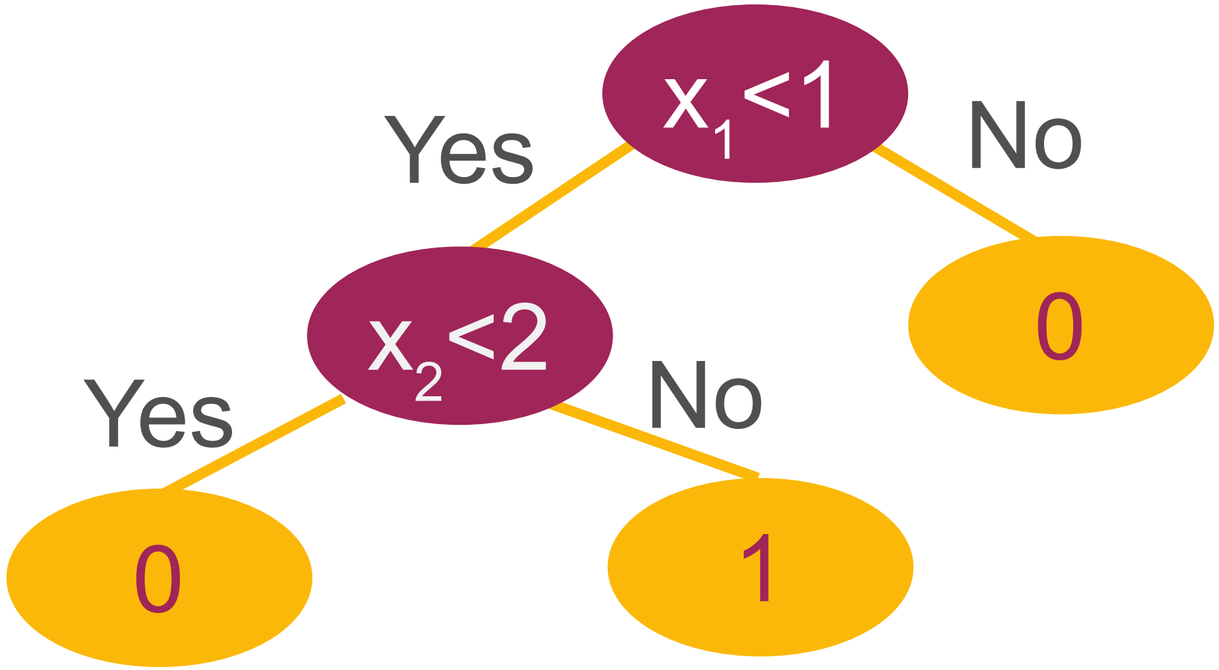
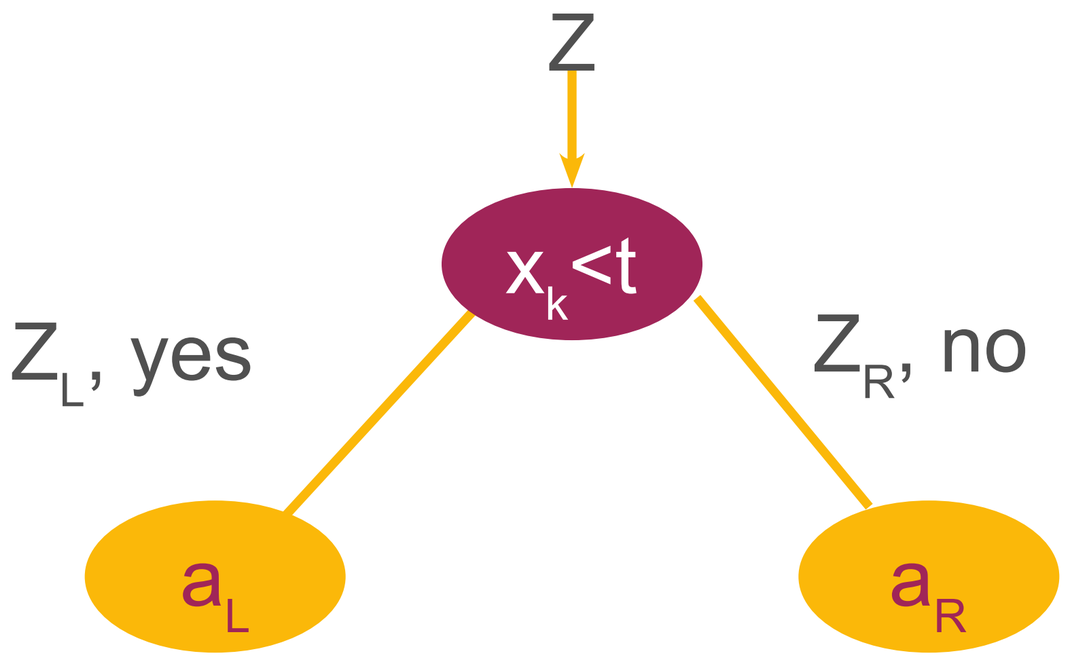
Decision Trees
for Classification

Decision Tree for Classification

Class labels in leaves



How to find the best split



Find the best split:

Maximize the information gain (IG):

$$IG = \mathbf{Impurity}(Z) - \left(\frac{|Z_L|}{|Z|} \mathbf{Impurity}(Z_L) + \frac{|Z_R|}{|Z|} \mathbf{Impurity}(Z_R) \right)$$

with respect to k, t (splitting criteria x_k<t)

Decision Tree for Classification

Gini impurity $= \sum_{i=1}^{C} f_i(1 - f_i)$

f_i – frequency of label ***i*** in node,
C – number of classes

Gini impurity(Z_L) $= \sum_{i=1}^{C} f_{i,L}(1 - f_{i,L})$

$$f_{i,L} = \frac{number\ of\ examples\ in\ Z_L with\ label\ i}{|Z_L|}$$

Binary classification (2 classes)

Gini impurity = $2f_1\ (1-f_1)$

(because $f_0+f_1=1$)

f_1 – frequency of class 1

Gini impurity, binary classification