Statistics and Data Analysis in MATLAB

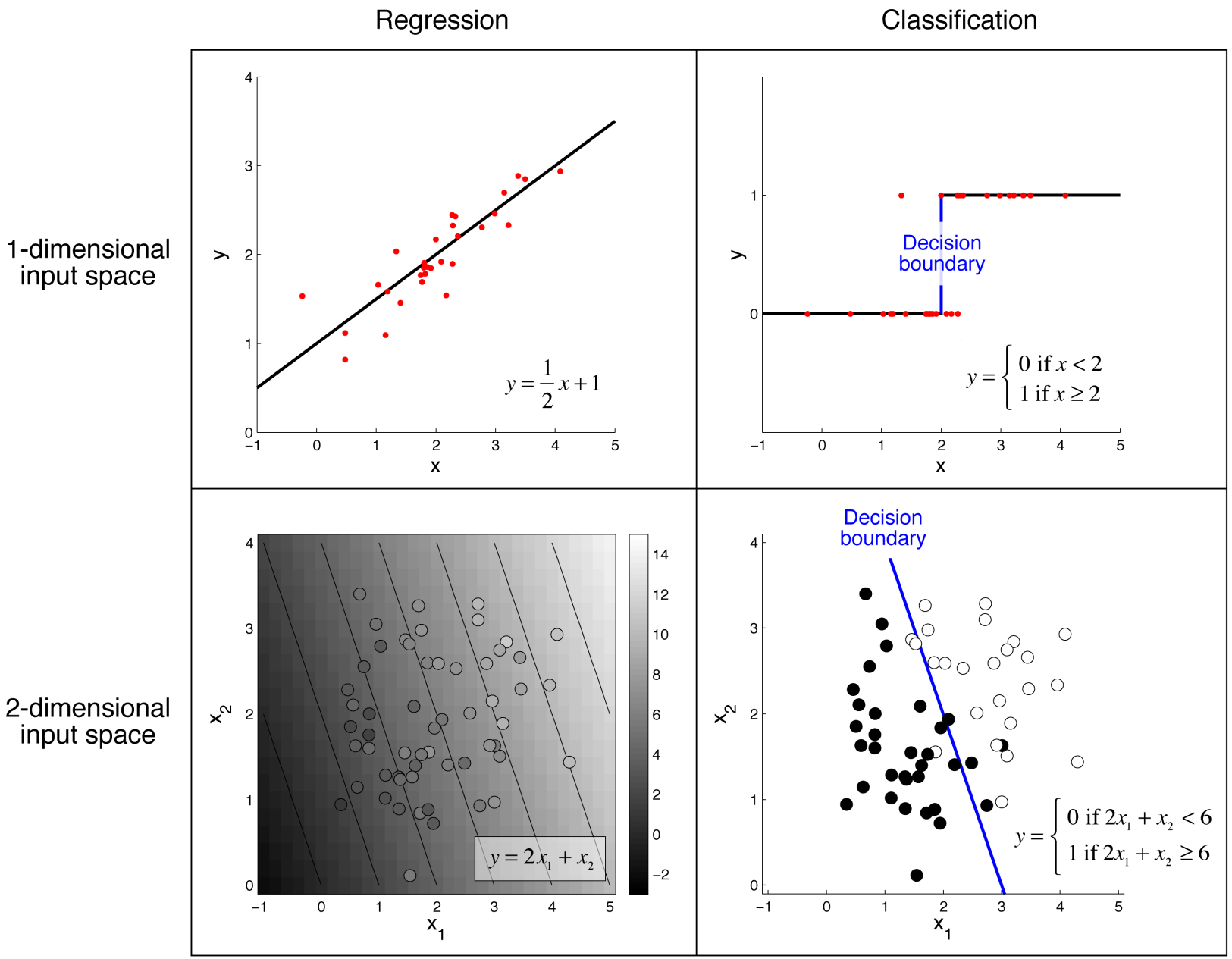
Lecture 7: Classification

Kendrick Kay
Washington University in St. Louis
April 11, 2014

Linear classification model

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

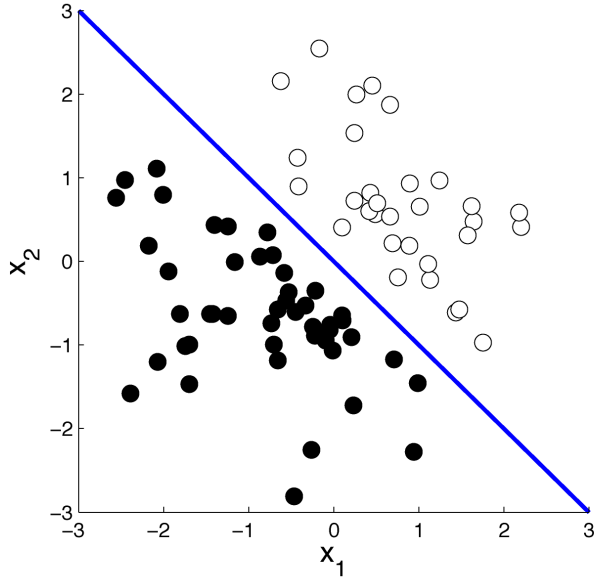
Comparing regression and classification



Kendrick Kay, Washington Univ. in St. Louis

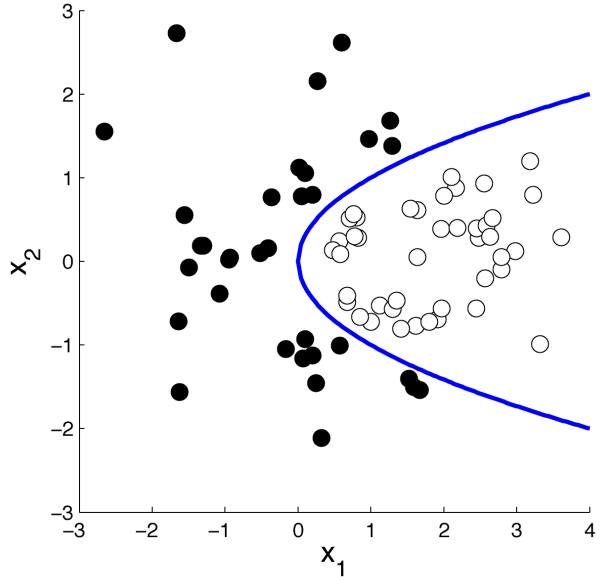
Nonlinear decision boundaries through input space expansion

Linear decision boundary



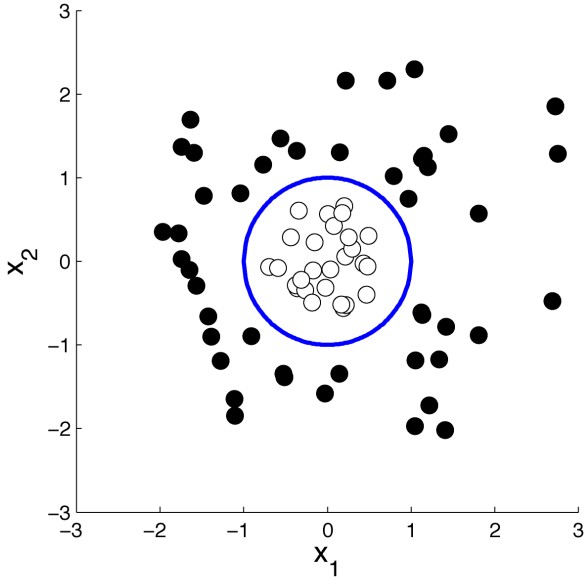
$$y = \begin{cases} 0 \text{ if } x_1 + x_2 < 0 \\ 1 \text{ if } x_1 + x_2 \geq 0 \end{cases}$$

Nonlinear decision boundary



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

Nonlinear decision boundary

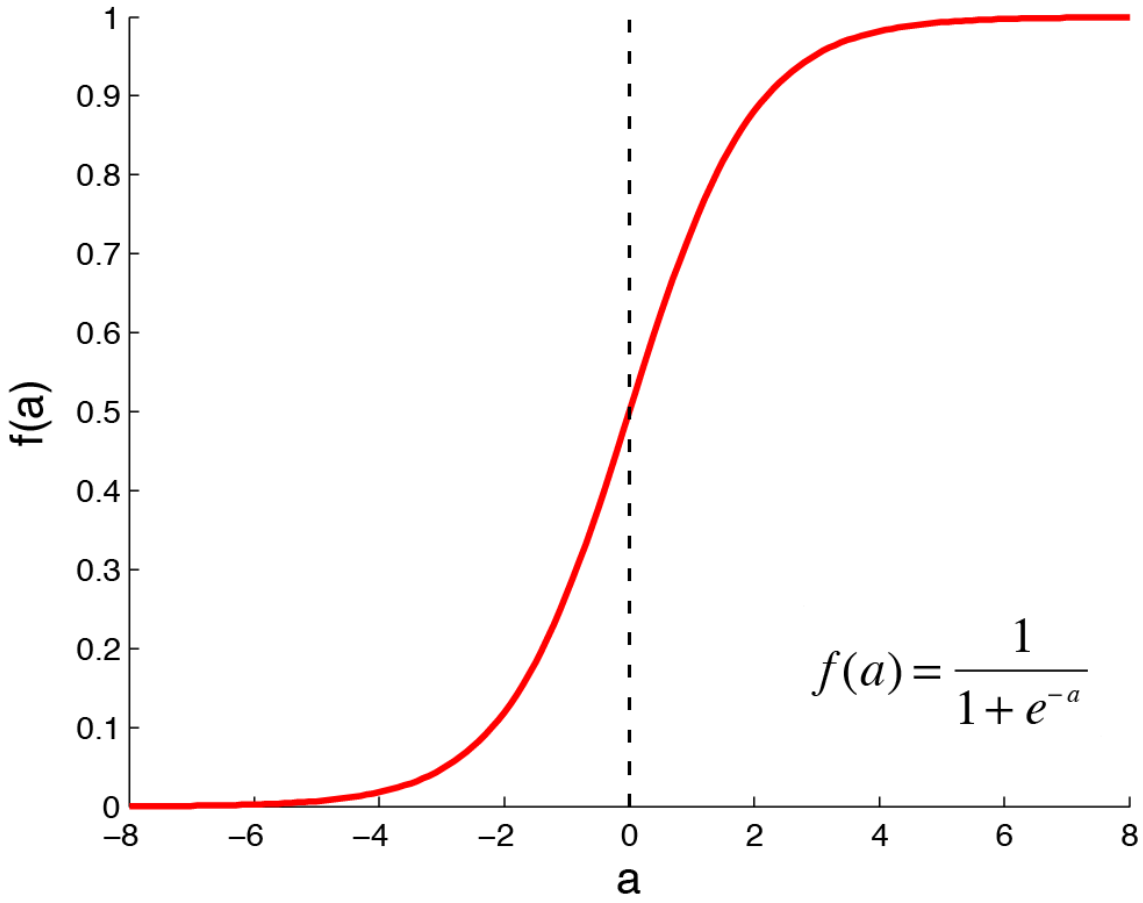


$$y = \begin{cases} 0 \text{ if } x_1^2 + x_2^2 < 1 \\ 1 \text{ if } x_1^2 + x_2^2 \geq 1 \end{cases}$$

Decision boundaries are linear in expanded input space:

$$\{x_1, x_2, x_1^2, x_2^2, x_1 x_2\}$$

Logistic function



Kendrick Kay, Washington Univ. in St. Louis

Logistic regression

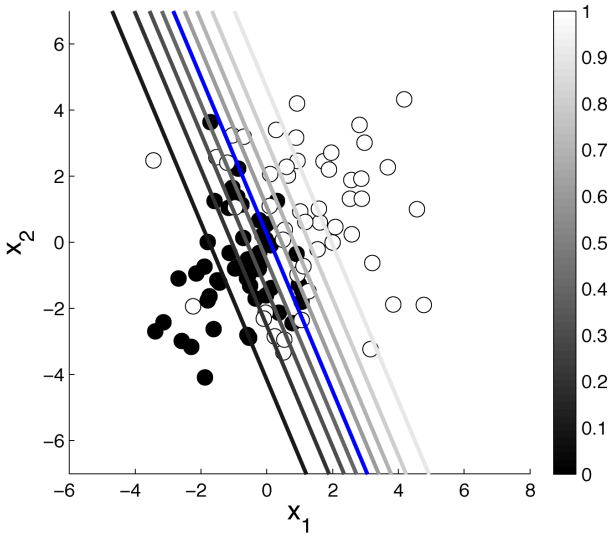
$$y = f\left(\sum_{i=1}^{n} w_i x_i\right) = \frac{1}{1 + e^{\left(-\sum_{i=1}^{n} w_i x_i\right)}}$$

$$\text{likelihood}(d \mid m) = \prod_{j=1}^{m} p(d_j) = \prod_{j=1}^{m} \left(y_j^{d_j} \left(1 - y_j\right)^{1-d_j}\right)$$

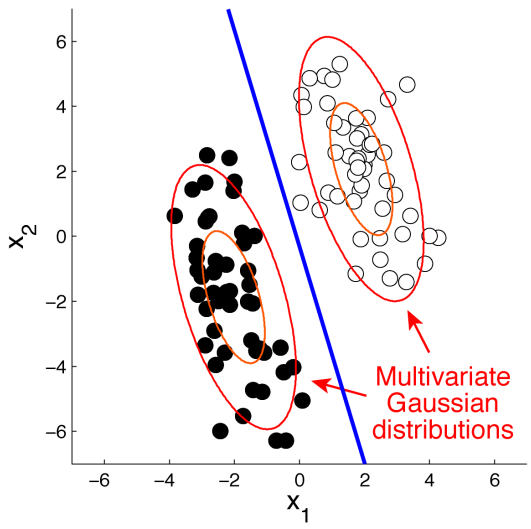
$$\text{negative-log-likelihood}(d \mid m) = -\sum_{j=1}^{m} \left(d_j \log(y_j) + (1 - d_j) \log(1 - y_j)\right)$$

Some classification techniques

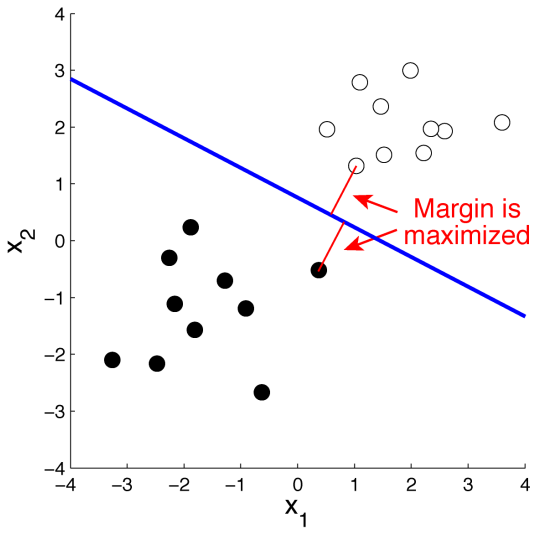
Logistic regression



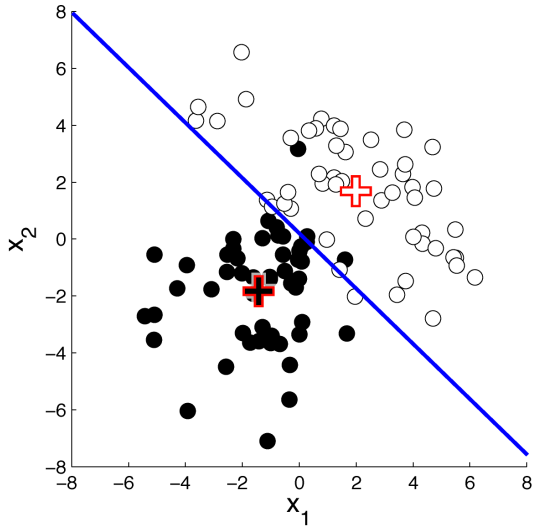
Linear discriminant analysis (LDA)



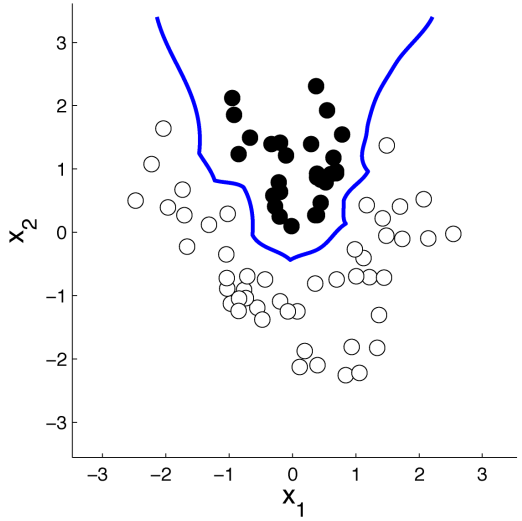
Support vector machines (SVM)



Nearest-prototype classification



Nearest-neighbor classification



Kendrick Kay, Washington Univ. in St. Louis