ELTE | FACULTY OF INFORMATICS

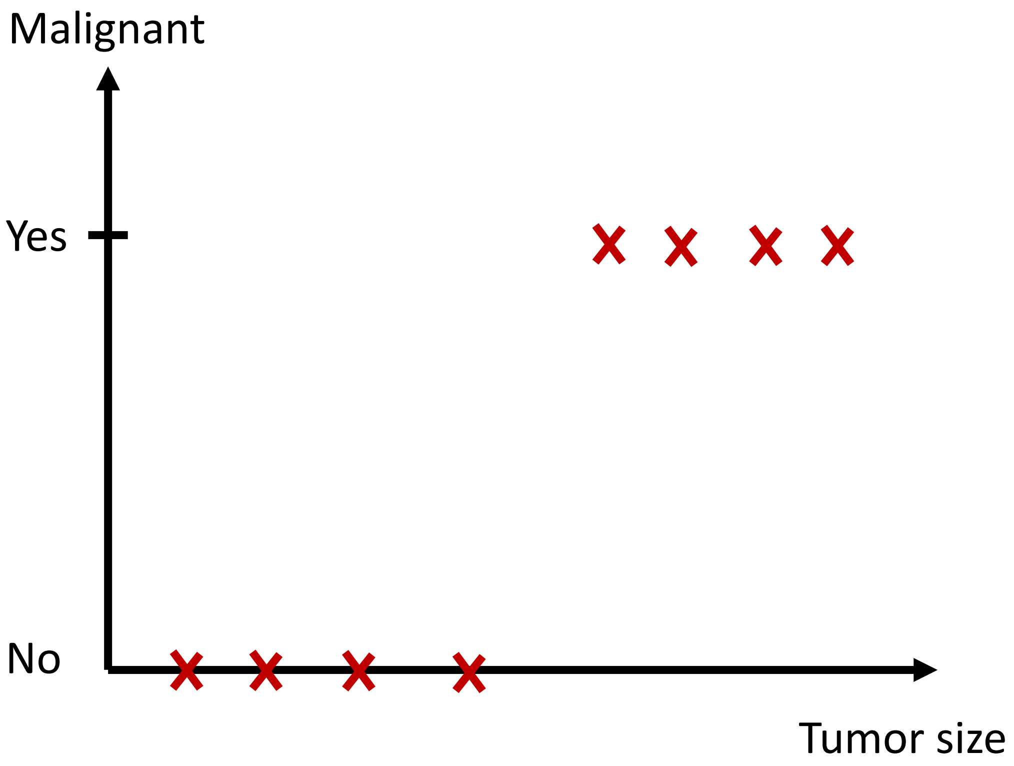
LOGISTIC REGRESSION

Machine Learning Course
Balázs Nagy, PhD

ELTE | IK
DEPARTMENT OF ARTIFICIAL INTELLIGENCE

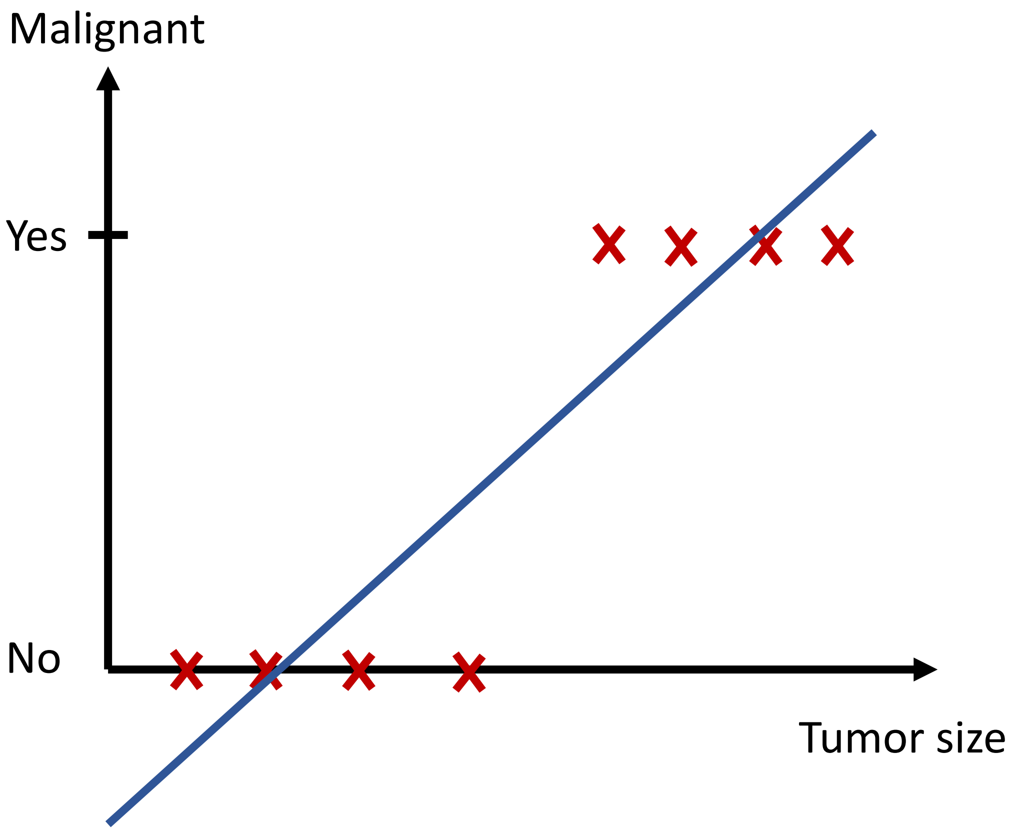
Classification problem

- **Example problem:**
  Classify tumors by their size into two class (malignant, non-malignant)



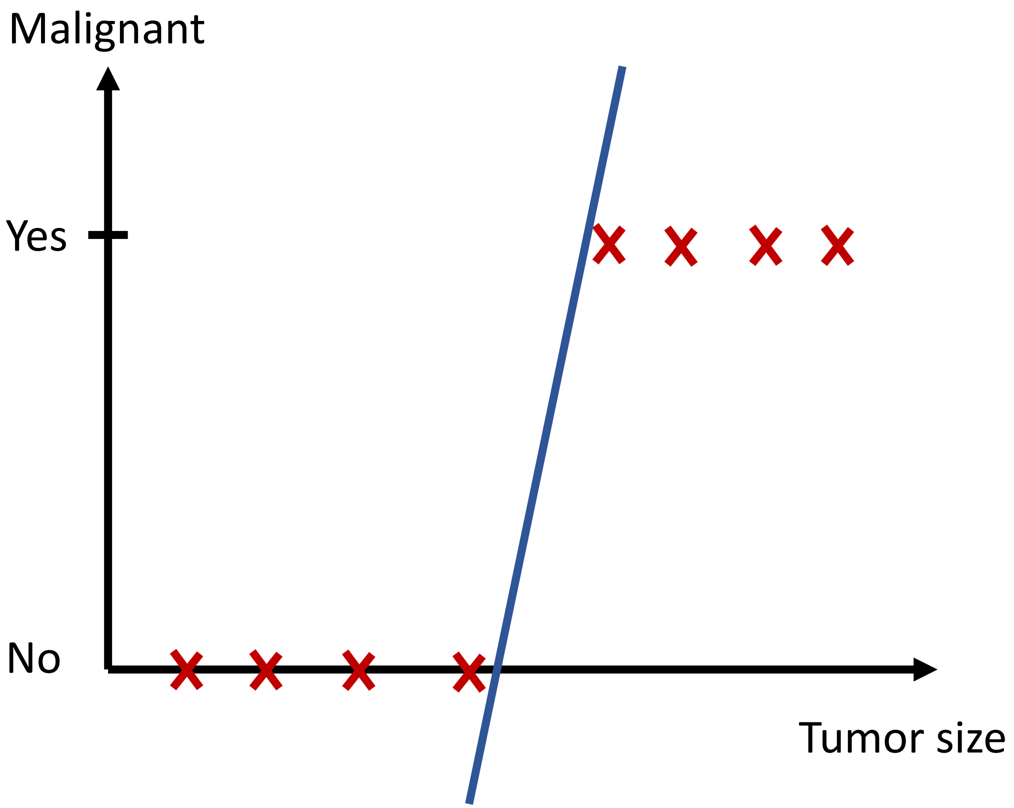
Classification problem

- **Example problem:** Classify tumors by their size into two class (malignant, non-malignant)
- Linea regression is not sufficient here



Classification problem

- **Example problem:** Classify tumors by their size into two class (malignant, non-malignant)
- Linea regression is not sufficient here



Vertical line as a treshold would be okey, but previous modell not designed for that

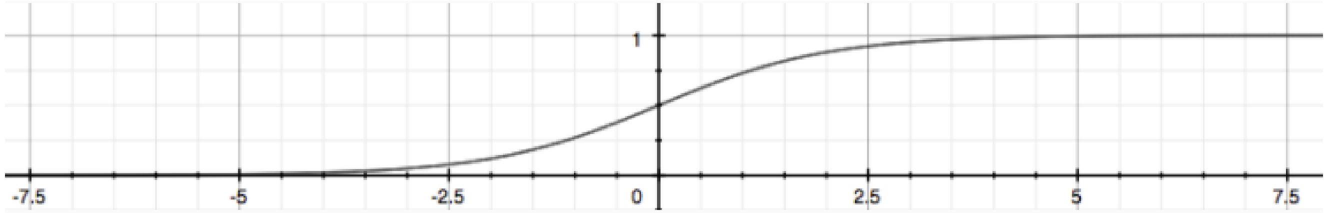
Classification

- Need a function with two outputs (y =0 or 1)
- Sigmoid function (also called Logistic Function)

$$h_\theta(x) = g(\theta^T x)$$

$$z = \theta^T x$$

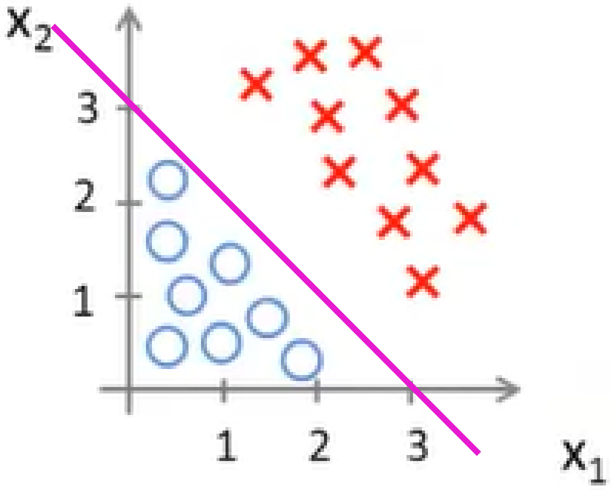
$$g(z) = \frac{1}{1 + e^{-z}}$$



- h(x) is a probability that the output is 1

$$h_\theta(x) = P(y = 1|x;\theta) = 1 - P(y = 0|x;\theta)$$
$$P(y = 0|x;\theta) + P(y = 1|x;\theta) = 1$$

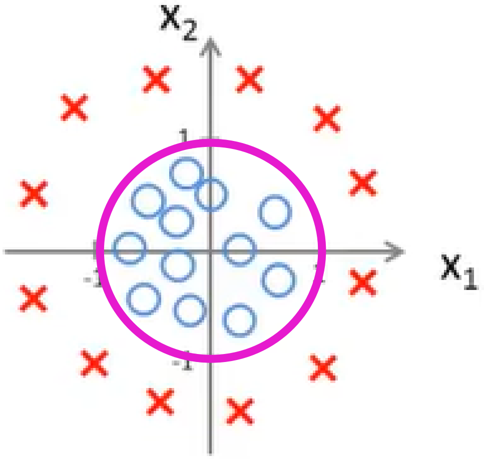
Decision Boundary - Linear



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

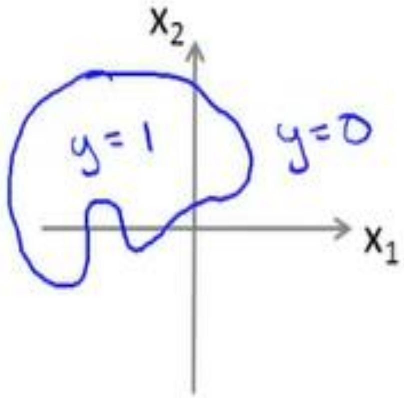
Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

Decision Boundary – Non-Linear



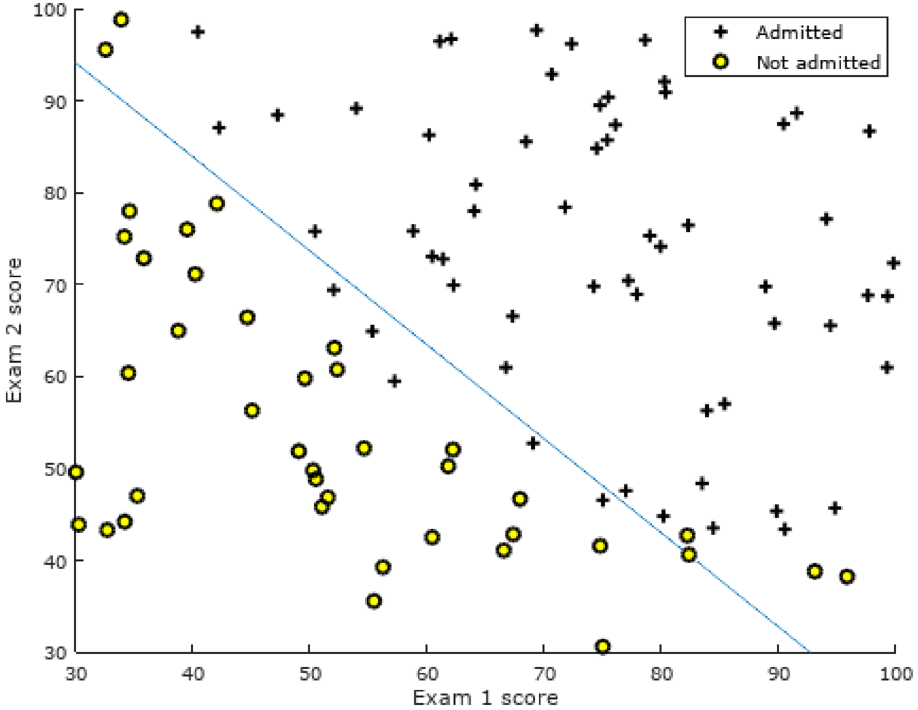
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Predict “$y = 1$“ if $-1 + x_1^2 + x_2^2 \geq 0$

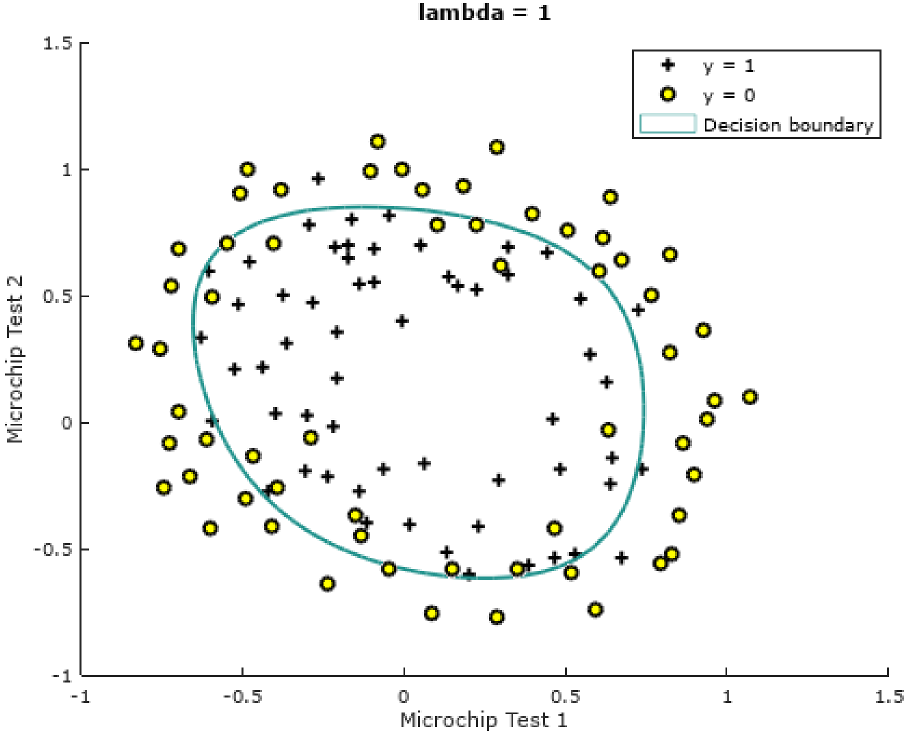


$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

Example



Admitted
Not admitted
Exam 2 score
Exam 1 score
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100



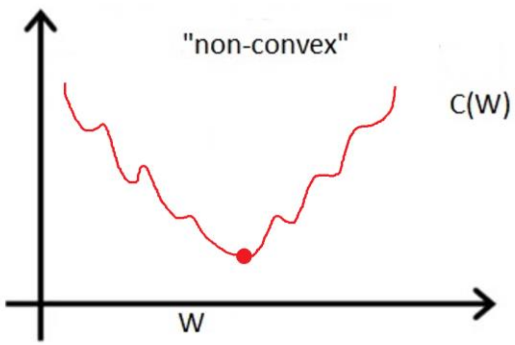
lambda = 1
y = 1
y = 0
Decision boundary
Microchip Test 2
Microchip Test 1
1.5
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1
1.5

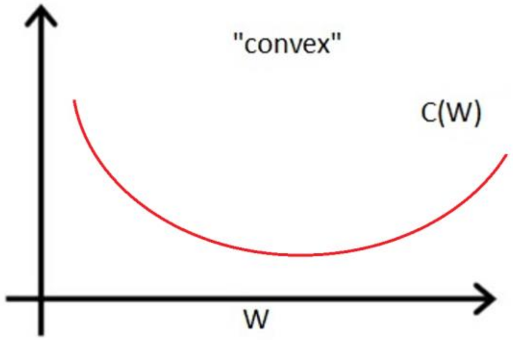
Cost function – Logistic Regression

- We cannot use the same cost function that we use for linear regression because the Logistic Function will cause the output to be wavy, causing many local optima and it will not be a convex function

$$C = \frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i)^2$$

?



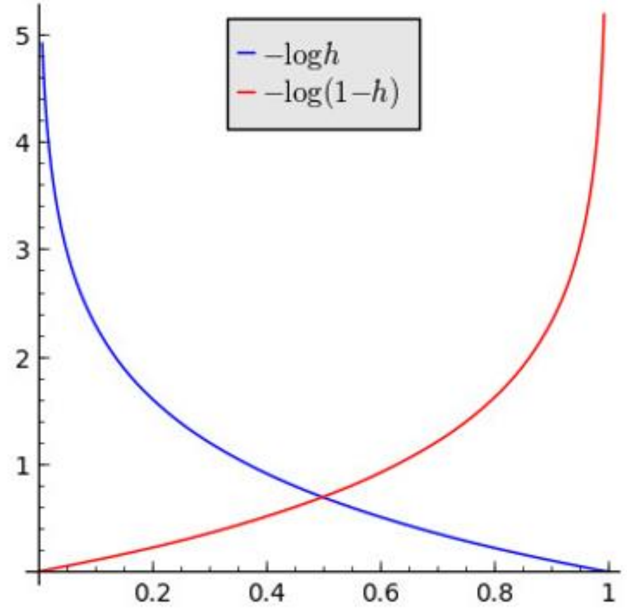


Cost function – Logistic Regression

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if y} = 1$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if y} = 0$$

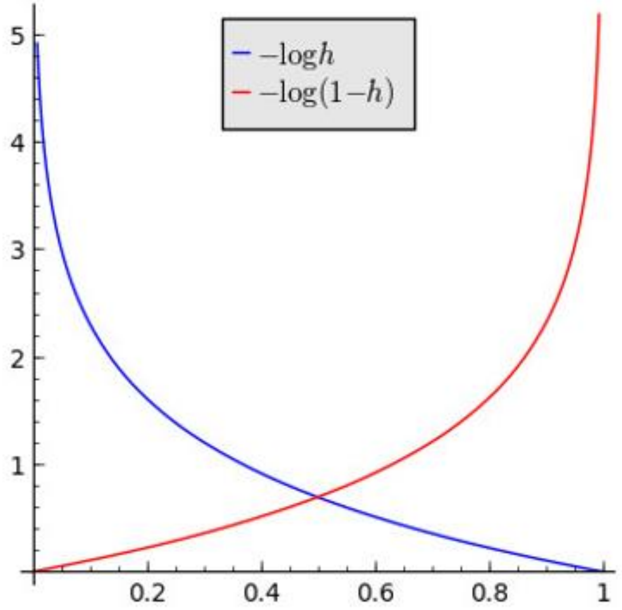


Cost function – Logistic Regression

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(x^{(i)})) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right],$$

and the gradient of the cost is a vector of the same length as θ where the j^{th} element (for $j = 0, 1, \ldots, n$) is defined as follows:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

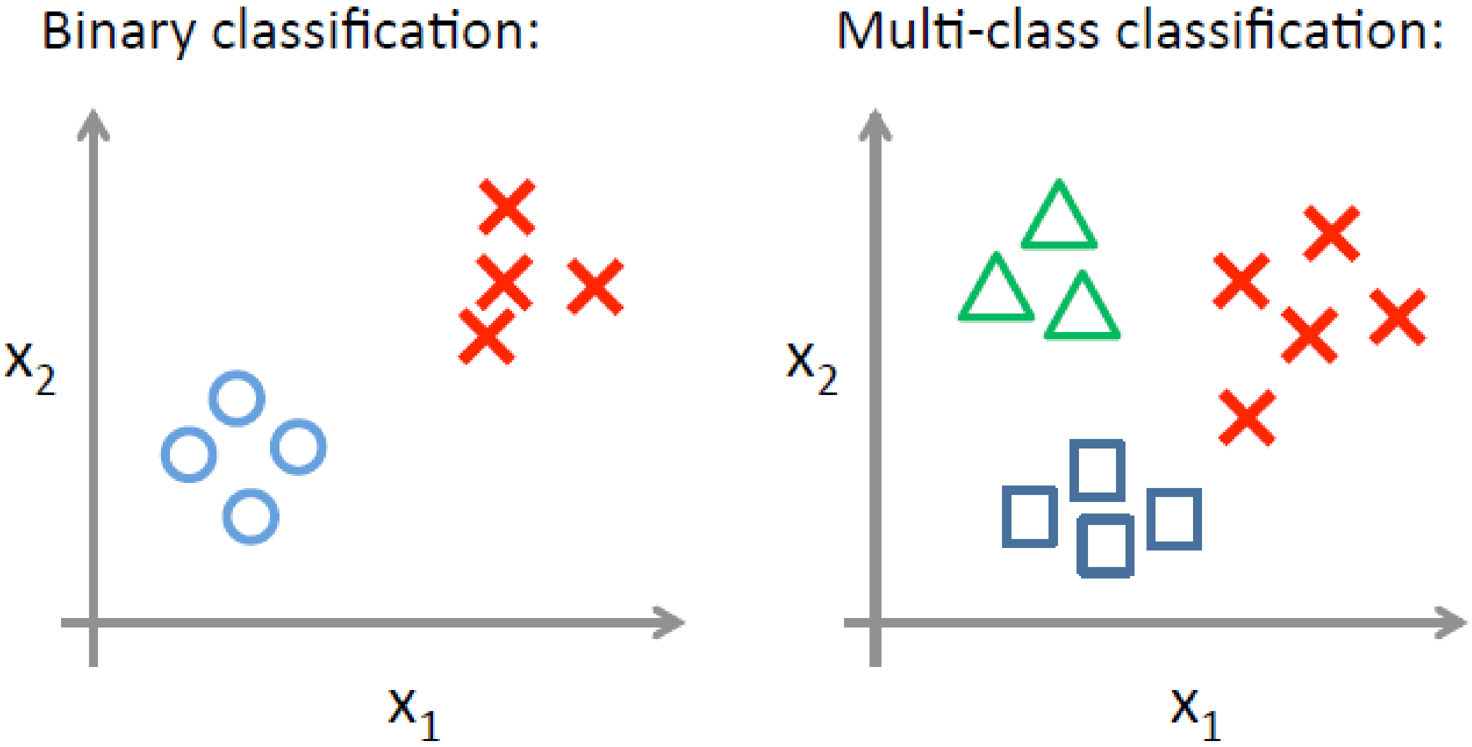


Optimization

- Optimization algorithms:
  - Gradient descent
  - Conjugate gradient
  - BFGS
  - L-BFGS

- Advantages:
  - No need to manually pick α
  - Often faster than gradient descent
- Disadvantages:
  - More complex

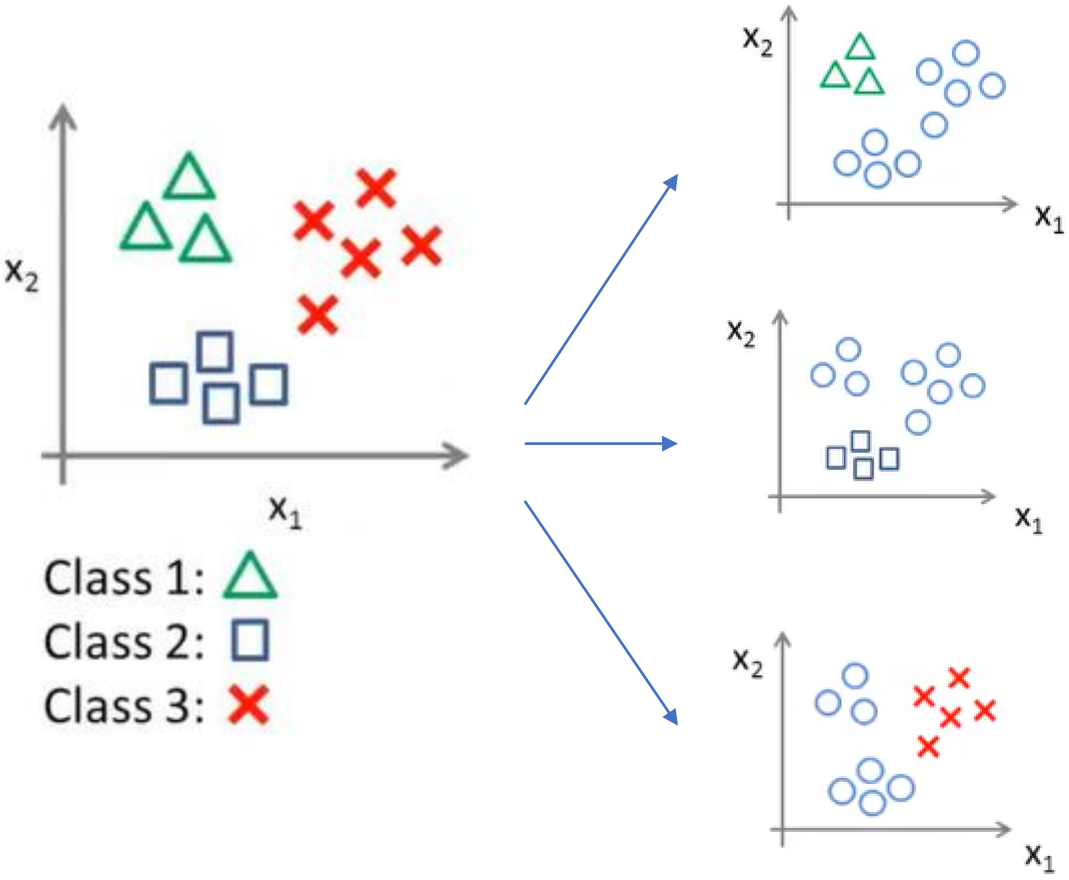
Multi Class Classification



Binary classification:

Multi-class classification:

Multi Class Classification



- Train a logistic regression classifier for each class *i* to predict the probability that *y* = *i*.
- On a new input *x*, to make a prediction, pick the class *i* that maximizes.

$$
\begin{aligned}
& y \in \{0, 1 \ldots n\} \\
& h_\theta^{(0)}(x) = P(y=0|x;\theta) \\
& h_\theta^{(1)}(x) = P(y=1|x;\theta) \\
& \cdots \\
& h_\theta^{(n)}(x) = P(y=n|x;\theta) \\
& \mathrm{prediction} = \max_i (h_\theta^{(i)}(x))
\end{aligned}
$$

ELTE | FACULTY OF INFORMATICS

Thank you for your attention!