Experimental Data Analysis

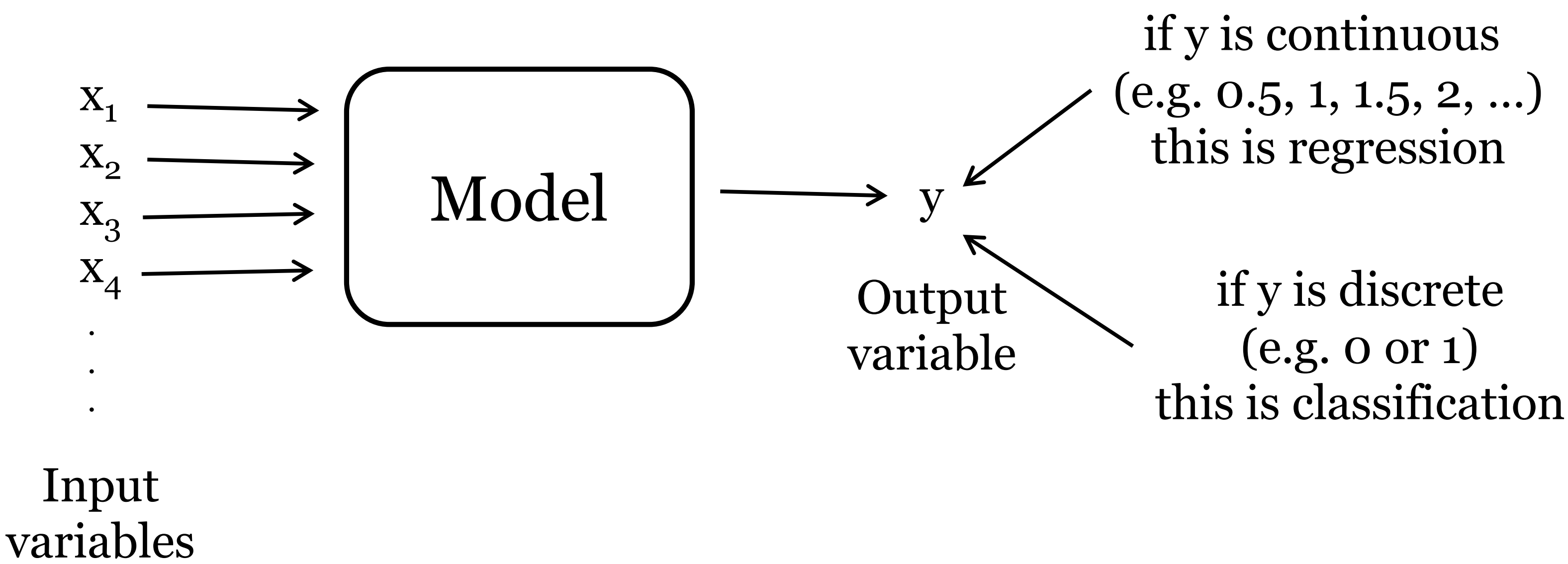
in ©MATLAB

Lecture 9:

Classification, logistic regression, linear discriminant analysis, support vector machine

Jan Rusz
Czech Technical University in Prague

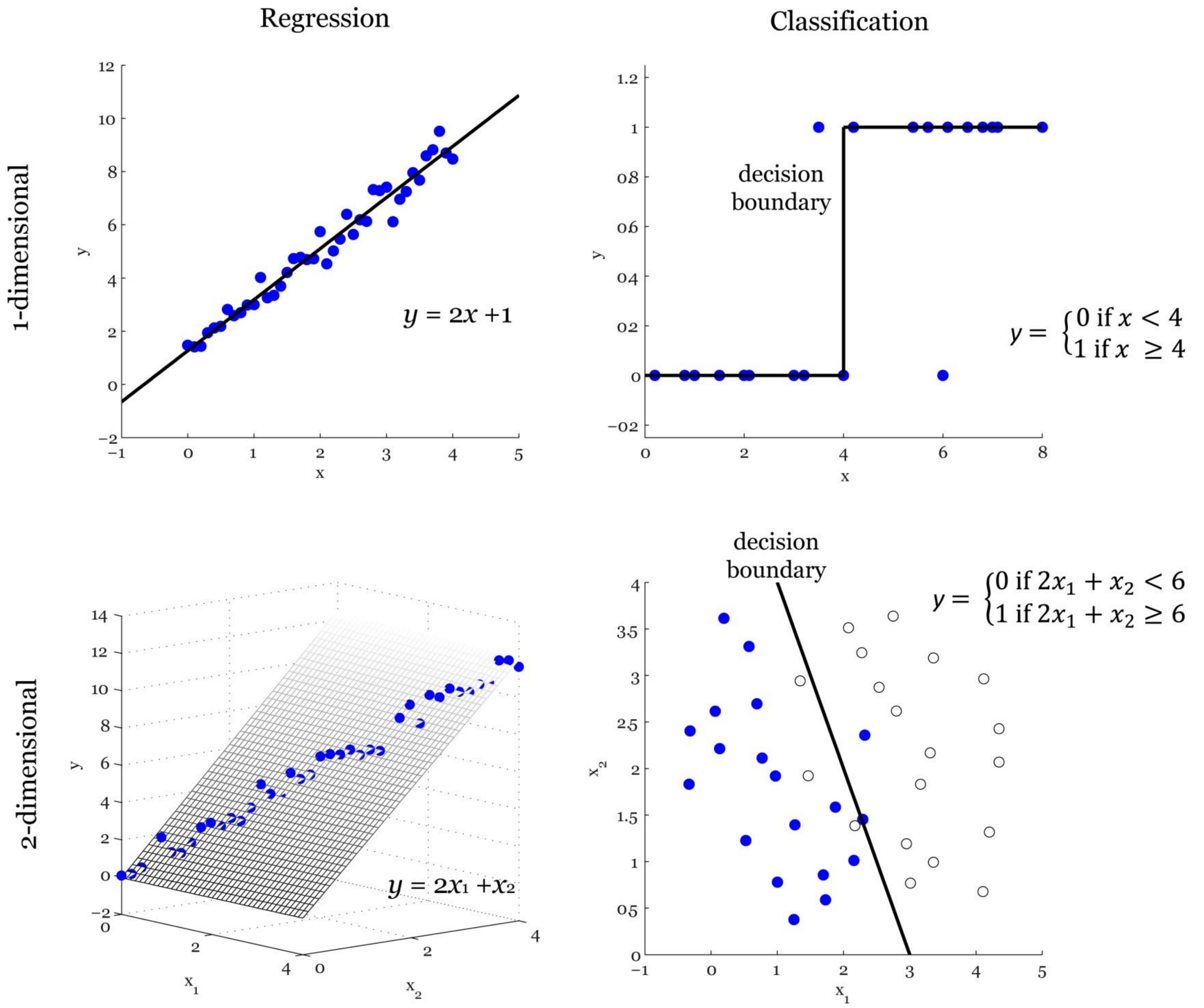
Supervised learning



Linear classification model

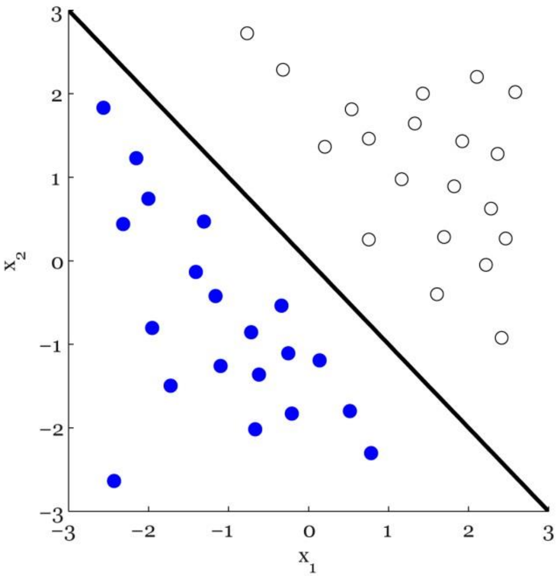
$$y = \begin{cases} 0 \text{ if } \sum_{i=1}^{n} w_i x_i < c \\ 1 \text{ if } \sum_{i=1}^{n} w_i x_i \geq c \end{cases}$$

Comparing regression and classification



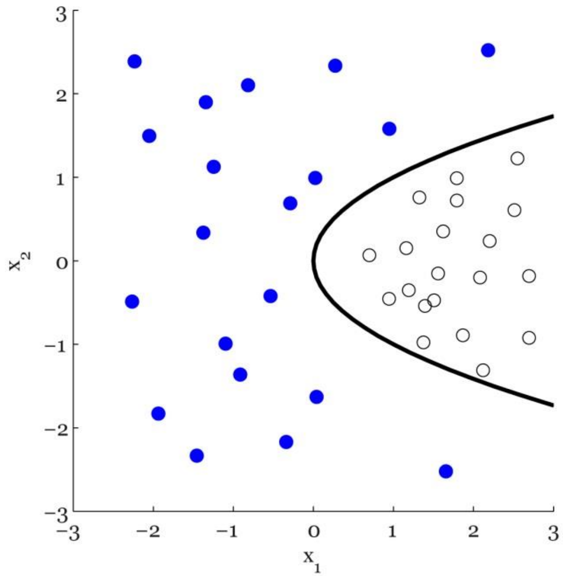
Different examples of decision boundaries

Linear



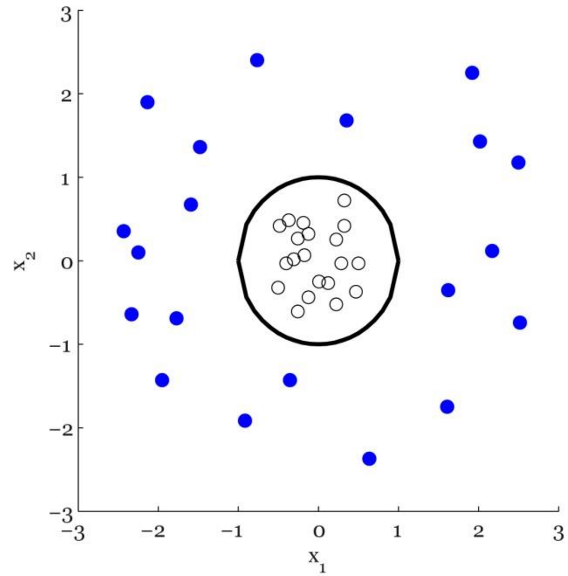
$$y = \begin{cases} 0 \text{ if } x_1 + x_2 < 0 \\ 1 \text{ if } x_1 + x_2 \geq 0 \end{cases}$$

Nonlinear



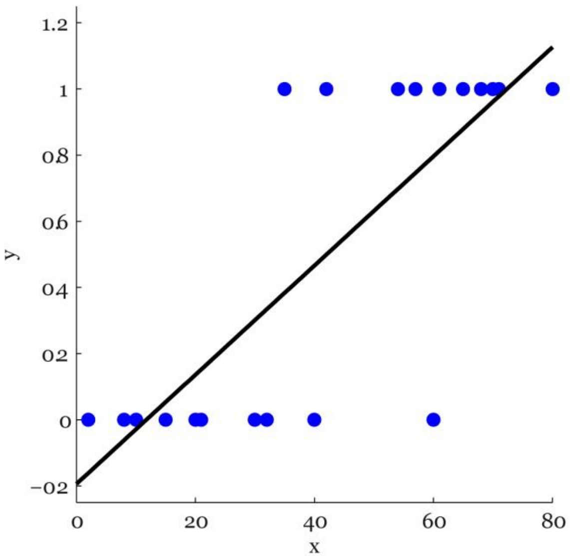
$$y = \begin{cases} 0 \text{ if } x_1 - x_2^2 < 0 \\ 1 \text{ if } x_1 - x_2^2 \geq 0 \end{cases}$$

Nonlinear

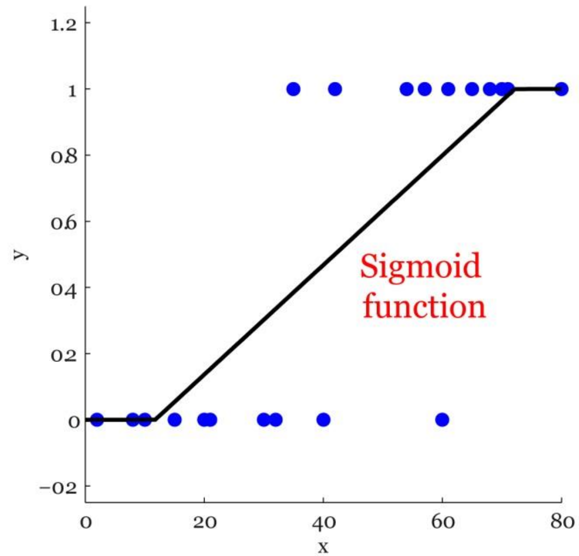


$$y = \begin{cases} 0 \text{ if } x_1^2 + x_2^2 < 1 \\ 1 \text{ if } x_1^2 + x_2^2 \geq 1 \end{cases}$$

Logistic regression

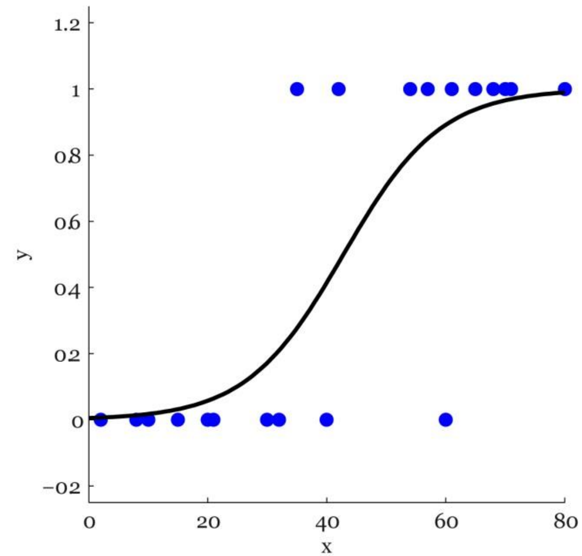


$y = w_1 x + w_2$



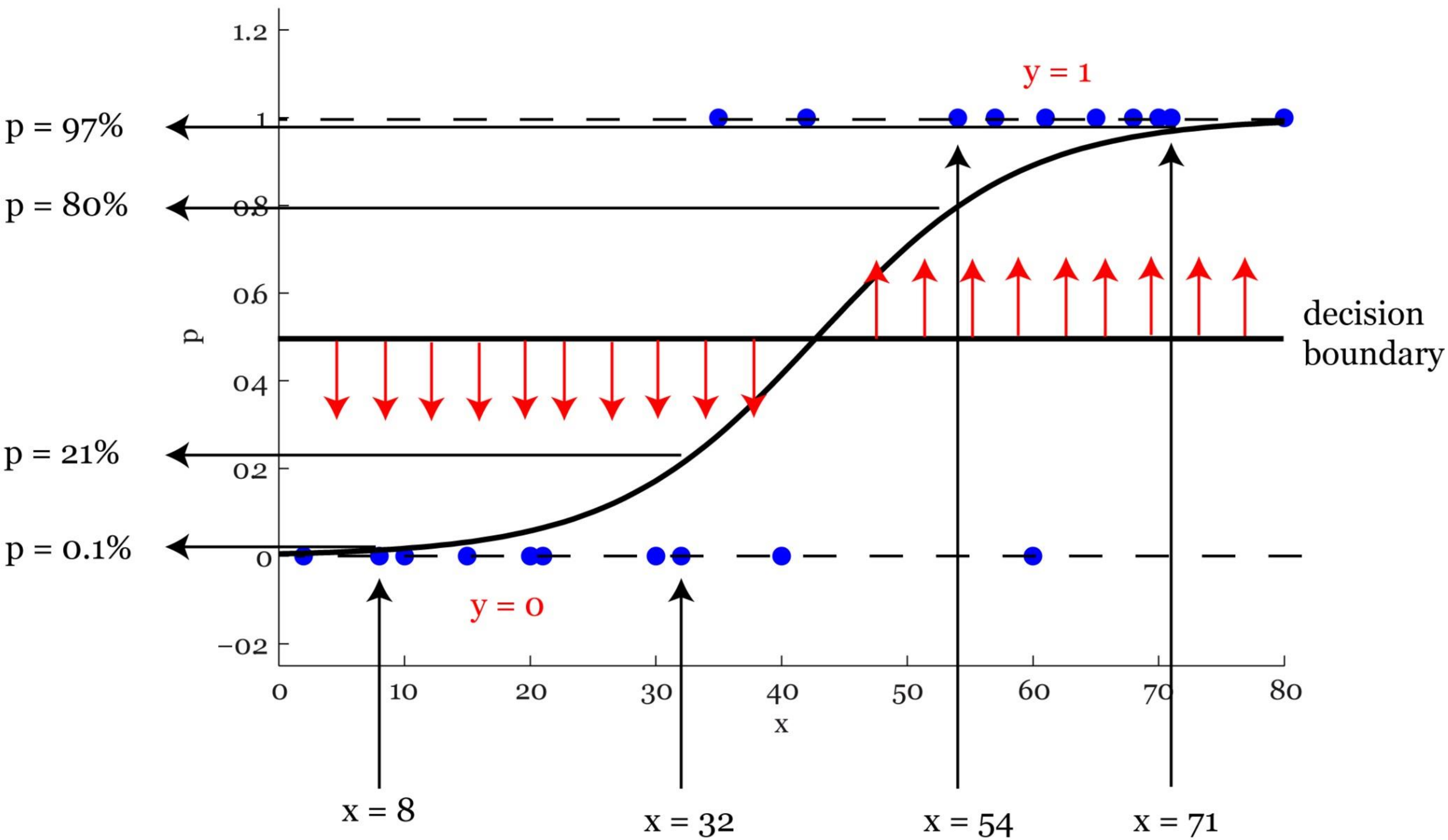
$$y = \frac{1}{1+e^{-(w_1 x + w_2)}} = \frac{1}{1+e^{-\sum_{i=1}^{n} w_i x_i}}$$

$$\frac{y}{1-y} = e^{w_1 x + w_2}$$



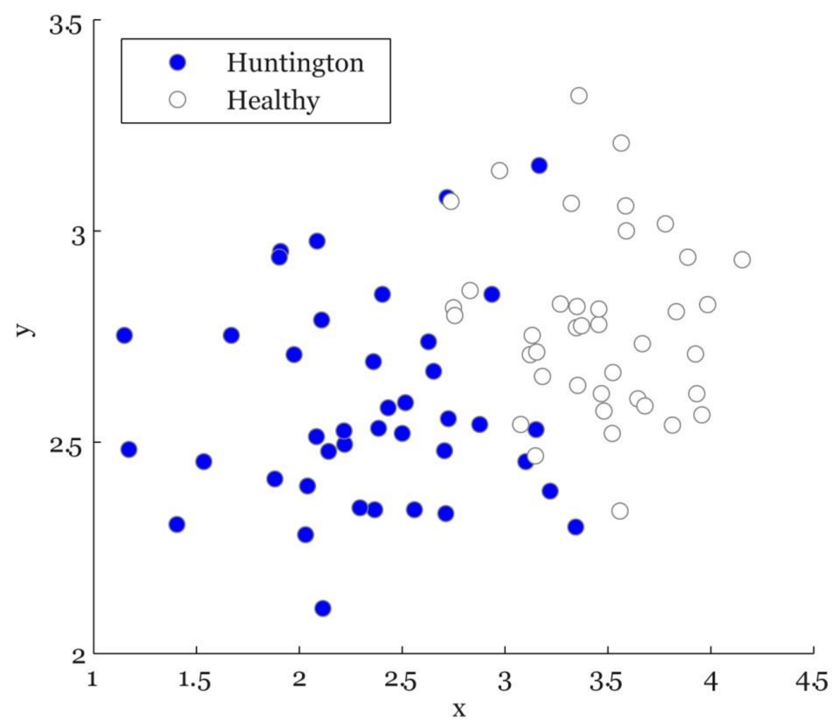
$$\ln\left(\frac{y}{1-y}\right) = w_1 x + w_2$$

Logistic regression

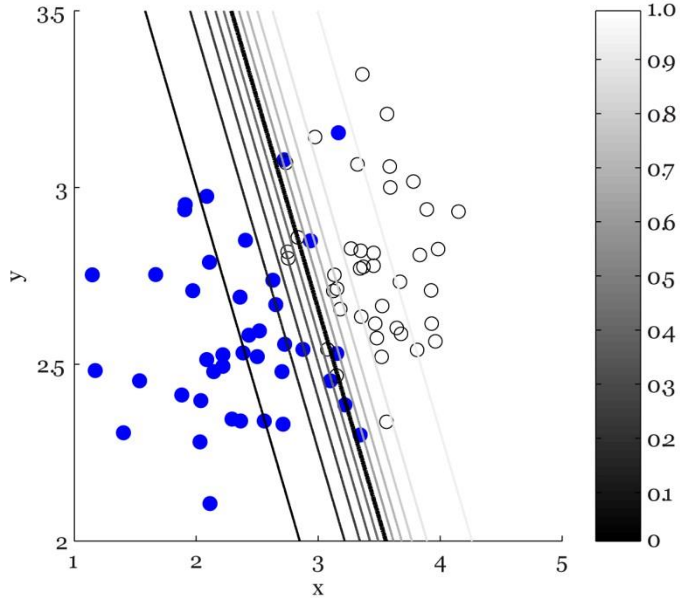


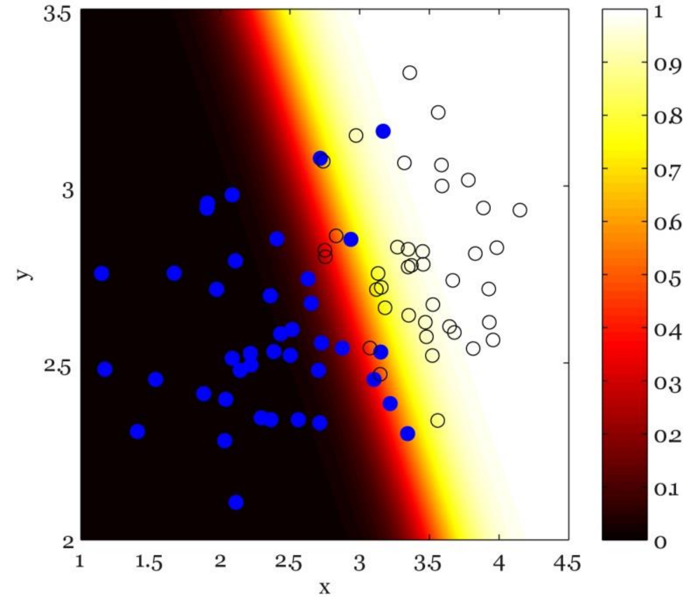
Data

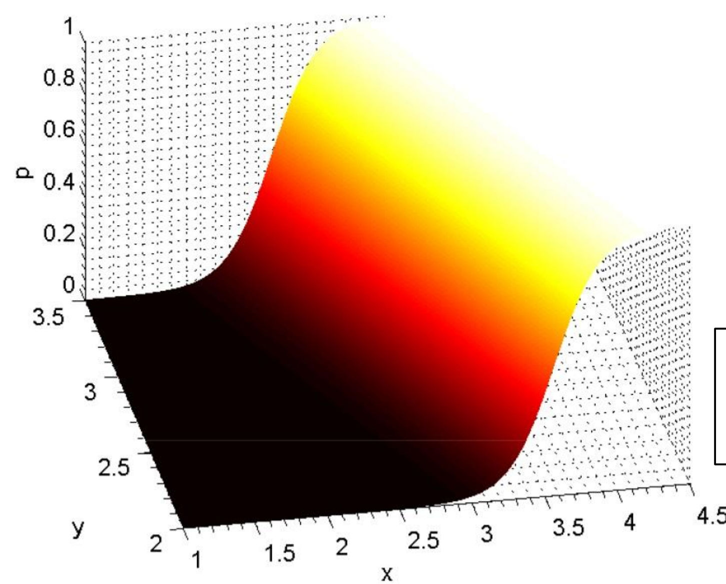
Among other speech problems, speakers with Huntington's disease (HD) typically manifest slower articulation rate and imprecise articulation of vowels. We collected reading passages from 40 speakers with HD and 40 age- and sex-matched healthy controls. Subsequently, we extracted features related to articulation rate (feature x) and vowel articulation quality (feature y). We would like to know how combination of these 2 features is able to contribute to correct diagnosis and thus robustly separate HD from controls.



Logistic regression







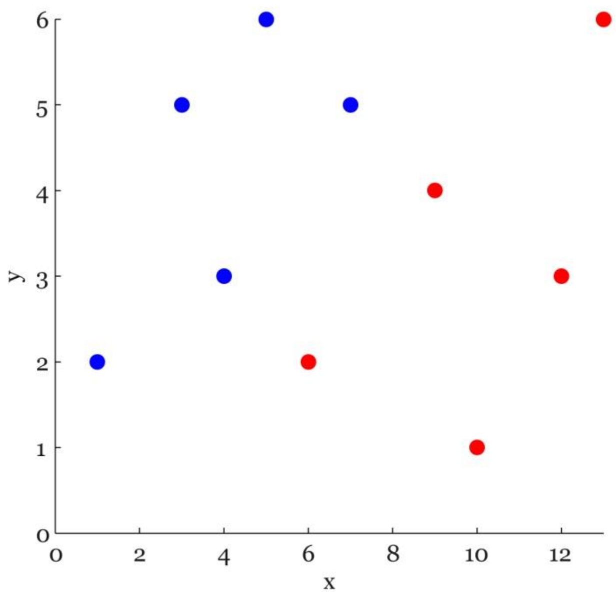
Accuracy = 88.8%

logistic regression: glmfit
class assignment: glmval

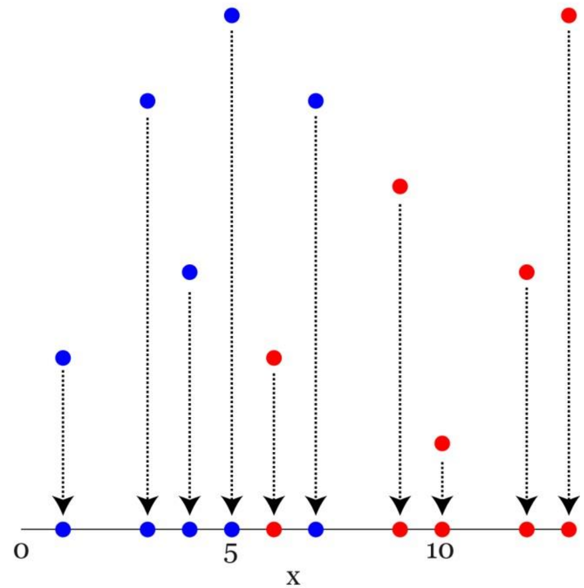
Matlab example 1

Linear discriminant analysis (aka Fisher discriminant analysis)

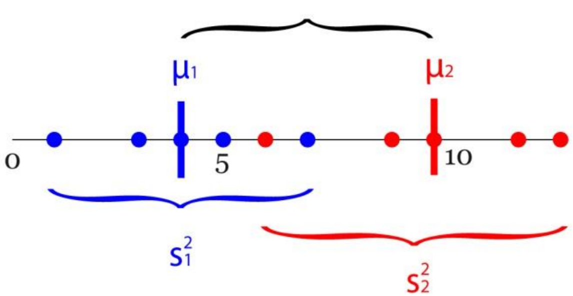
Simple example of 2D data



Reducing 2D data to 1D data



Maximize the distance between means



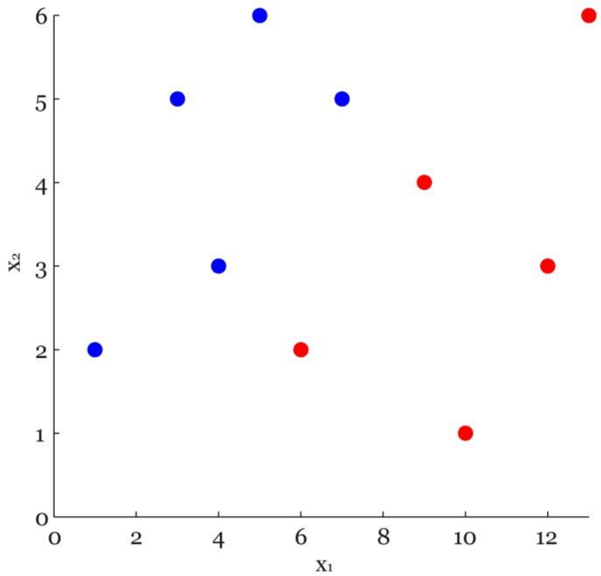
Minimize the variation (called „scatter“ in LDA) within each category

Fisher linear discriminant

$$J(w) = \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

Measure of the difference *between*-class means normalized by a measure of *within*-class scatter matrix

How to compute Linear Discrimination projection?



The classes are:

Sample for class ω_1:
$X_1=(x_1,x_2)=\{(1, 2),(3,5),(4,3),(5,6),(7,5)\}$

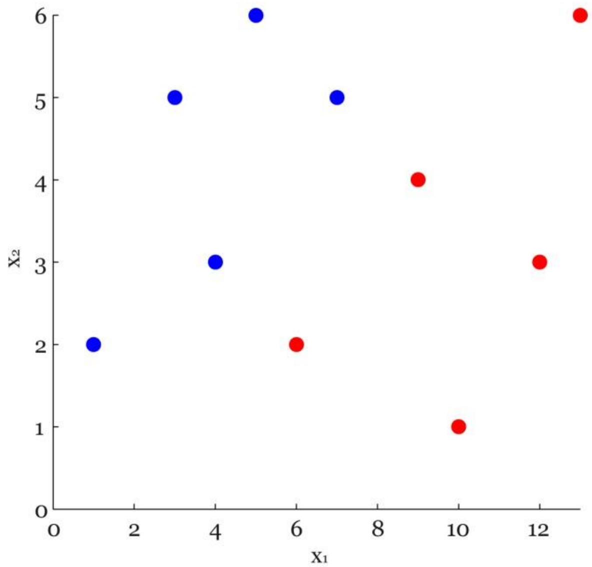
Sample for class ω_2:
$X_2=(x_1,x_2)=\{(6, 2),(9,4),(10,1),(12,3),(13,6)\}$

The classes means are:

$$\mu_1 = \frac{1}{N}\sum_{x\in\omega_1} x = \frac{1}{5}\left[\begin{pmatrix}1\\2\end{pmatrix}+\begin{pmatrix}3\\5\end{pmatrix}+\begin{pmatrix}4\\3\end{pmatrix}+\begin{pmatrix}5\\6\end{pmatrix}+\begin{pmatrix}7\\5\end{pmatrix}\right]=\begin{pmatrix}4\\4.2\end{pmatrix}$$

$$\mu_2 = \frac{1}{N}\sum_{x\in\omega_2} x = \frac{1}{5}\left[\begin{pmatrix}6\\2\end{pmatrix}+\begin{pmatrix}9\\4\end{pmatrix}+\begin{pmatrix}10\\1\end{pmatrix}+\begin{pmatrix}12\\3\end{pmatrix}+\begin{pmatrix}13\\6\end{pmatrix}\right]=\begin{pmatrix}10\\3.2\end{pmatrix}$$

How to compute Linear Discrimination projection?



The classes are:

Sample for class ω_1:
$X_1=(x_1,x_2)=\{(1, 2),(3,5),(4,3),(5,6),(7,5)\}$

Sample for class ω_2:
$X_2=(x_1,x_2)=\{(6, 2),(9,4),(10,1),(12,3),(13,6)\}$

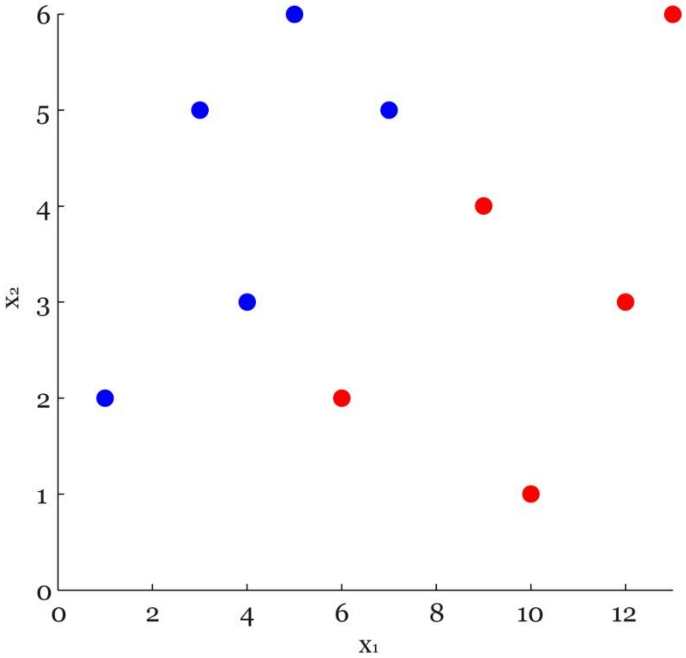
Covariance matrix of the first class:

$$S_1 = \sum_{x\in\omega_1}(x-\mu_1)(x-\mu_1)^T$$
$$= \left[\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}4\\4.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}3\\5\end{pmatrix}-\begin{pmatrix}4\\4.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}4\\3\end{pmatrix}-\begin{pmatrix}4\\4.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}5\\6\end{pmatrix}-\begin{pmatrix}4\\4.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}7\\5\end{pmatrix}-\begin{pmatrix}4\\4.2\end{pmatrix}\right]^2$$
$$= \begin{pmatrix}5 & 2.5\\2.5 & 2.7\end{pmatrix}$$

Covariance matrix of the second class:

$$S_2 = \sum_{x\in\omega_2}(x-\mu_2)(x-\mu_2)^T$$
$$= \left[\begin{pmatrix}6\\2\end{pmatrix}-\begin{pmatrix}10\\3.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}9\\4\end{pmatrix}-\begin{pmatrix}10\\3.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}10\\1\end{pmatrix}-\begin{pmatrix}10\\3.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}12\\3\end{pmatrix}-\begin{pmatrix}10\\3.2\end{pmatrix}\right]^2+\left[\begin{pmatrix}13\\6\end{pmatrix}-\begin{pmatrix}10\\3.2\end{pmatrix}\right]^2$$
$$= \begin{pmatrix}7.5 & 3\\3 & 3.7\end{pmatrix}$$

How to compute Linear Discrimination projection?



The main parameters are:

$$\mu_1 = \begin{pmatrix} 4 \\ 4.2 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 10 \\ 3.2 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 5 & 2.5 \\ 2.5 & 2.7 \end{pmatrix} \quad S_2 = \begin{pmatrix} 7.5 & 3 \\ 3 & 3.7 \end{pmatrix}$$

Within-class scatter matrix:

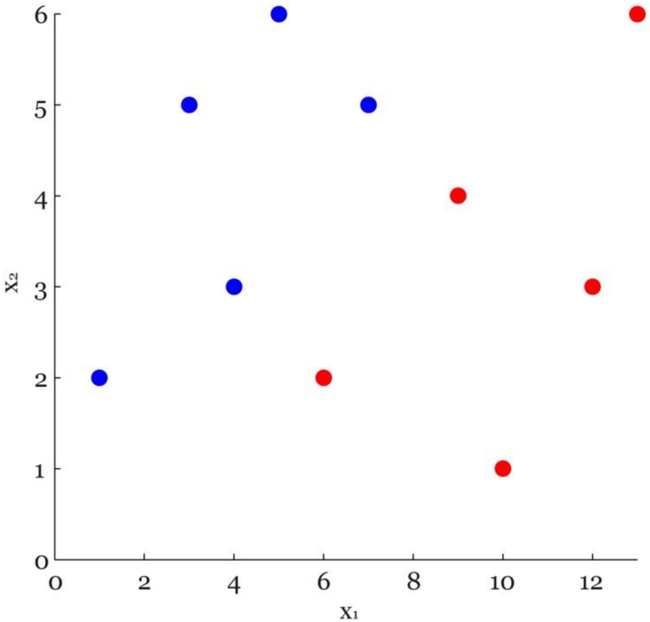
$$S_w = S_1 + S_2 = \begin{pmatrix} 5 & 2.5 \\ 2.5 & 2.7 \end{pmatrix} + \begin{pmatrix} 7.5 & 3 \\ 3 & 3.7 \end{pmatrix} = \begin{pmatrix} 12.5 & 5.5 \\ 5.5 & 6.4 \end{pmatrix}$$

Between-class scatter matrix:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \left[\begin{pmatrix} 4 \\ 4.2 \end{pmatrix} - \begin{pmatrix} 10 \\ 3.2 \end{pmatrix}\right]\left[\begin{pmatrix} 4 \\ 4.2 \end{pmatrix} - \begin{pmatrix} 10 \\ 3.2 \end{pmatrix}\right]^T$$

$$= \begin{pmatrix} -6 \\ 1 \end{pmatrix}(-6 \quad 1) = \begin{pmatrix} 36 & -6 \\ -6 & 1 \end{pmatrix}$$

How to compute Linear Discrimination projection?



The updated parameters are:

$$\mu_1 = \begin{pmatrix} 4 \\ 4.2 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 10 \\ 3.2 \end{pmatrix}$$

$$S_w = \begin{pmatrix} 12.5 & 5.5 \\ 5.5 & 6.4 \end{pmatrix} \quad S_B = \begin{pmatrix} 36 & -6 \\ -6 & 1 \end{pmatrix}$$

LDA projection is obtained as the solution of generalized eigen value problem:

$${S_w}^{-1} S_B w = \lambda w \qquad \textit{Highest eigen value corresponds to the best projection}$$

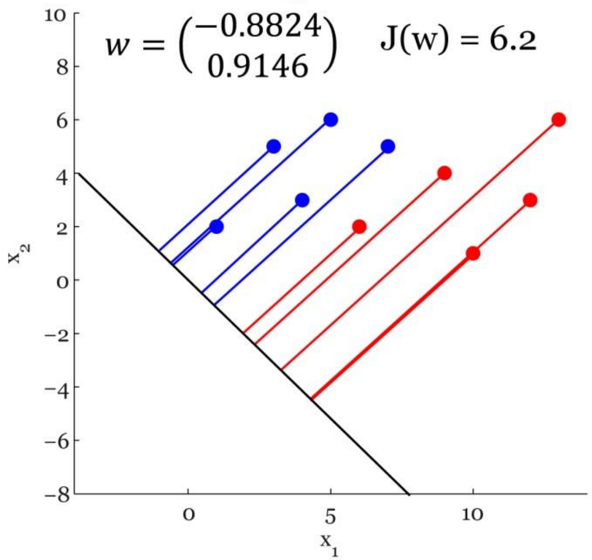
Or directly:

$$w = {S_w}^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 12.5 & 5.5 \\ 5.5 & 6.4 \end{pmatrix}^{-1} \left[\begin{pmatrix} 4 \\ 4.2 \end{pmatrix} - \begin{pmatrix} 10 \\ 3.2 \end{pmatrix} \right]$$

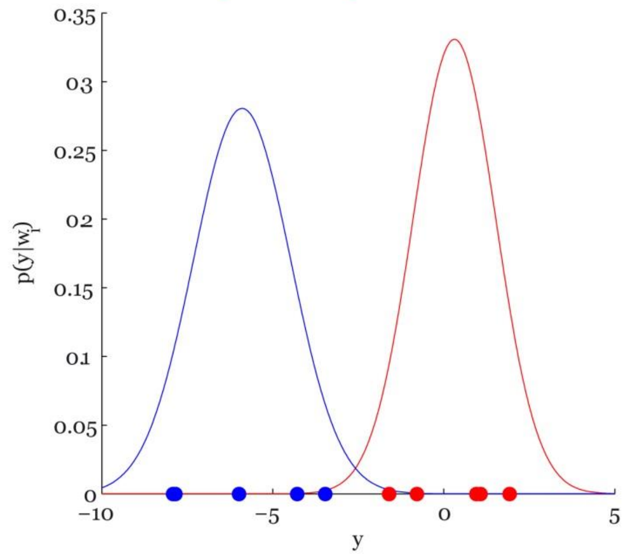
$$= \begin{pmatrix} 0.1286 & -0.1106 \\ -0.1106 & 0.2513 \end{pmatrix} \begin{pmatrix} -6 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.8824 \\ 0.9146 \end{pmatrix}$$

LDA projection

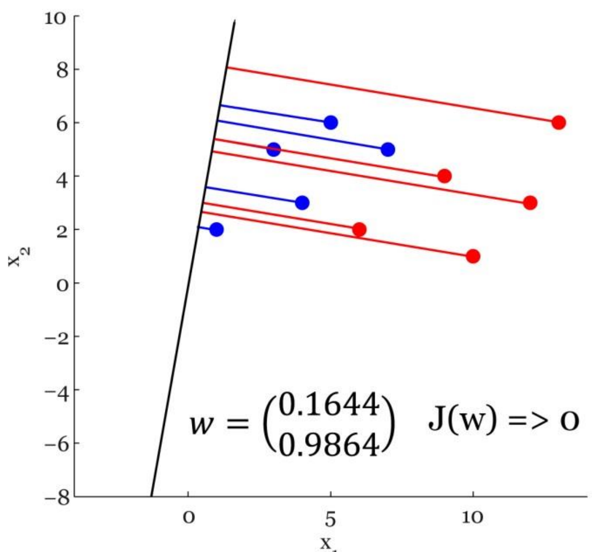
Vector with highest eigen value



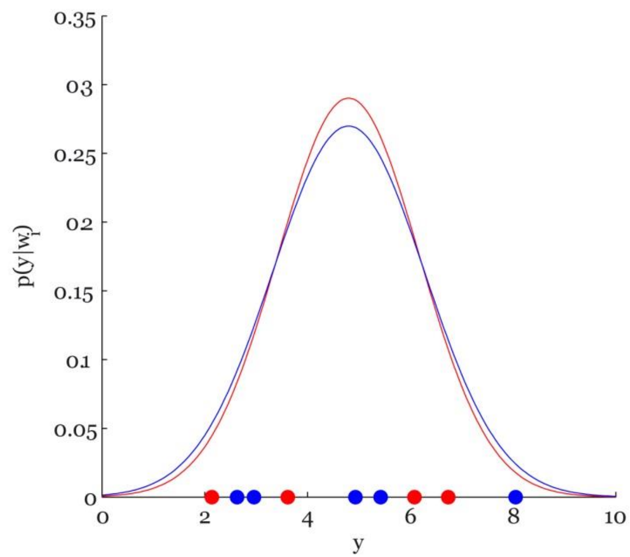
Good separability between two classes



Vector with lowest eigen value

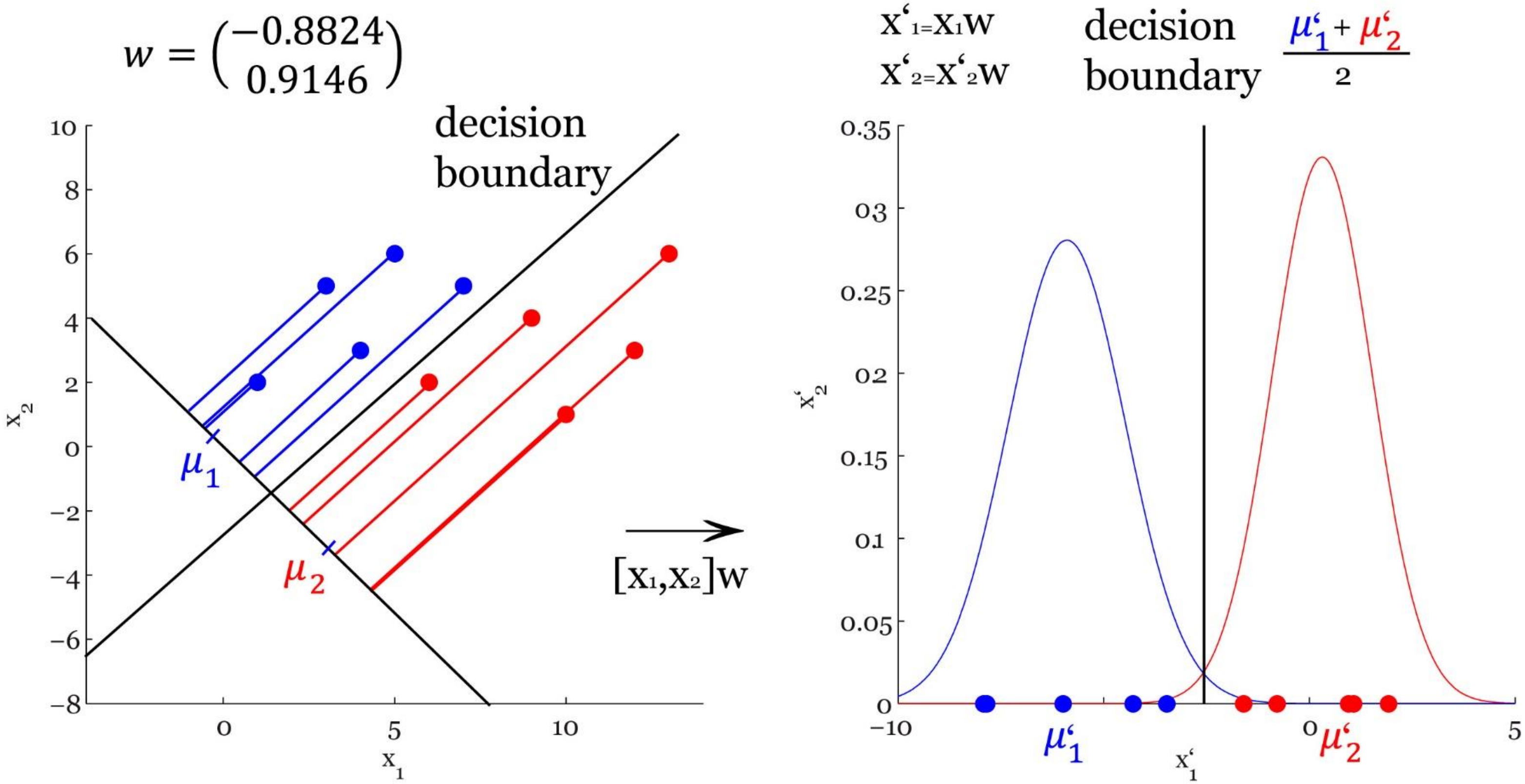


Bad separability between two classes

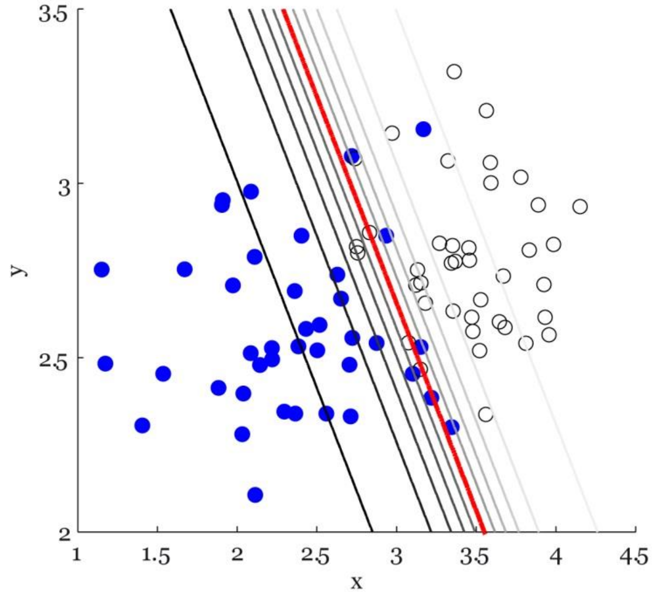


LDA decision

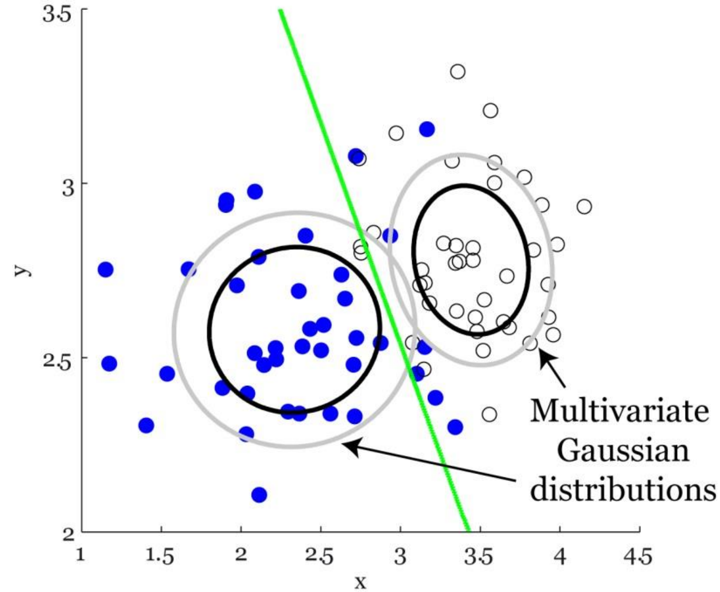
$$w = \begin{pmatrix} -0.8824 \\ 0.9146 \end{pmatrix}$$



Logistic regression



Linear discrimination analysis

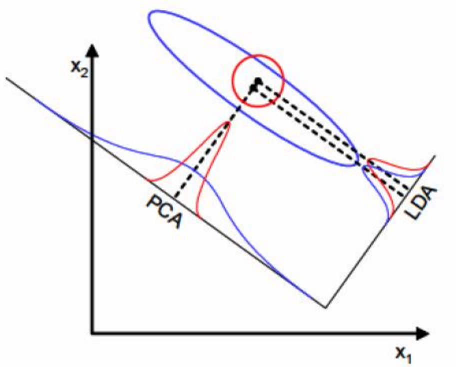


Logistic regression: 88.8%
Linear discriminant analysis: 88.8%

LDA: fitcdiscr
class assignment: predict

Discriminant analysis

+ perform a dimensionality reduction while preserve as much of the class discriminatory information as possible
+ can be easily extended to classify more than two classes
- it is parametric method that assumes unimodal Gaussian distributions
- it will fail when the discriminatory information is not in the mean but rather in the variance of the data



Logistic regression

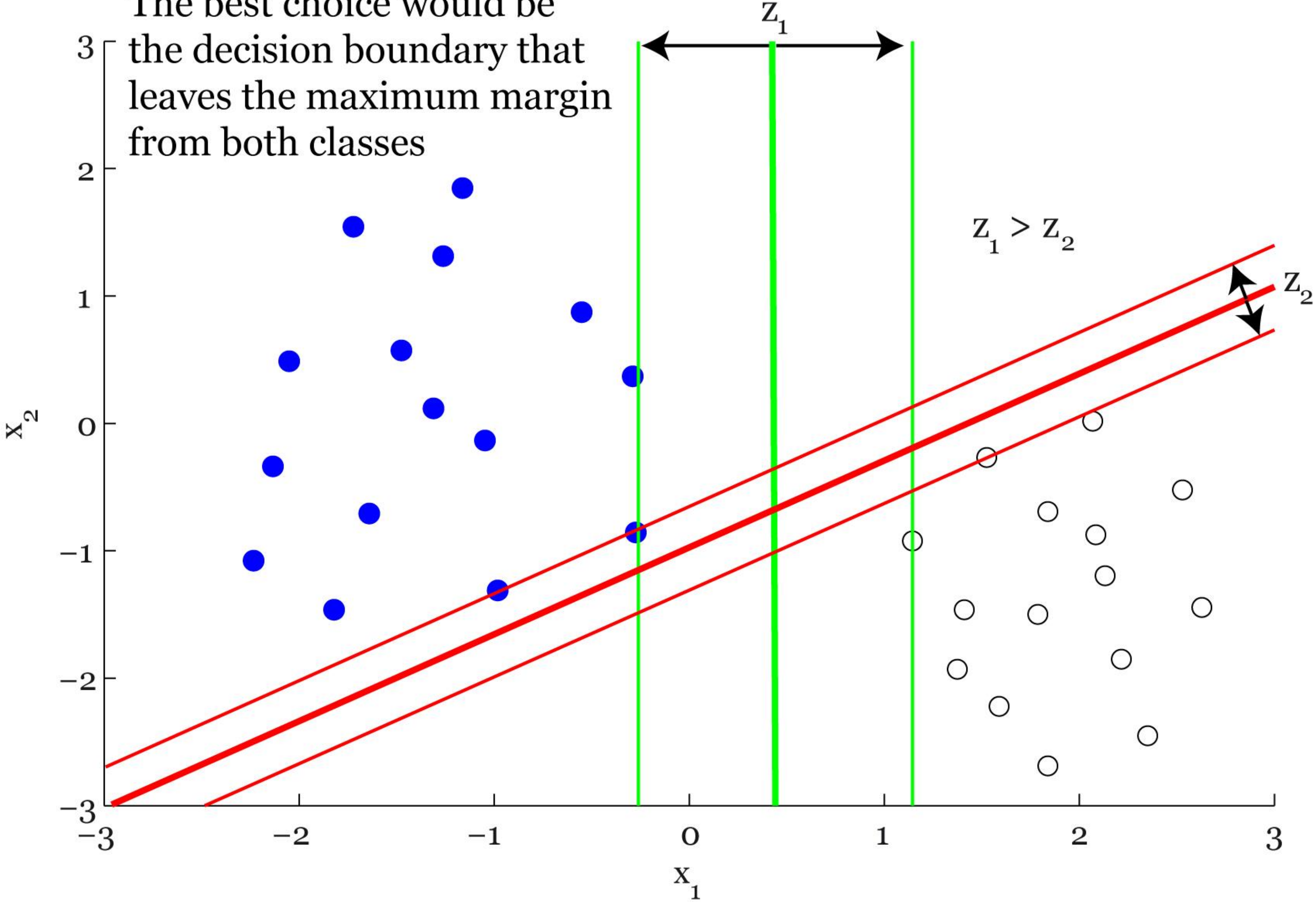
+ more robust
+ variable do not need to be normally distributed
+ there is no homogeneity of variance assumption
+ it may handle nonlinear effects

Why to even consider discriminant analysis over logistic regression?

- logistic regression requires greater data sample to achieve stable and meaningful results - at least 50 data points per predictor are necessary
+ typically 20 data points per predictor are considered the lower bound for discriminant analysis

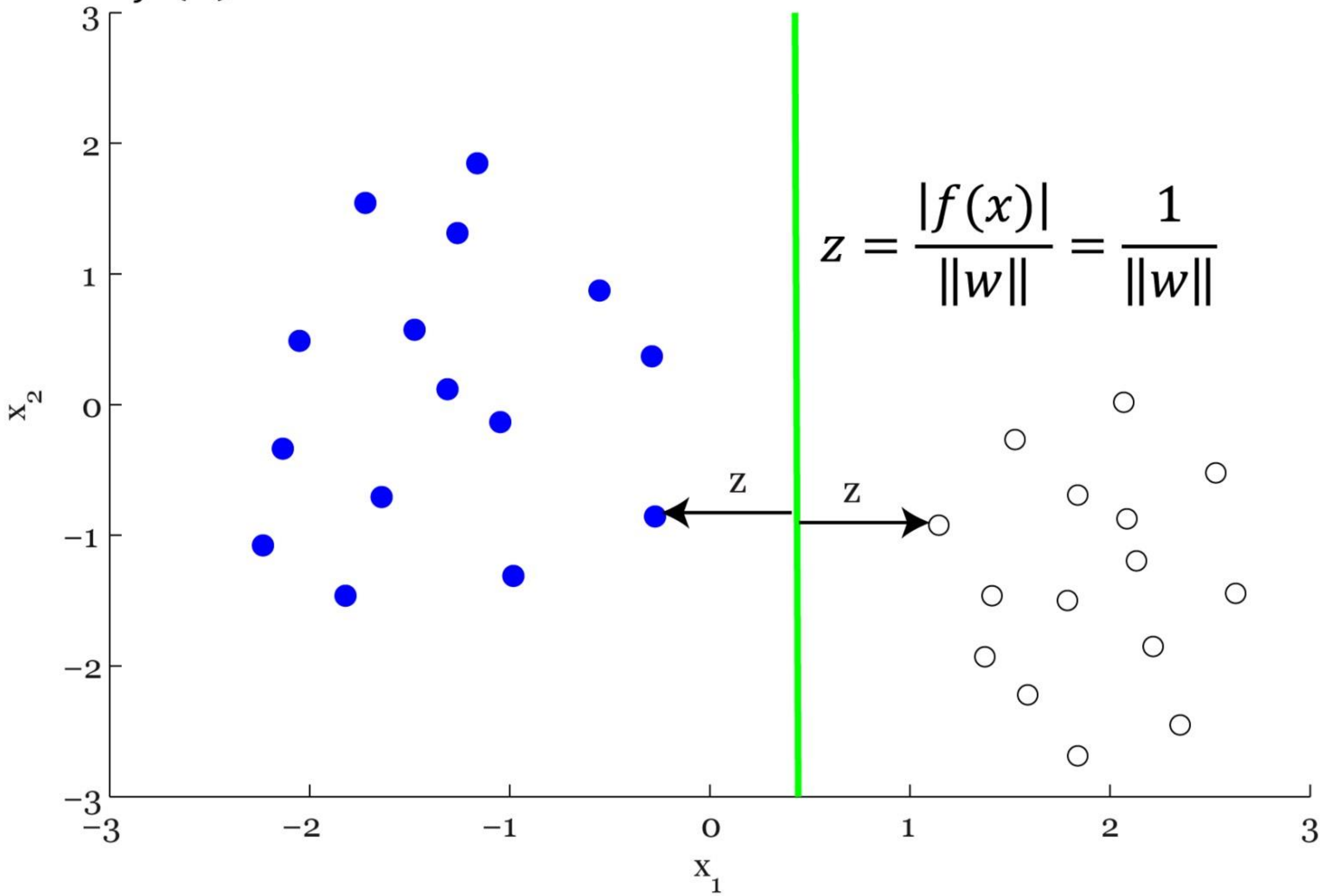
Support Vector Machine

The best choice would be the decision boundary that leaves the maximum margin from both classes



Support Vector Machine

$f(x) \geq 1$ class ● $f(x) = w^T x + b$ $f(x) \leq -1$ class ○



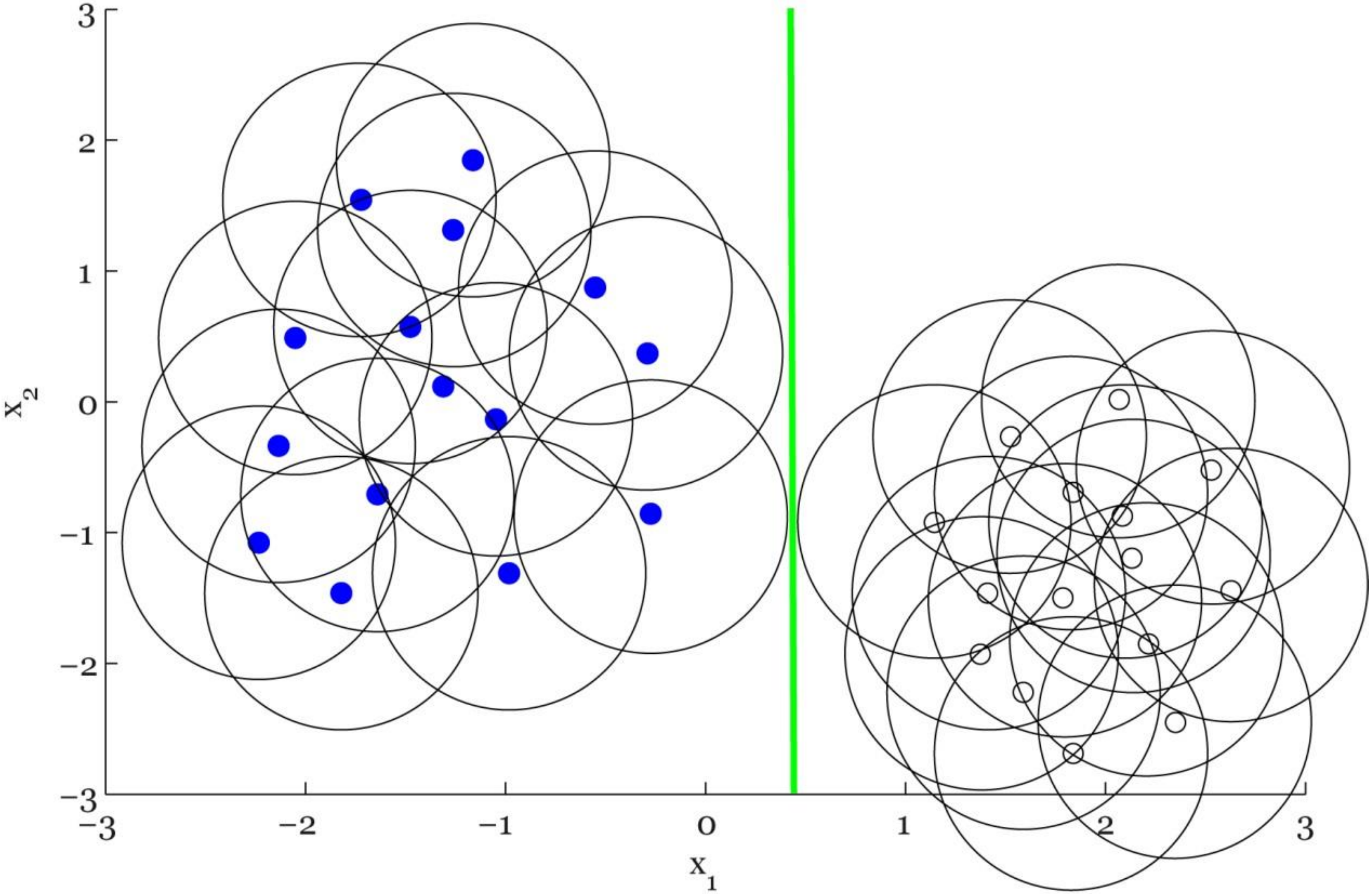
Total margin is computed by:

$$\frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|}$$

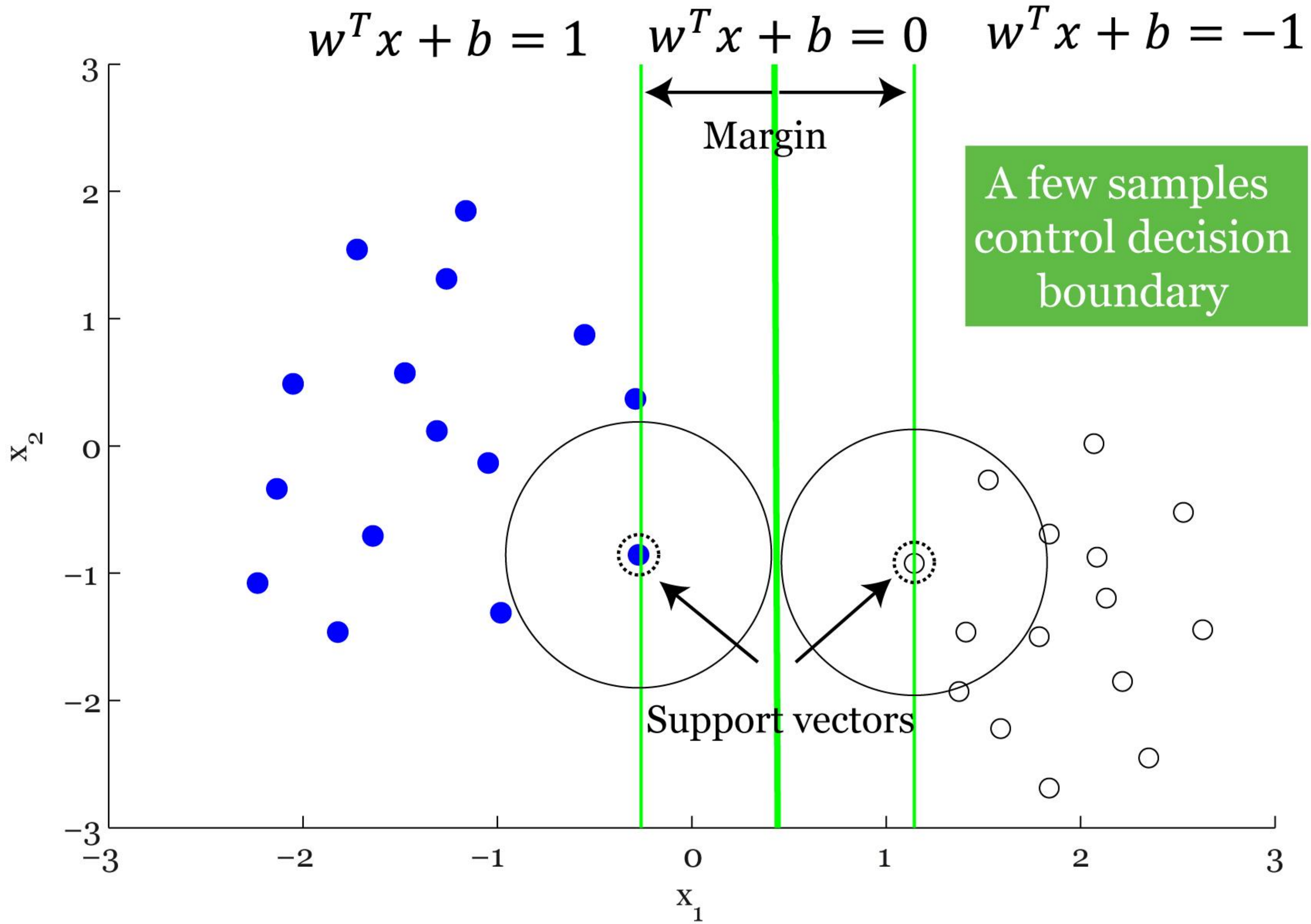
Minimazing this term will maximize separability

Can be solved by quadratic programming (*quadprog*)

Margin: bubles around samples

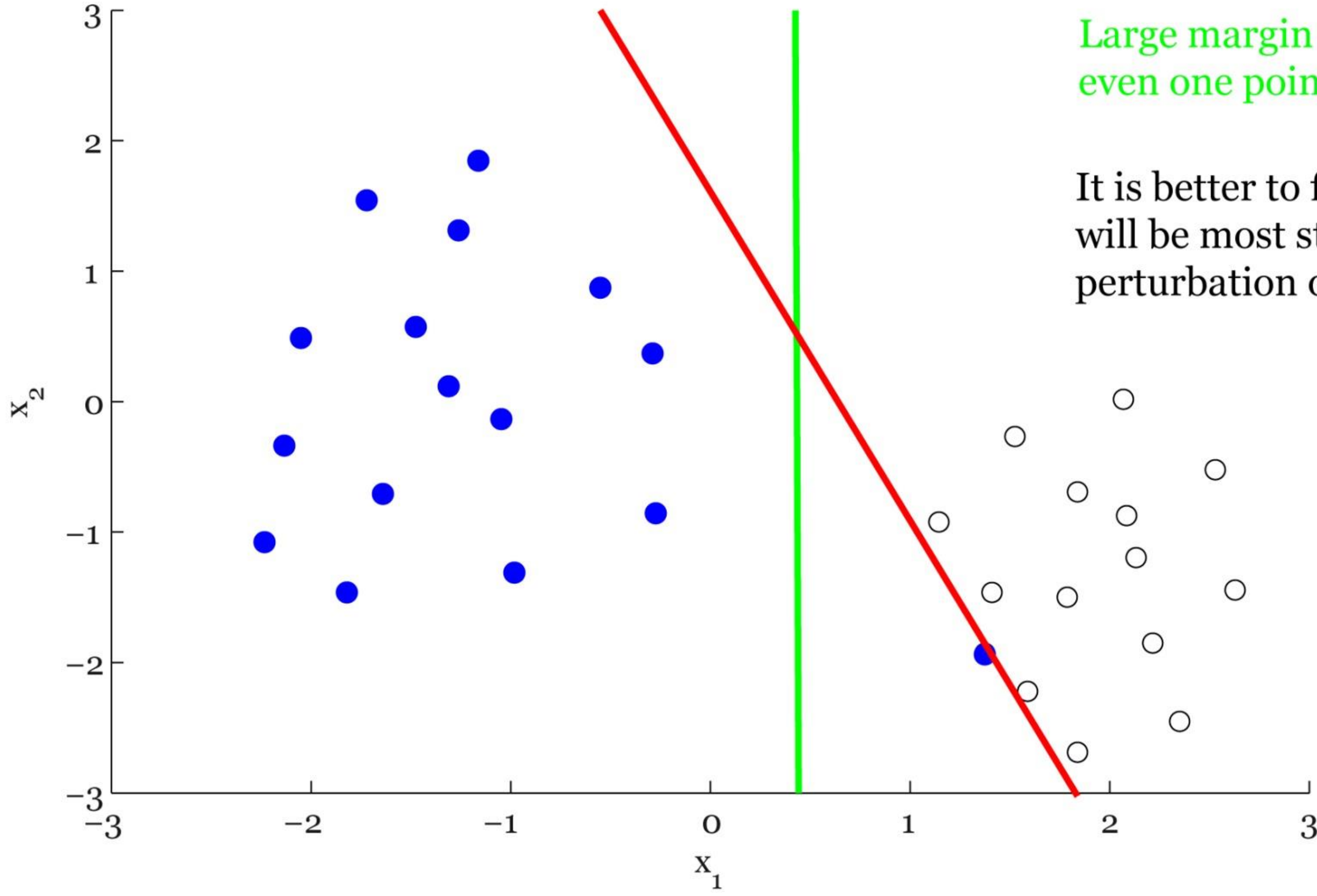


Support vectors



What is the best weight?

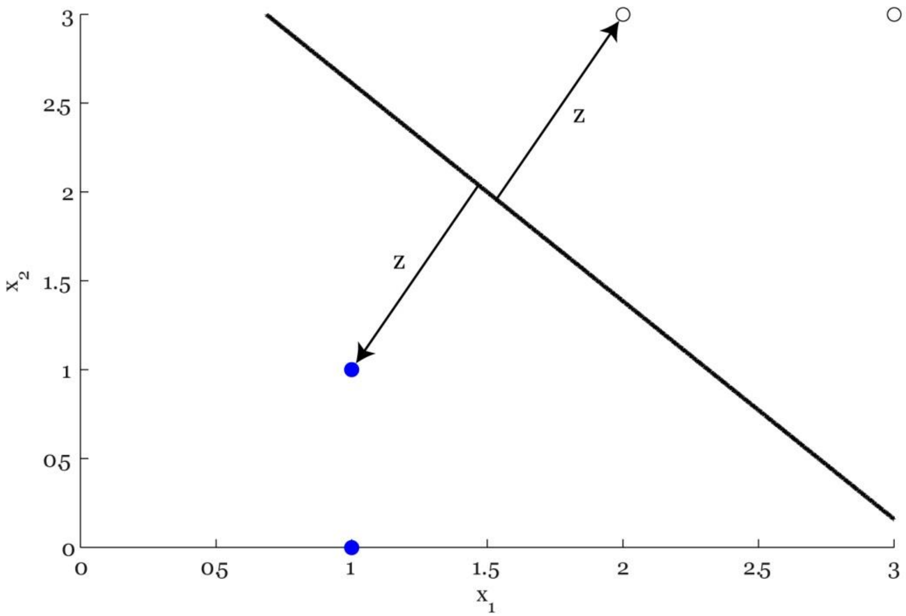
Data can be linearly separated but margin is pretty narrow



Large margin solution is likely better, even one point is classified incorrectly

It is better to find solution that will be most stable under perturbation of the inputs

SVM basic example for linearly separable data



Sample for class ω_1:
$X_1=(x_1,x_2)=\{(1,0),(1,1)\}$

Sample for class ω_2:
$X_2=(x_1,x_2)=\{(2,3),(3,3)\}$

By visual inspection:
support vectors$=(x_1,x_2)=\{(1,1),(2,3)\}$

Weight vector:
w = (2,3) - (1,1) = (a,2a)

$$w^T x + b = 1$$

$$w^T x + b = -1$$

a + 2a + b = -1 using point (1,1)
2a + 6a + b = 1 using point (2,3)

b = 1 - 8a
3a + (1 - 8a) = -1
a = 2/5

b = 1 -8*(2/5)
b = -11/5

Weight vector:

$$w = (a, 2a) = \left(\frac{2}{5}, \frac{4}{5}\right)$$

Decision boundary:

$$f(x) = w^T x + b$$

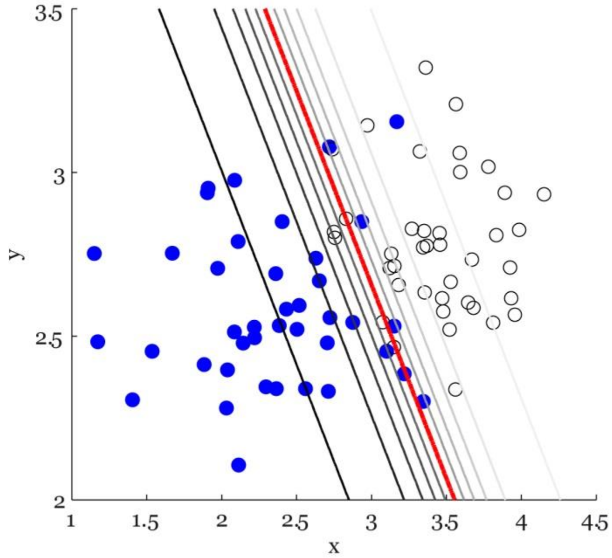
$$f(x) = \frac{2}{5}x_1 + \frac{4}{5}x_2 - \frac{11}{5}$$

$$f(x) = x_1 + 2x_2 - 5.5$$

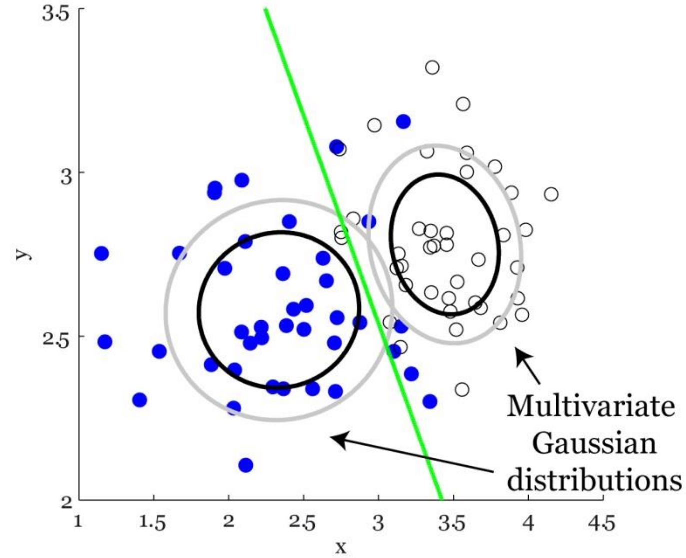
Margin:

$$z = \frac{2}{\|w\|} = \frac{2}{\sqrt{\frac{4}{25} + \frac{16}{25}}} = \sqrt{5}$$

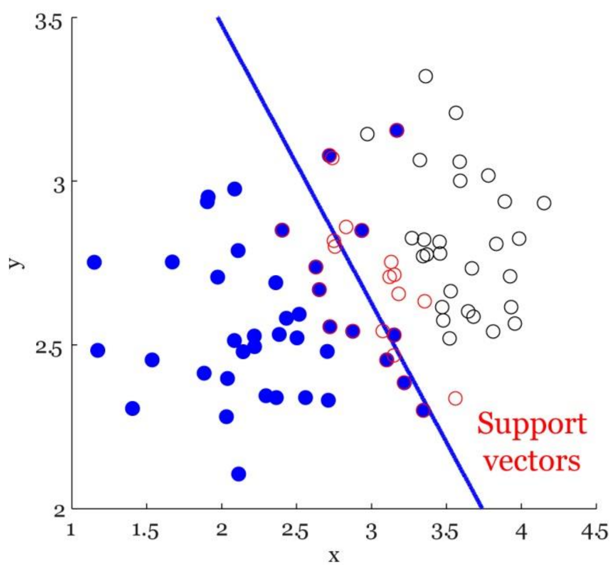
Logistic regression



Linear discrimination analysis



Support vector machine



Logistic regression: 88.8%
Linear discriminant analysis: 88.8%
Support vector machine: 90.0%

SVM: fitcsvm
class assignment: predict

LIBSVM

Support vector machine

\+ better performance in most cases
\+ works well for high number of dimensions and not linearly separated data
\+ computationally cheaper $O(N^2*K)$ where K is number of support vectors, whereas logistic regression $O(N^3)$
\+ classifier depends only on a subset of points unlike logistic regression

\- kernel models can be quite sensitive to over-fitting