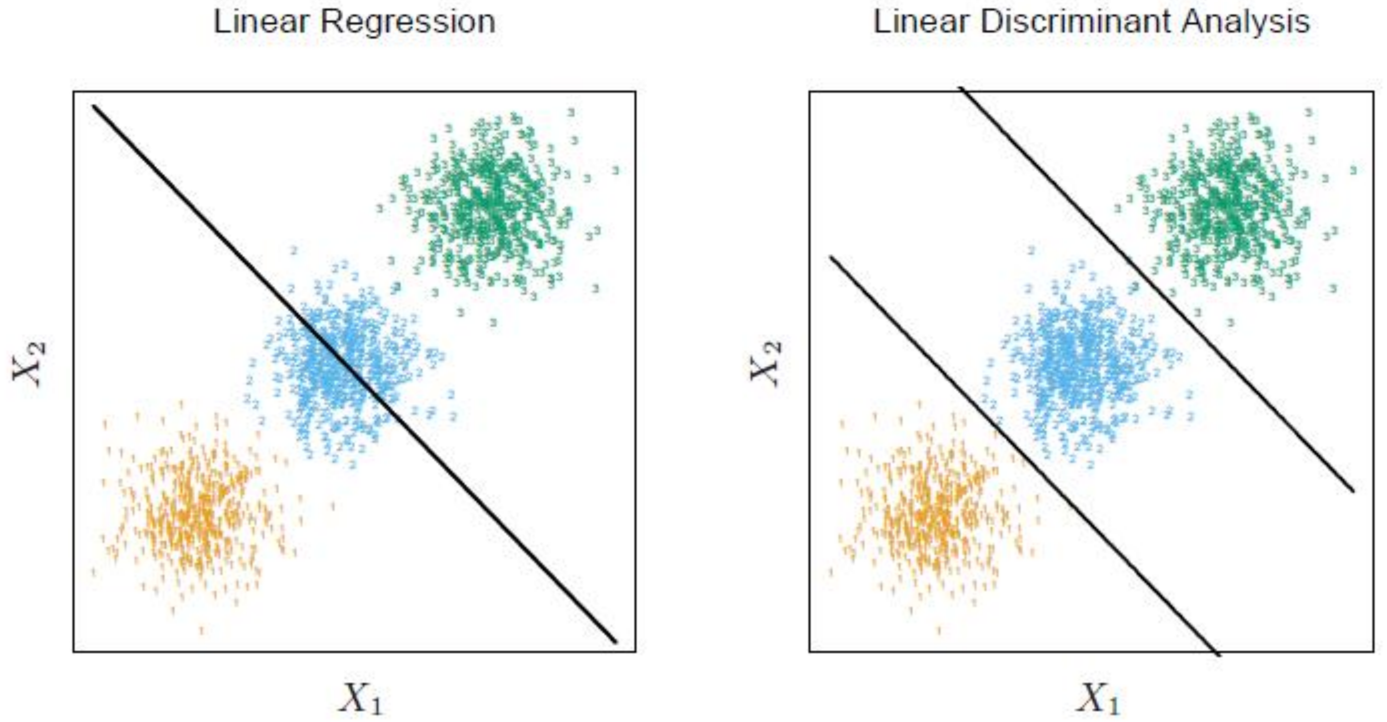


FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle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

Linear Classifiers

Bayes Rule:

Let $(\pi_1,\ldots,\pi_k)$ be the prior probabilities for the k groups.

Then $L_i(\mathbf{y})$ becomes: $\mathbf{a}_i'\mathbf{y} + a_{i0} + \ln \pi_i$

Misclassification Rate:

Build a table of y vs predicted

MSR= $$\frac{\sum_{i \neq j} n_{ij}}{\sum n_i}$$

Linear Classifiers

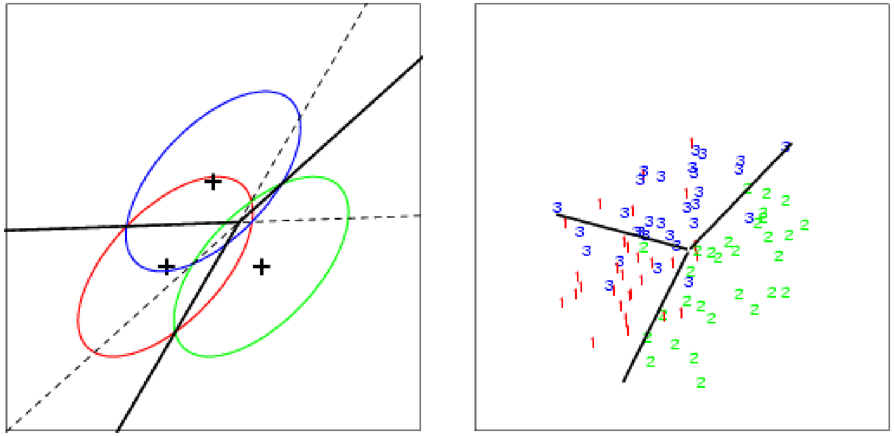


Figure 4.5: *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

Linear Vs Quadratic

Lets not assume that the covariances are equal. Then the discriminant functions are quadratic:

$$Q_i(y) = \ln \pi_i - \ln |S_i| - \frac{1}{2}(y_i - \bar{y}_i)' S_i^{-1} (y_i - \bar{y}_i)$$

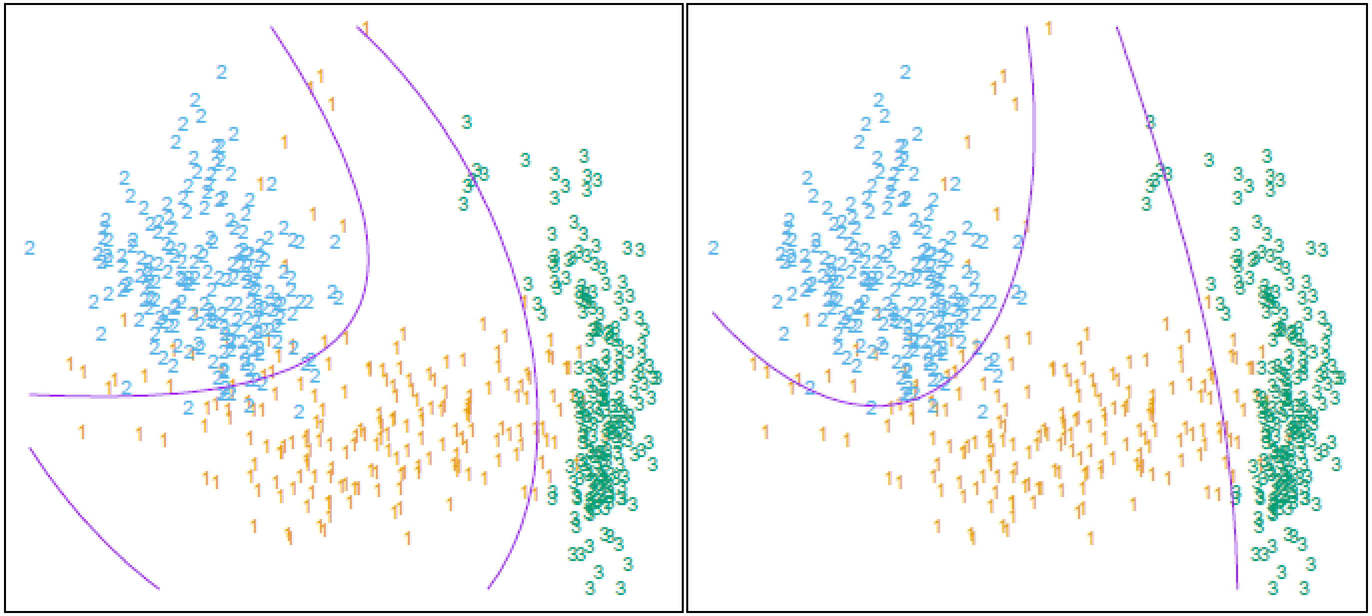


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences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

dataset crops: Example in R

```
CORN        16 27 31 33          SUGARBEETS25 25 24 26
CORN        15 23 30 30          SUGARBEETS34 25 16 52
CORN        16 27 27 26          SUGARBEETS54 23 21 54
CORN        18 20 25 23          SUGARBEETS25 43 32 15
CORN        15 15 31 32          SUGARBEETS26 54  2 54
CORN        15 32 32 15          CLOVER      12 45 32 54
CORN        12 15 16 73          CLOVER      24 58 25 34
SOYBEANS    20 23 23 25          CLOVER      87 54 61 21
SOYBEANS    24 24 25 32          CLOVER      51 31 31 16
SOYBEANS    21 25 23 24          CLOVER      96 48 54 62
SOYBEANS    27 45 24 12          CLOVER      31 31 11 11
SOYBEANS    12 13 15 42          CLOVER      56 13 13 71
SOYBEANS    22 32 31 43          CLOVER      32 13 27 32
COTTON      31 32 33 34          CLOVER      36 26 54 32
COTTON      29 24 26 28          CLOVER      53 08 06 54
COTTON      34 32 28 45          CLOVER      32 32 62 16
COTTON      26 25 23 24          ;
COTTON      53 48 75 26
```

Generalized Squared Distance to CROP

From CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CLOVER	2.37125	7.52830	4.44969	6.16665	5.07262
CORN	6.62433	3.27522	5.46798	4.31383	6.47395
COTTON	3.23741	5.15968	3.58352	5.01819	4.87908
SOYBEANS	4.95438	4.00552	5.01819	3.58352	4.65998
SUGARBEETS	3.86034	6.16564	4.87908	4.65998	3.58352

Logistic Regression

Note that LDA is linear in x:

$$\log\frac{p(c_k\mid x)}{p(c_0\mid x)} = \log\frac{p(c_k)}{p(c_0)} - \frac{1}{2}(\mu_k+\mu_0)^T\Sigma^{-1}(\mu_k-\mu_0) + x^T\Sigma^{-1}(\mu_k-\mu_0)$$

$$= \alpha_{k0} + \alpha_k^T x$$

Linear logistic regression looks the same:

$$\log\frac{p(c_k\mid x)}{p(c_0\mid x)} = \beta_{k0} + \beta_k^T x$$

But the estimation procedure for the co-efficients is different. LDA maximizes joint likelihood [y,X]; logistic regression maximizes conditional likelihood [y|X]. Usually similar predictions.

Logistic Regression MLE

For the two-class case, the likelihood is:

$$l(\beta) = \sum_{i=1}^{n} \left\{ y_i \log p(x_i;\beta) + (1 - y_i) \log(1 - p(x_i;\beta)) \right\}$$

$$\log\left(\frac{p(x;\beta)}{1 - p(x;\beta)} \right) = \beta^T x \qquad \log p(x;\beta) = \beta^T x - \log(1 + \exp(\beta^T x))$$

$$\Rightarrow l(\beta) = \sum_{i=1}^{n} \left\{ y_i \beta^T x + \log(1 + \exp(\beta^T x)) \right\}$$

The maximize need to solve (non-linear) score equations:

$$\frac{dl(\beta)}{d\beta} = \sum_{i=1}^{n} x_i (y_i - p(x_i;\beta)) = 0$$

Regularized Logistic Regression

- Ridge/LASSO logistic regression

$$\hat{w} = \arg\inf_{w} \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-w^T x_i y_i)) + \lambda w^2$$

$$\hat{w} = \arg\inf \frac{1}{n} \sum_{i=1}^{N} \log(1 + \exp(-w^T x_i y_i)) + \lambda \sum_{j} \left| w_j \right|$$

- Successful implementation with over 100,000 predictor variables
- Can also regularize discriminant analysis

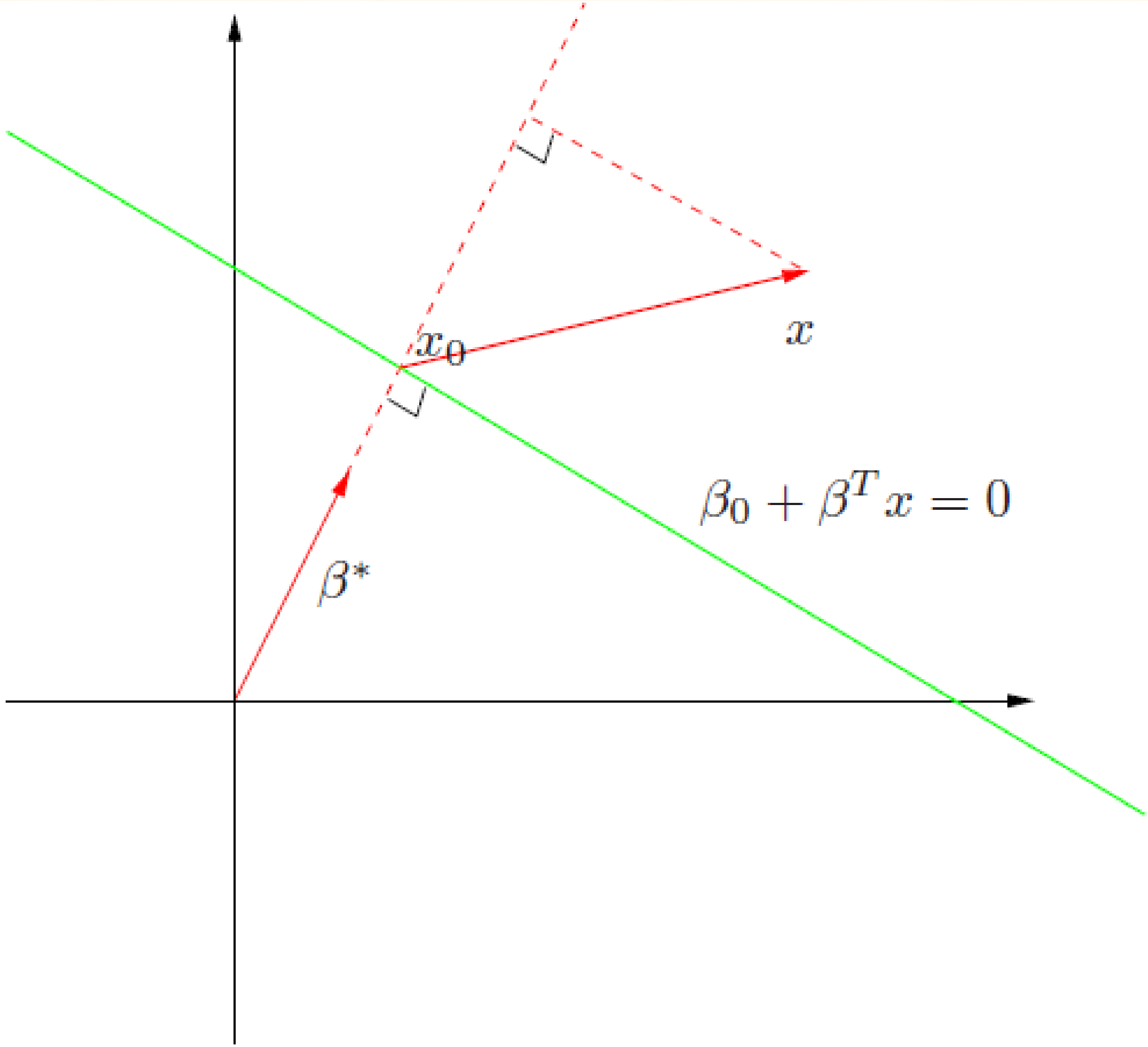


FIGURE 4.15. *The linear algebra of a hyperplane (affine set)*

Simple Two-Class Perceptron

Define: $h(x) = x^T \beta + \beta_0$ $\quad$ w = $\begin{pmatrix} \beta \\ \beta_0 \end{pmatrix}$

Classify as class y=1 if $h(x)$>0, class y=-1 otherwise.

Score function: $D(\beta, \beta_0) = \sum_{i \in I} y_i (x^T \beta + \beta_0)$ $\quad$ $\partial D / \partial w = \begin{pmatrix} -\sum_{i \in I} y_i x_i \\ -\sum_{i \in I} y_i \end{pmatrix}$

Initialize weight vector

Repeat one or more times:

For each training data point $\boldsymbol{x}_i$

If point correctly classified, do nothing

Else $\quad w \leftarrow w + \lambda \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix}$

Guaranteed to converge to a separating hyperplane (if exists)

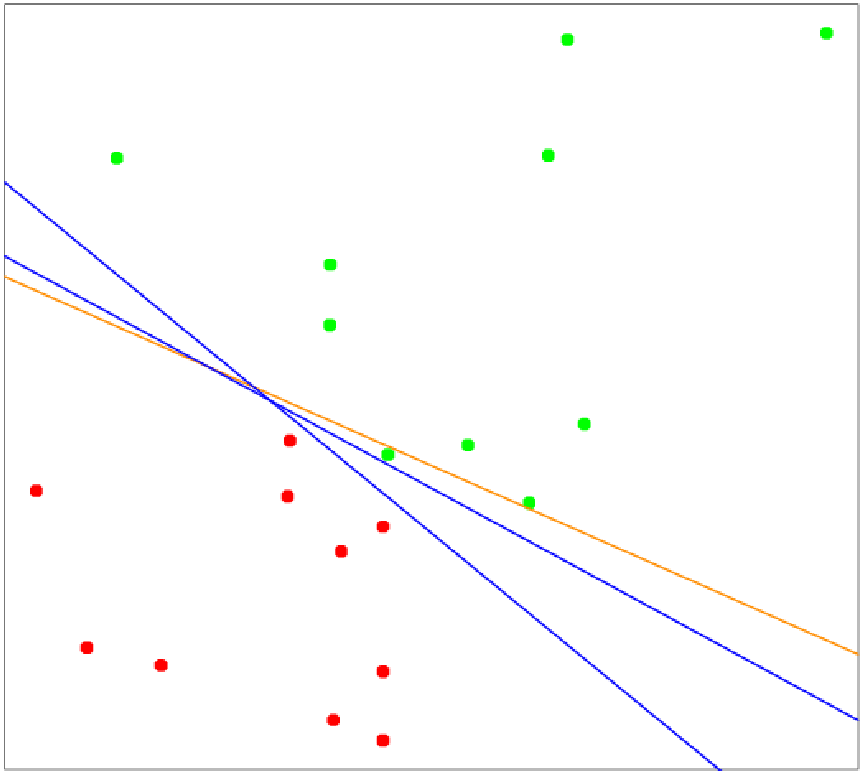


Figure 4.13: *A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the* perceptron learning algorithm *with different random starts.*

“Optimal” Hyperplane

The “optimal” hyperplane separates the two classes and maximizes the distance to the closest point from either class.

Finding this hyperplane is a convex optimization problem.

This notion plays an important role in support vector machines

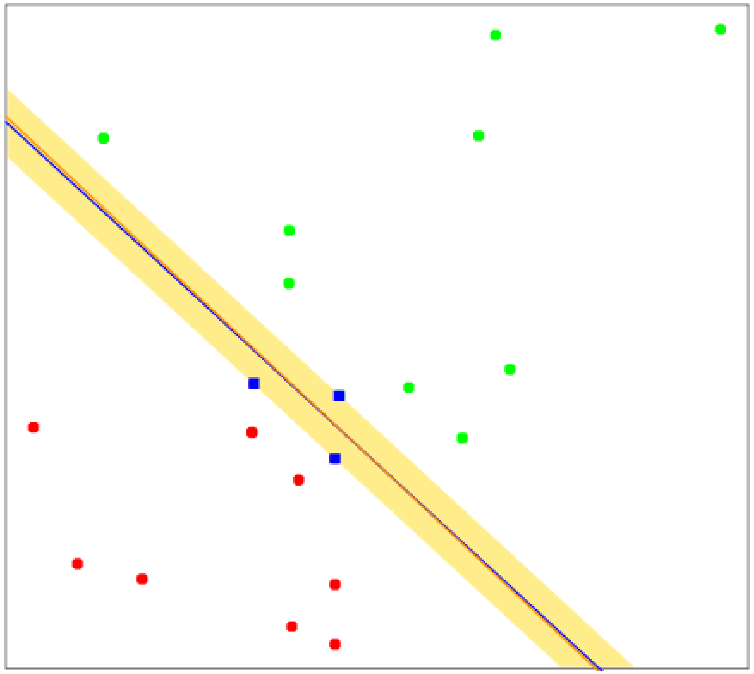
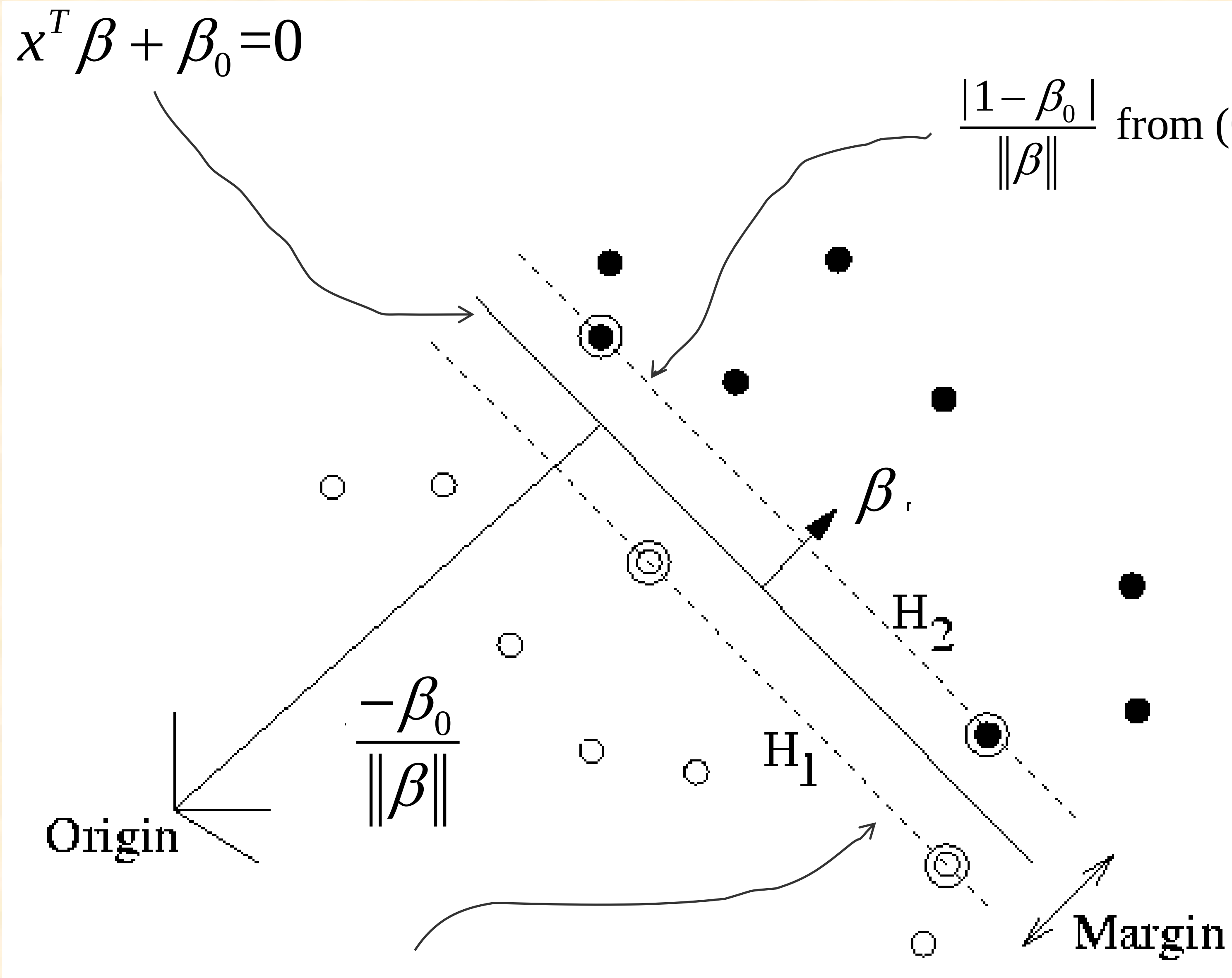


Figure 4.15: *The same data as in Figure 4.13. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Section 12.3.3).*



$\frac{|-1-\beta_0|}{\|\beta\|}$ from (0,0)

$\frac{2}{\|\beta\|}$