WEEKS 5-9

Introduction to Machine Learning

Dr Mykola Gordovskyy

Week 7

- Weighting
- Support Vector Machines
- Decision trees
- Supervised learning pipeline

Weighting: why?

- Some data entries may be more or less reliable than others – use of different measurement tools, a large dataset created from smaller datasets obtained in different ways etc etc

- We may want to make some data entries more or less important based on their properties

Weighting

Example: weighting in regression

Simple linear regression with weighting

Find the values of a and b so that

$$\frac{\partial E}{\partial a} = \frac{\partial}{\partial a}\sum_i (Y_i - a - bXi)^2 = -2\sum_i \boxed{(Y_i - a - bXi) = 0}$$

and

$$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b}\sum_i (Y_i - a - bXi)^2 = -2\sum_i \boxed{X_i(Y_i - a - bXi) = 0}$$

$$b = \frac{\sum_i X_i Y_i - \frac{1}{N}\sum_i X_i \sum_i Y_i}{\sum_i {X_i}^2 - \frac{1}{N}\left(\sum_i X_i\right)^2}$$

$$a = \frac{1}{N}\sum_i Y_i - b\frac{1}{N}\sum_{Ni} X_i$$

Example: weighting in regression

Simple linear regression with weighting

Find the values of a and b so that

$$\frac{\partial E}{\partial a} = \frac{\partial}{\partial a}\sum_i W_i (Y_i - a - bXi)^2 = -2\boxed{\sum_i W_i (Y_i - a - bXi) = 0}$$

and

$$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b}\sum_i W_i (Y_i - a - bXi)^2 = -2\boxed{\sum_i W_i X_i (Y_i - a - bXi) = 0}$$

$$\boxed{\begin{aligned} b &= \frac{\sum_i W_i X_i Y_i - \frac{1}{\sum_i W_i}\sum_i W_i X_i \sum_i W_i Y_i}{\sum_i W_i {X_i}^2 - \frac{1}{\sum_i W_i}\left(\sum_i W_i X_i\right)^2} \\ a &= \frac{1}{\sum_i W_i}\sum_i W_i Y_i - b\frac{1}{\sum_i W_i}\sum_{Ni} W_i X_i \end{aligned}}$$

Example: weighting in regression

Linear regression with weighting

$$b = \frac{\sum_i W_i X_i Y_i - \frac{1}{\sum_i W_i} \sum_i W_i X_i \sum_i W_i Y_i}{\sum_i W_i X_i^{\,2} - \frac{1}{\sum_i W_i} \left(\sum_i W_i X_i\right)^2}$$

$$a = \frac{1}{\sum_i W_i} \sum_i W_i Y_i - b \frac{1}{\sum_i W_i} \sum_{Ni} W_i X_i$$

$$\xrightarrow{W_i = 1}$$

$$b = \frac{\sum_i X_i Y_i - \frac{1}{N} \sum_i X_i \sum_i Y_i}{\sum_i X_i^{\,2} - \frac{1}{N} \left(\sum_i X_i\right)^2}$$

$$a = \frac{1}{N} \sum_i Y_i - b \frac{1}{N} \sum_{Ni} X_i$$

Weighting in kNN

- The effect of neighbours may change with distance. We might want neighbours which are further away to have lower significance. How do we do this? – Weighting!

- $P_t = \text{sign}(\sum_{j=0}^{k-1} W_j P_{i(j)})$

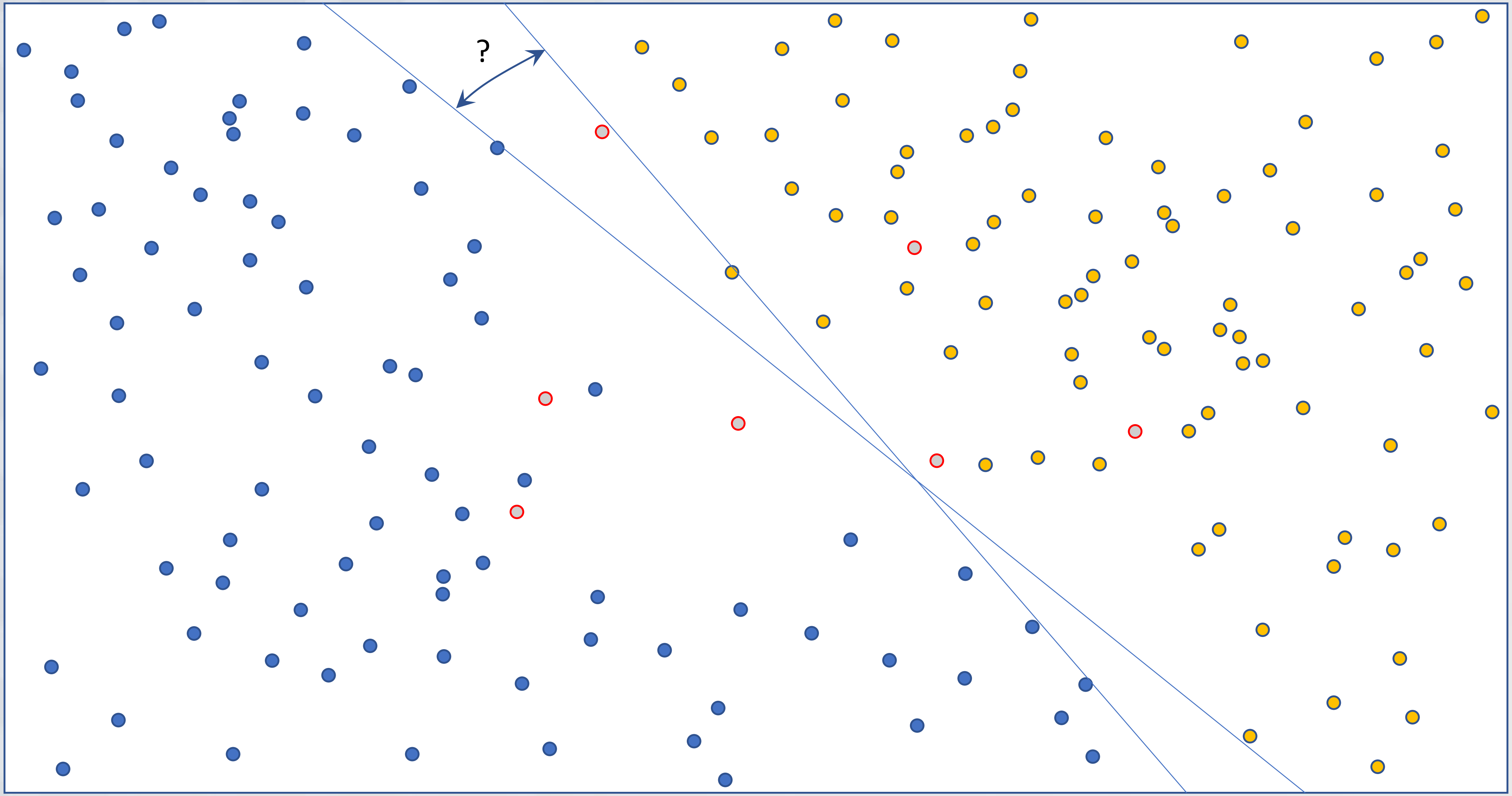
- E.g. $P_t = \text{sign}(\sum_{j=0}^{k-1} e^{-L_j} P_{i(j)})$ - in this case weights exponentially decrease with distance to the corresponding neighbour

Weighting in kNN

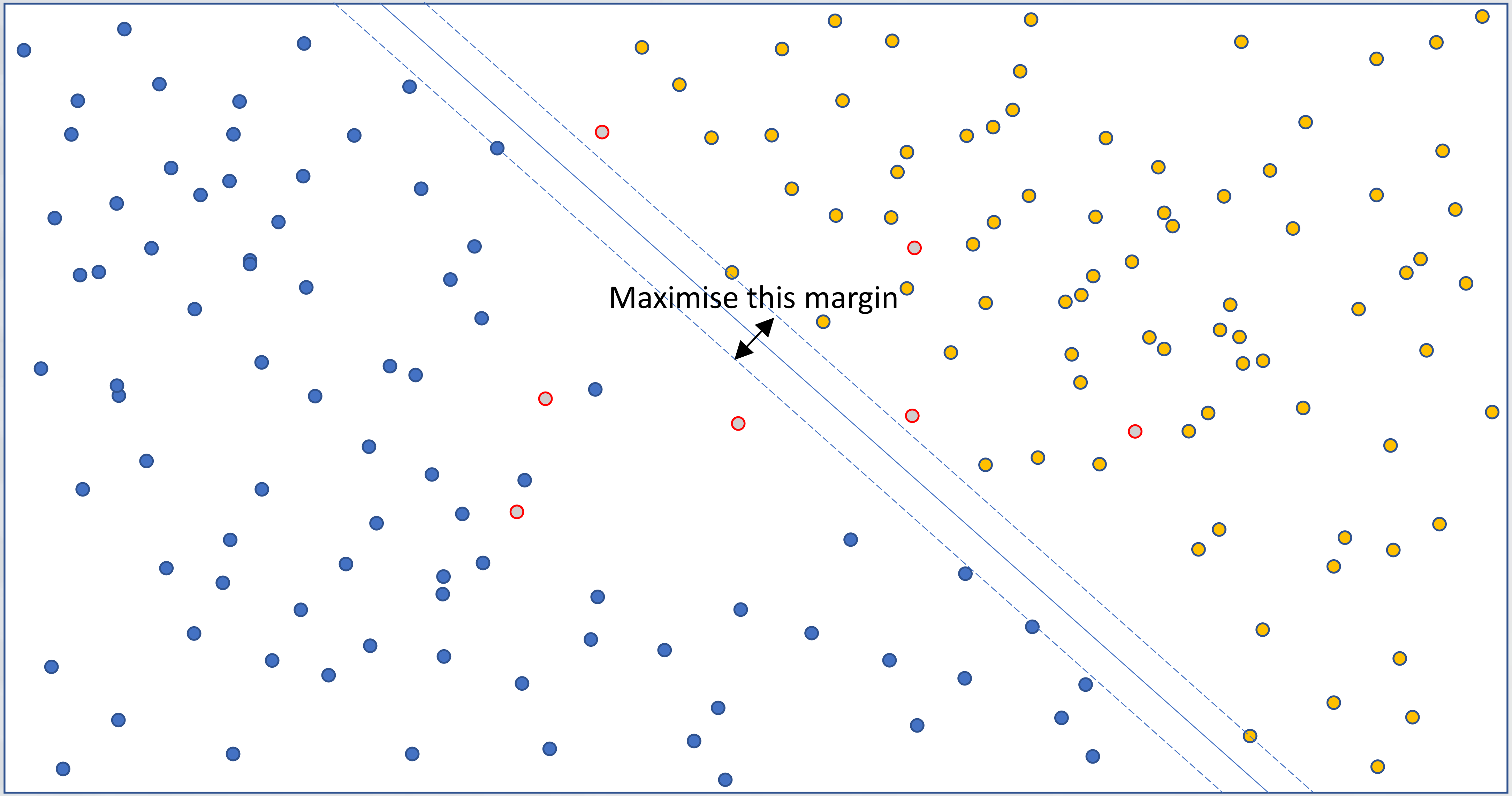
- The effect of neighbours may depend on their reliability. We can give different ‘voting weight’ to each training data entry

- $P_t = \text{sign}(\sum_{j=0}^{k-1} W_{i(j)} P_{i(j)})$ – in this case each entry in the training data has its own weight

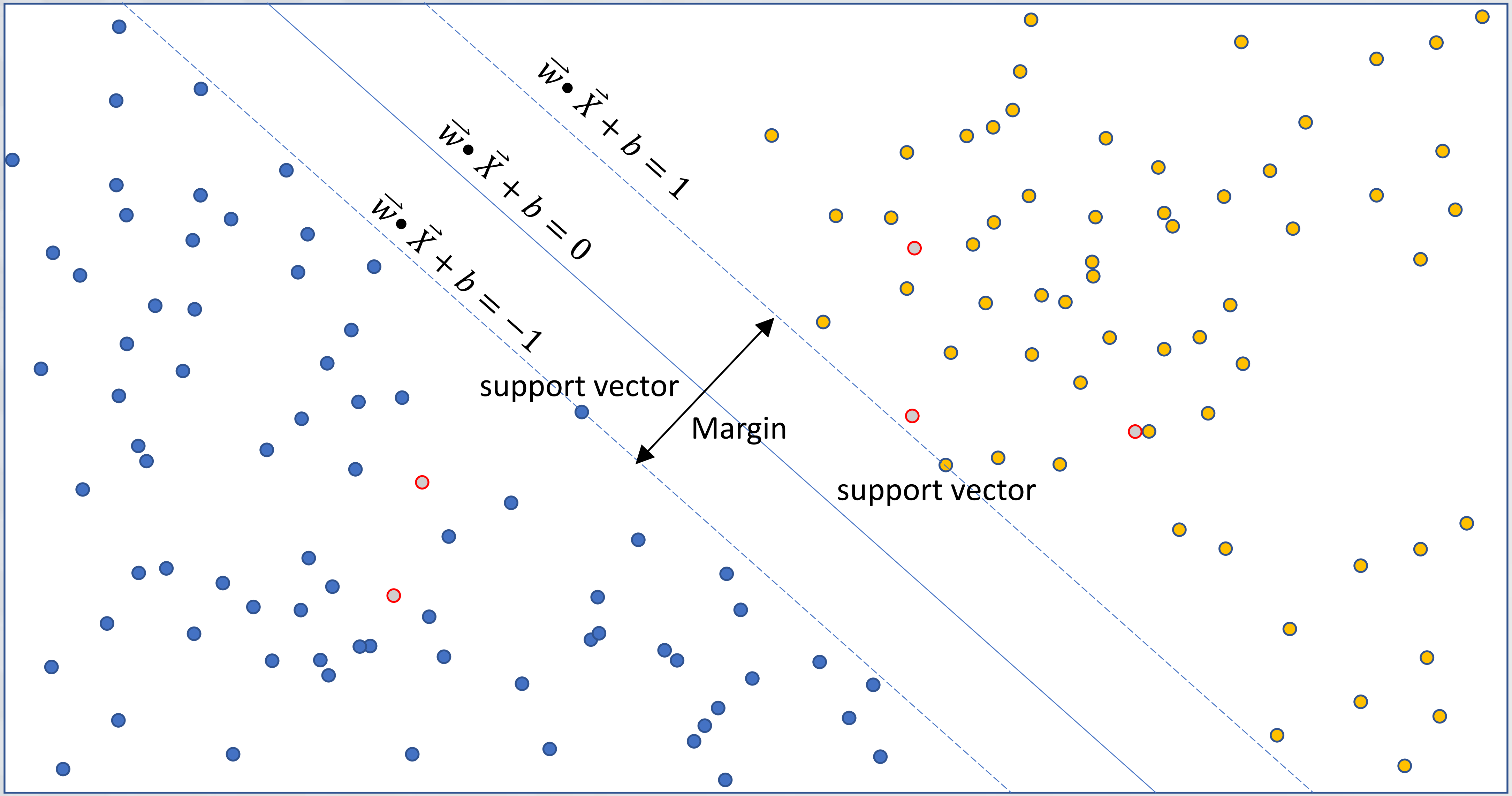
Example



Example



Support Vector Machine (SVM)



Support Vector Machine (SVM)

$$S = \begin{cases} +1 & \text{if } \vec{X}.\vec{w} + \mathrm{b} \geq 0 \\ -1 & \text{if } \vec{X}.\vec{w} + \mathrm{b} < 0 \end{cases}$$

- What is $\vec{w} \bullet \vec{X} + b = 0$?

  - this is another way of representing a line

$$\vec{X} = [x, y]$$
$$\vec{w} = [w_1, w_2]$$

$$xw_1 + yw_2 + b = 0$$

$$y = -\frac{w_1}{w_2}x - \frac{b}{w_2}$$

Support Vector Machine (SVM)

- We have two lines

$$xw_1 + yw_2 + b + 1 = 0$$

$$xw_1 + yw_2 + b - 1 = 0$$

- The distance Δ between them is

$$\Delta = \frac{2}{\sqrt{w_1^2 + w_1^2}} = \frac{2}{|\vec{w}|}$$

- **Maximise Δ = minimise absolute value of $\vec{w}$**

Support Vector Machine (SVM)

- **Maximise Δ = minimise absolute value of $\vec{w}$**

- Taking into account that

$$\vec{w} \bullet \vec{X} + b \geq 1 \quad \text{when S =1}$$
$$\vec{w} \bullet \vec{X} + b \leq 1 \quad \text{when S =-1}$$

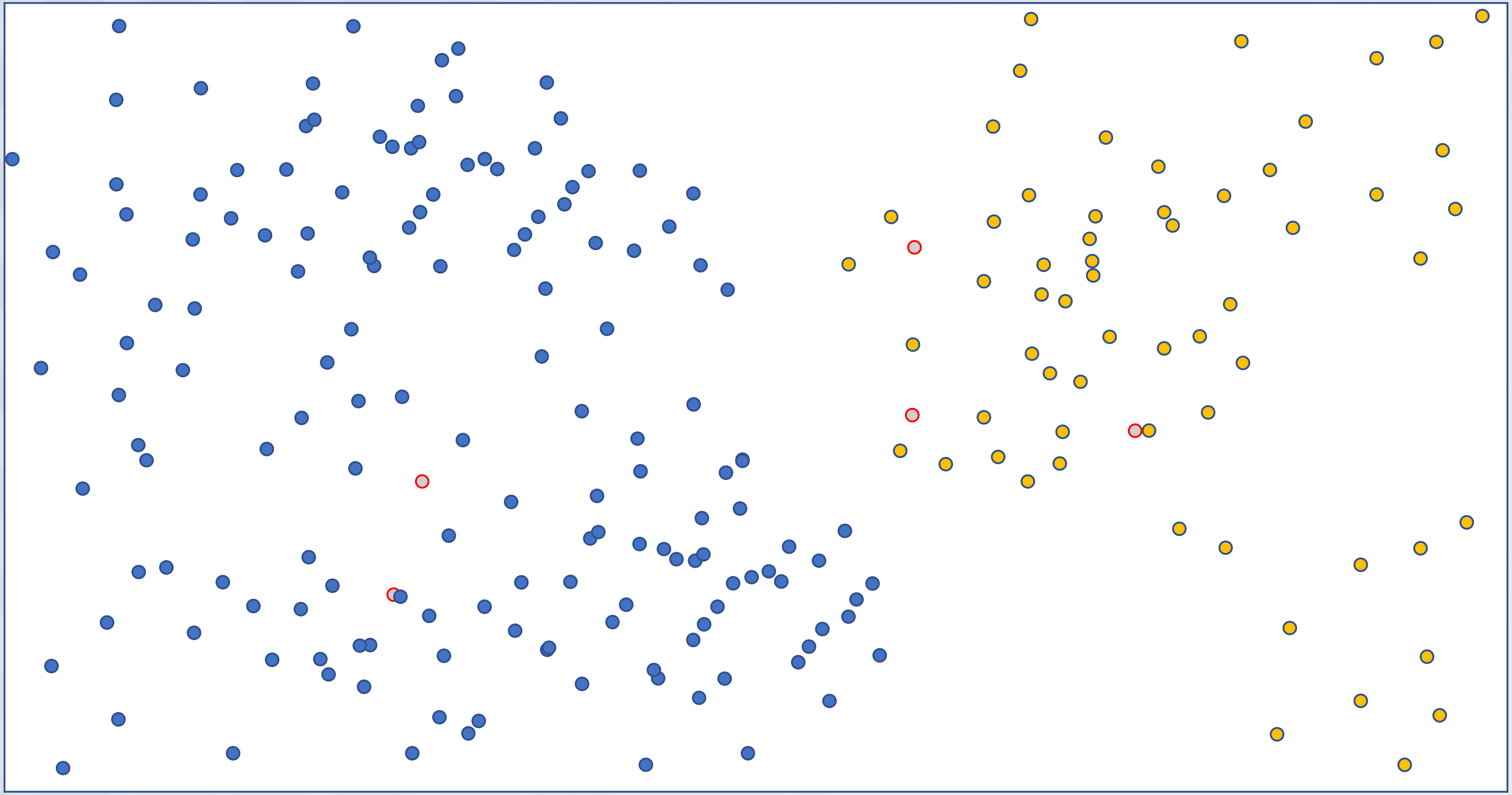
we can write

$$\mathrm{S}_{\mathrm{i}}(\vec{w} \bullet \vec{X_i}) \geq 1$$

Support Vector Machine (SVM)

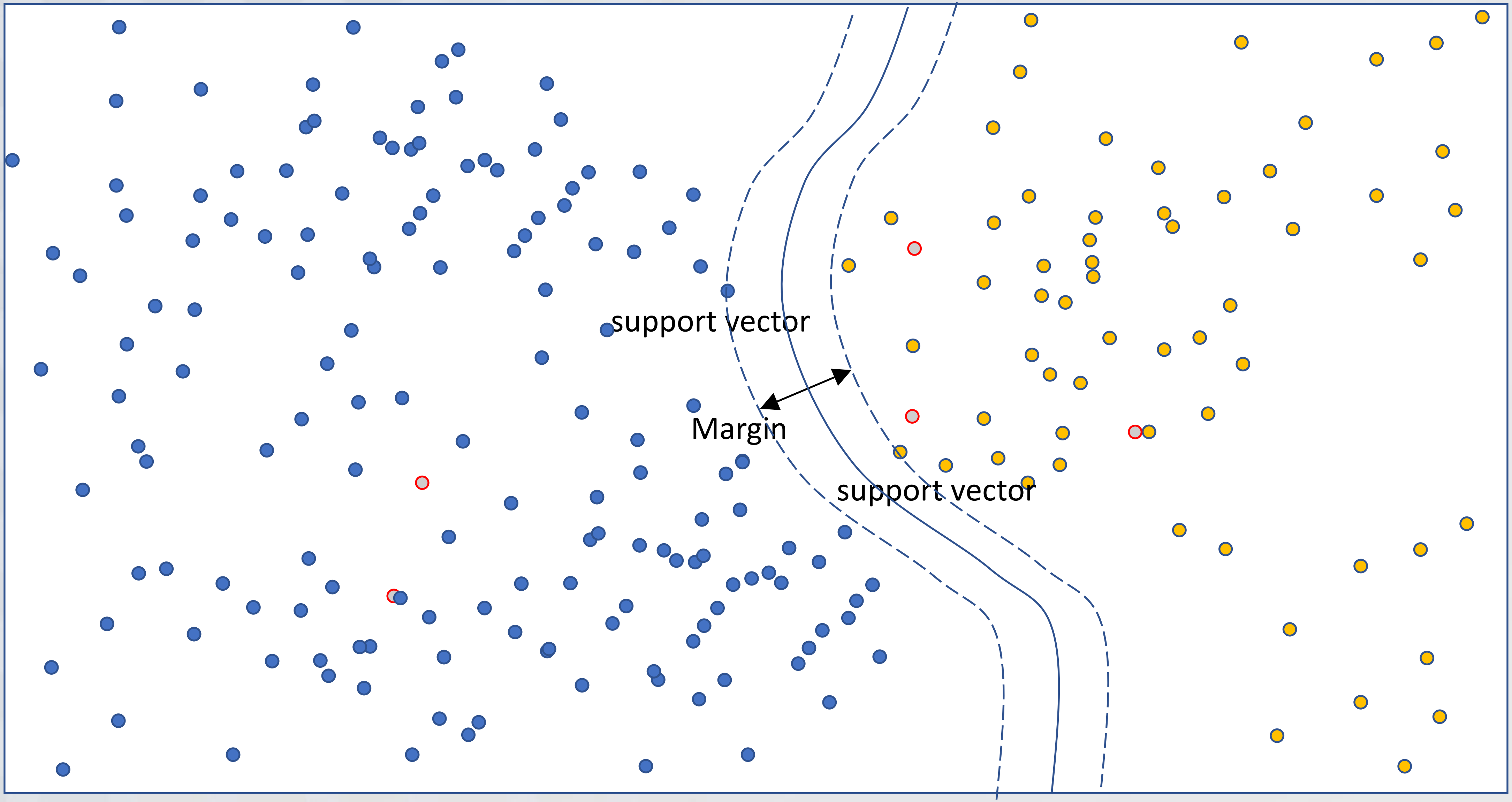
- Linear = separated by line
- Non-linear = separated by curve

Support Vector Machine (SVM)



Support Vector Machine (SVM)

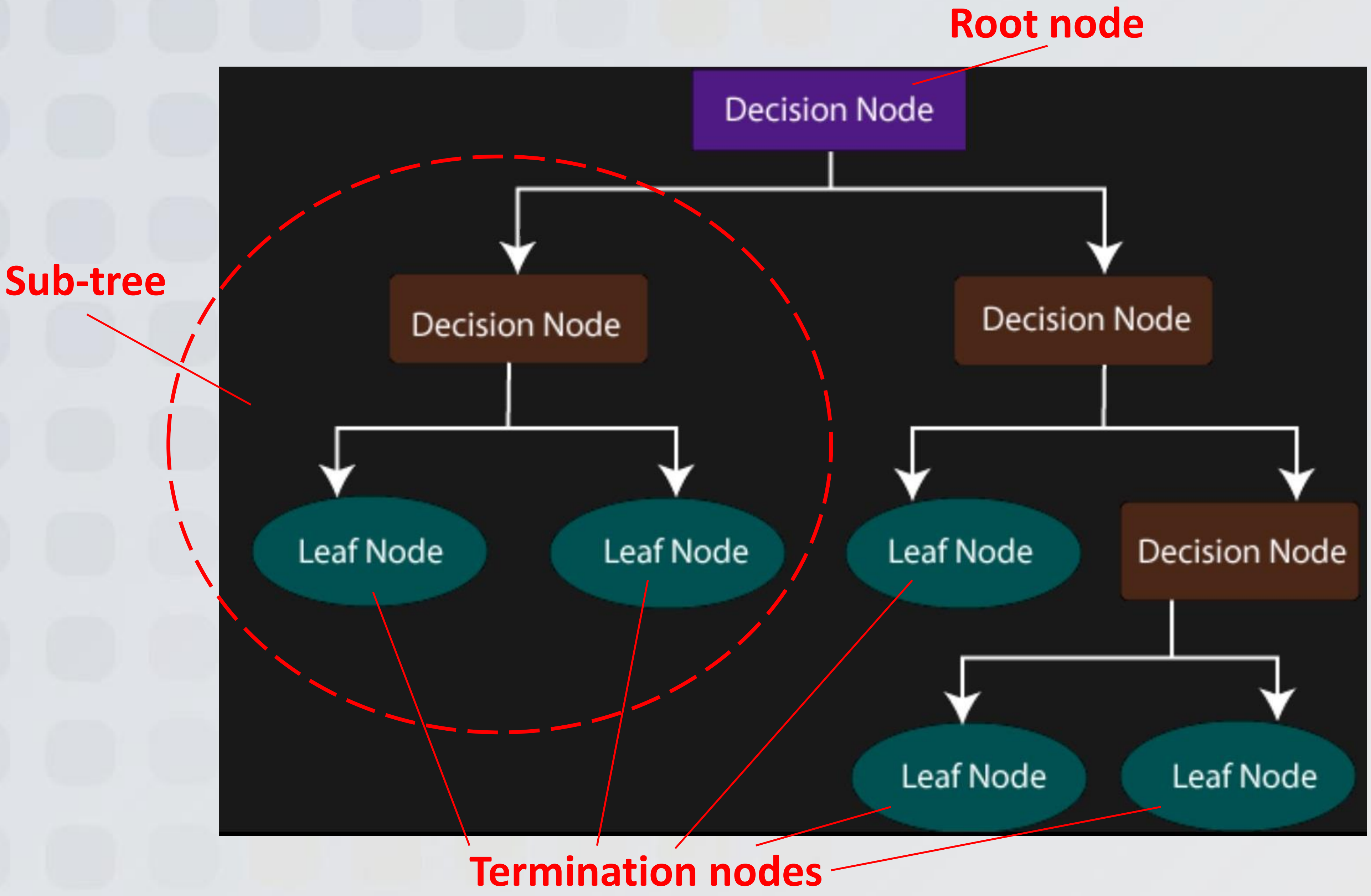
$$F(\vec{w} \bullet \vec{X} + \vec{w} \bullet \overrightarrow{X^2} + \cdots) \geq 1$$



Support Vector Machine (SVM)

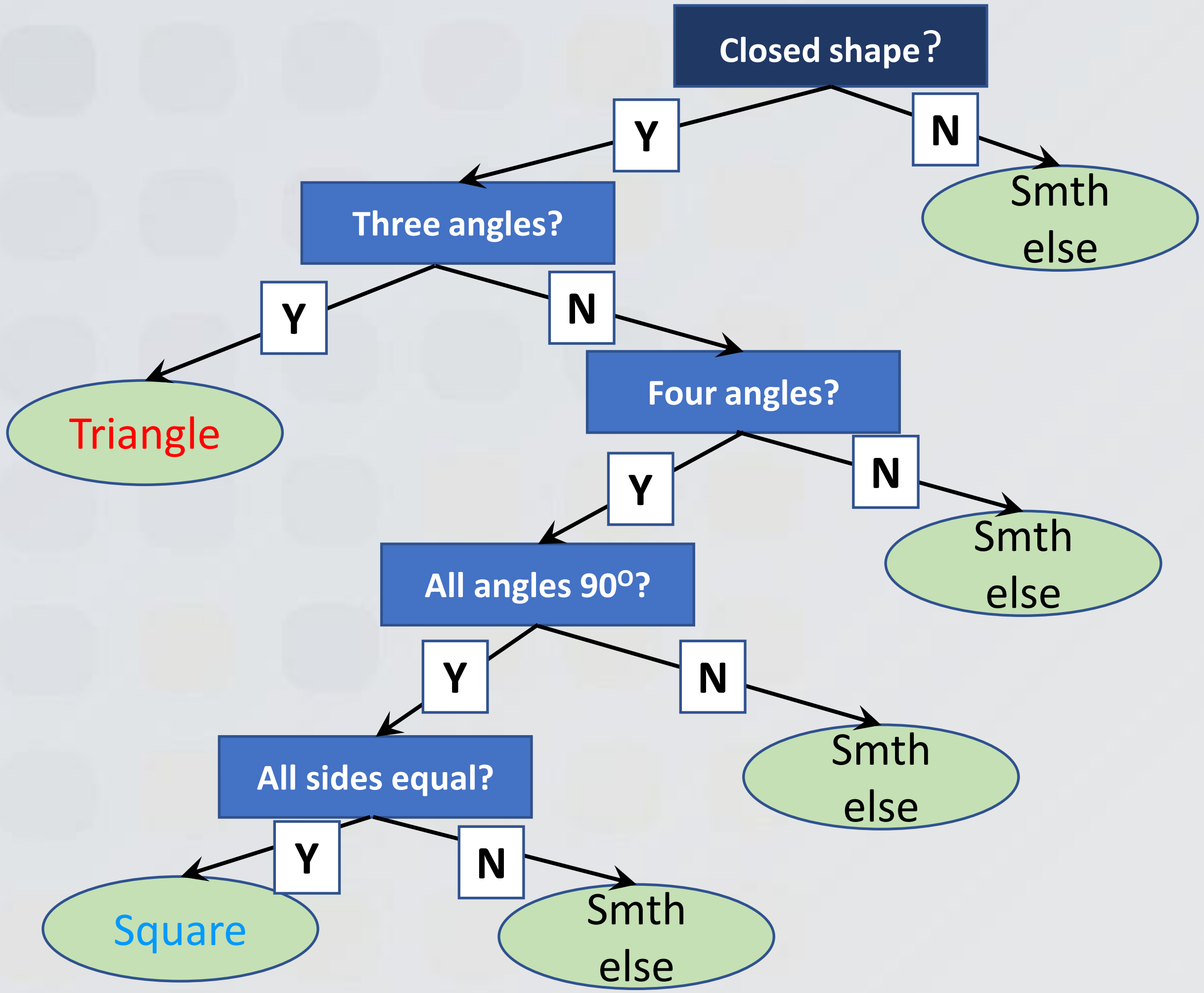
- Hard margin = clear separation between classes
- Soft margin = no clear separation, some entries will be in “the band”

Decision Tree Classifier

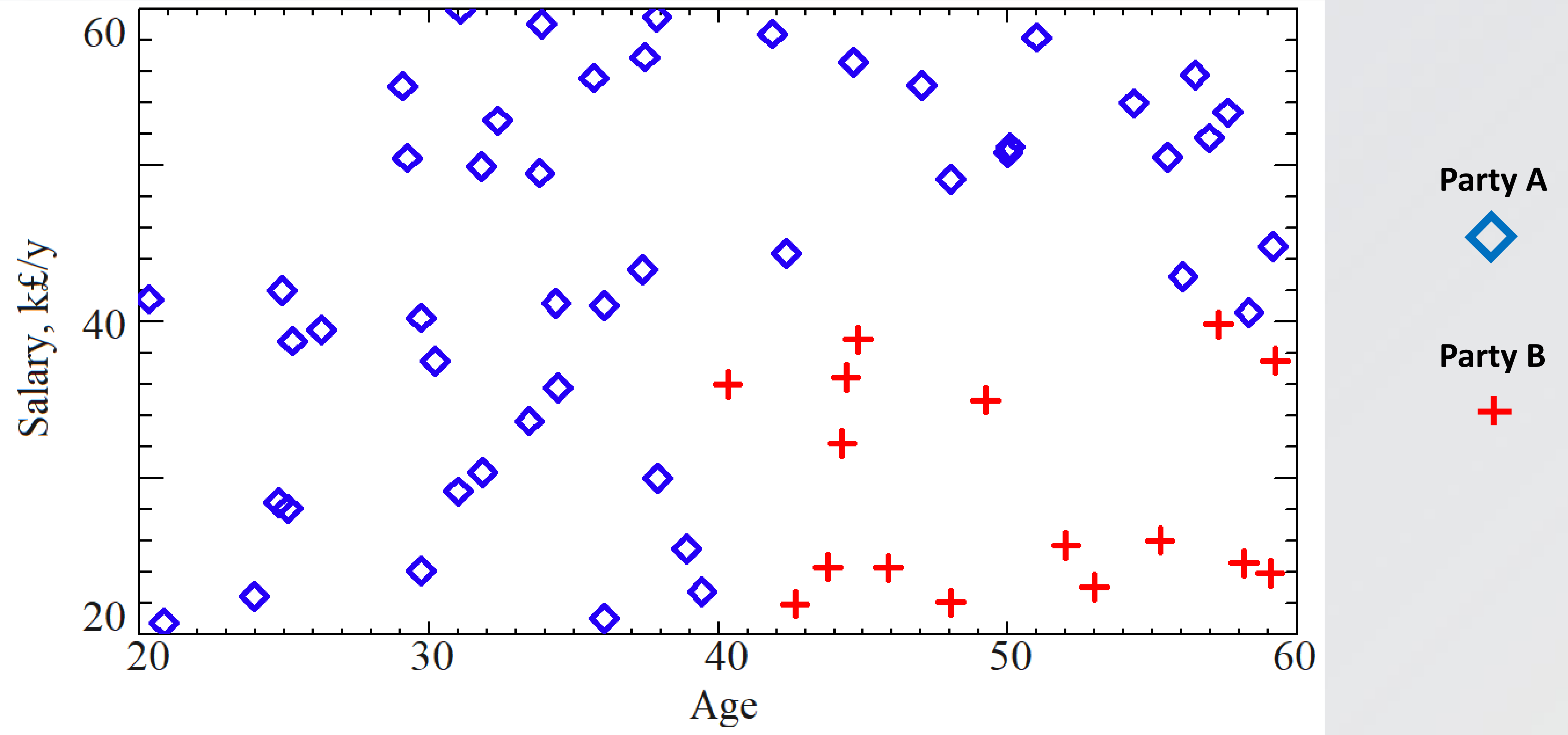


Example 1

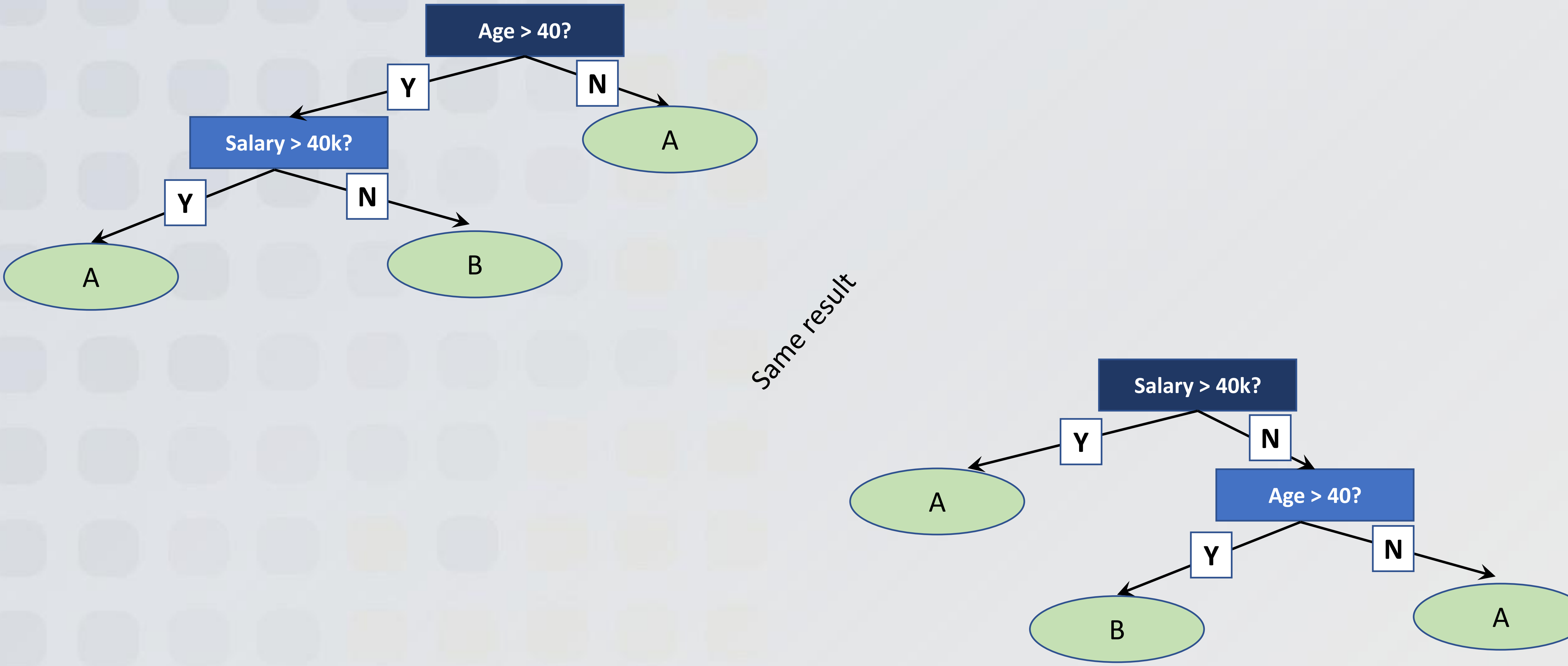
Triangle: has three angles and three sides
Square: has four angles and four sides, all angles are right angles, all sides are of equal length



Example 2

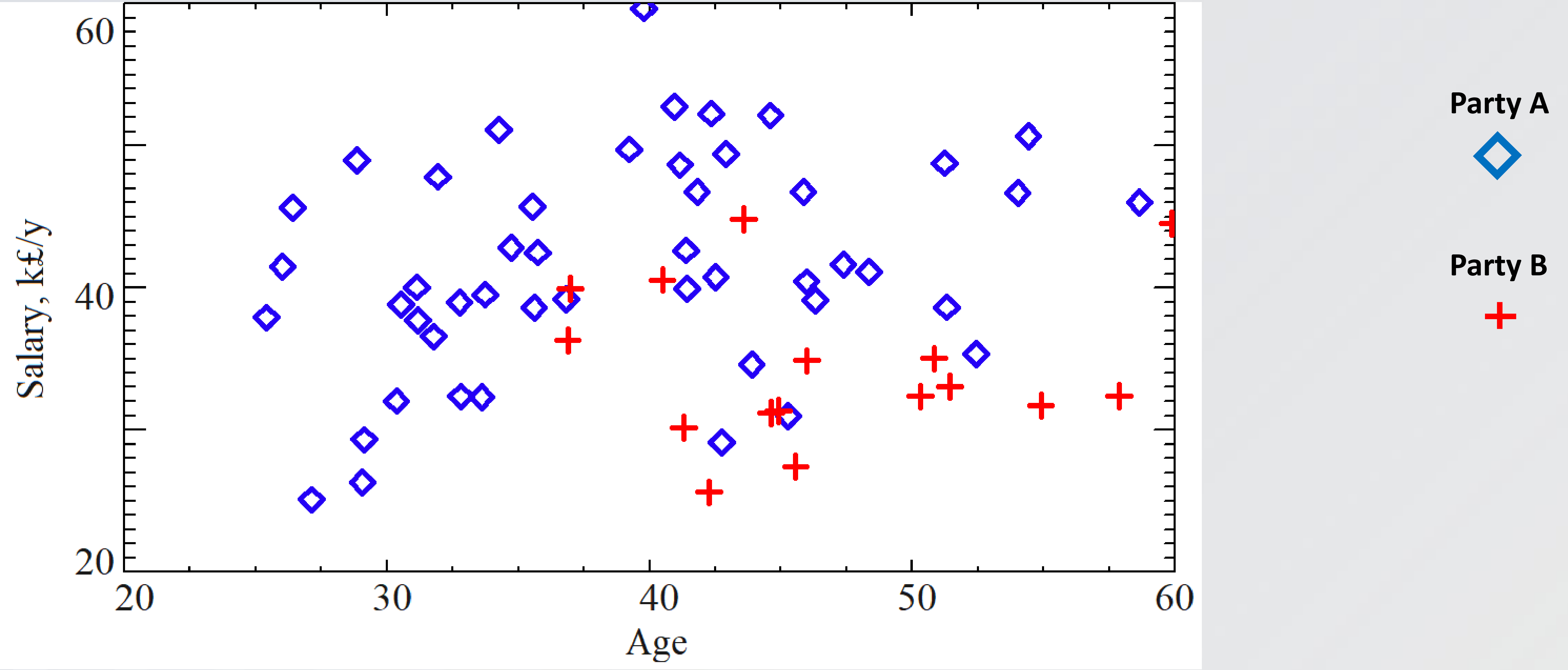


Example 2

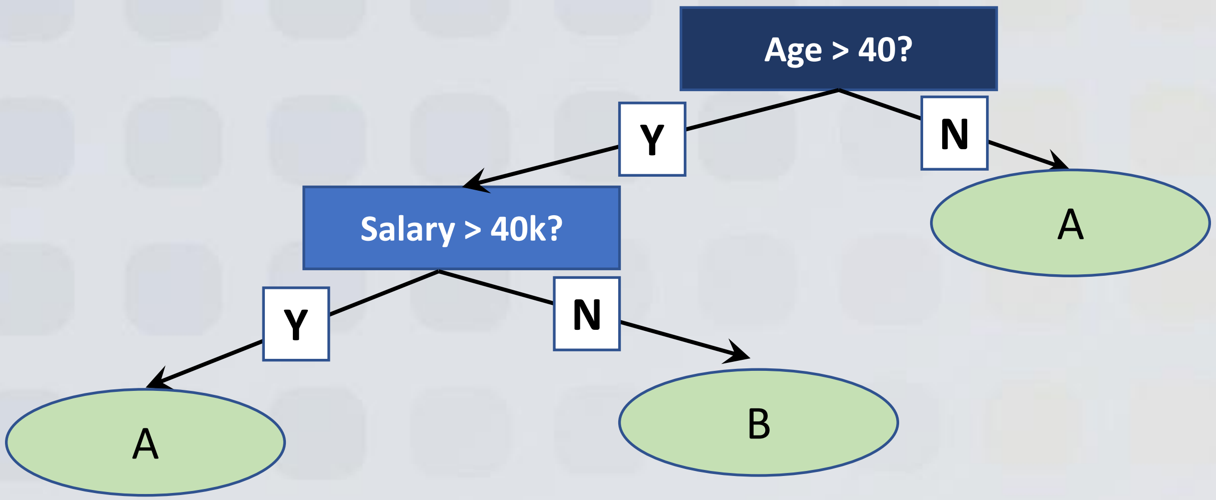


Age > 40?
Y
N
Salary > 40k?
A
Y
N
A
B
Same result
Salary > 40k?
Y
N
A
Age > 40?
Y
N
B
A

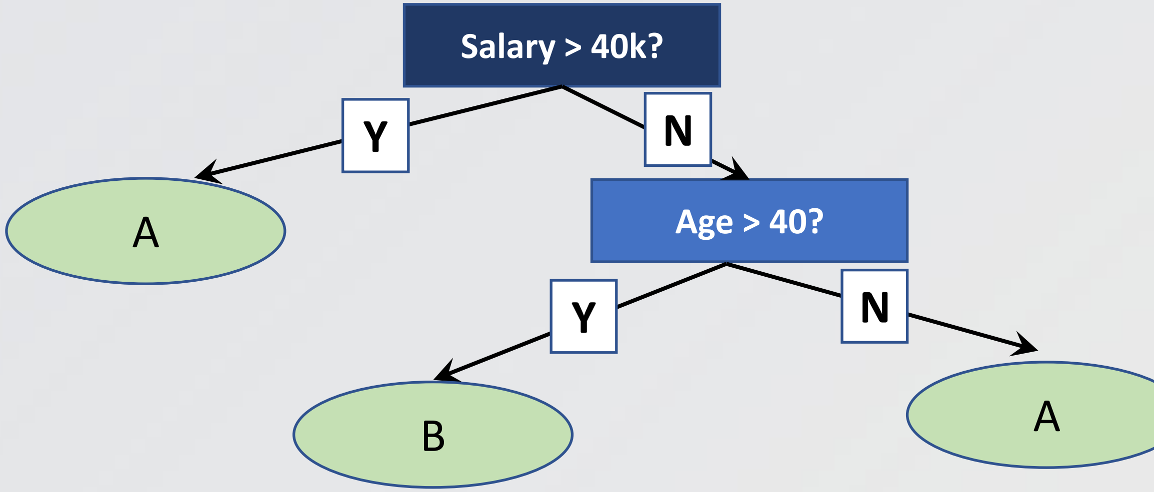
Example 2a



Example 2a



Different result



Recursive binary splitting

Split in terms of A, B or C?

B

OK, split for B

Split in terms of A or C?

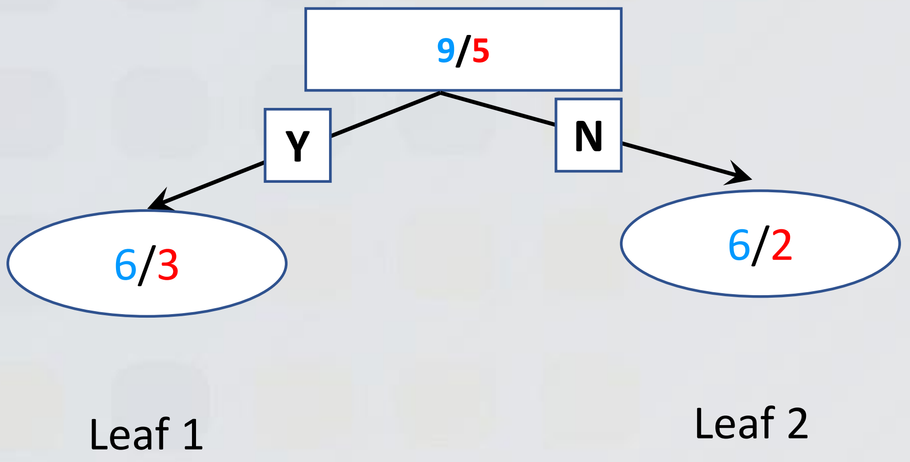
A

OK, split for A

Finally, split for C

Entropy

$$H = -\sum_{i=1}^{K} p_i \log_2(p_i)$$



Information gain

Information Gain = Information entropy (parent) – Information entropy (child split)

Information entropy (child split) = Fraction1 * Entropy1 + Fraction2 * Entropy2

Split criteria: maximise information gain