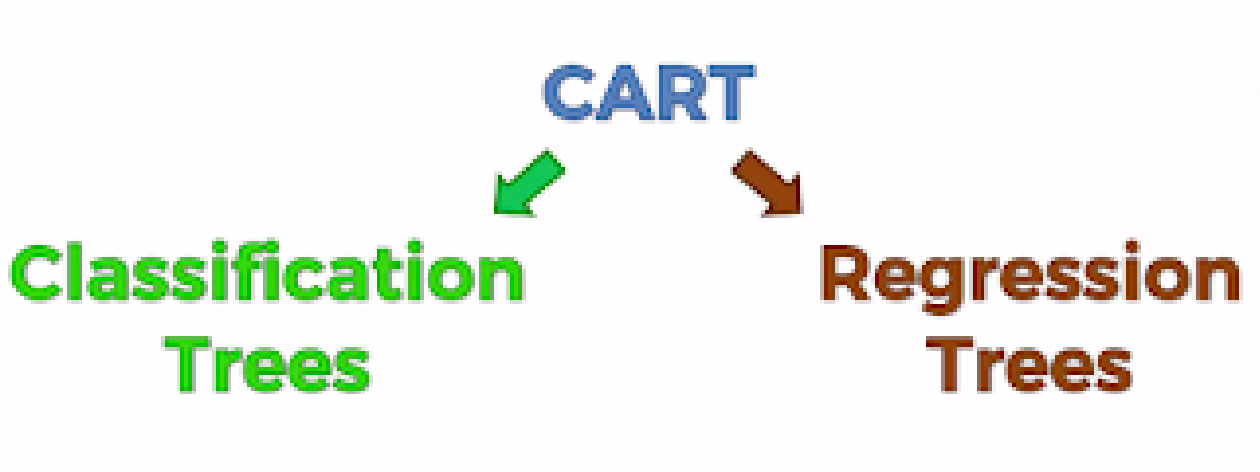
分类与回归树



赵海涛

haitaozhao@ecust.edu.cn

Josh Gordon

random-forests

Unfollow

Git is complicated ¯_(ツ)_/¯

NYC

twitter.com/random_forests

Block or report user

Organizations

CART

Training Data

Color	Diameter	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Training Data

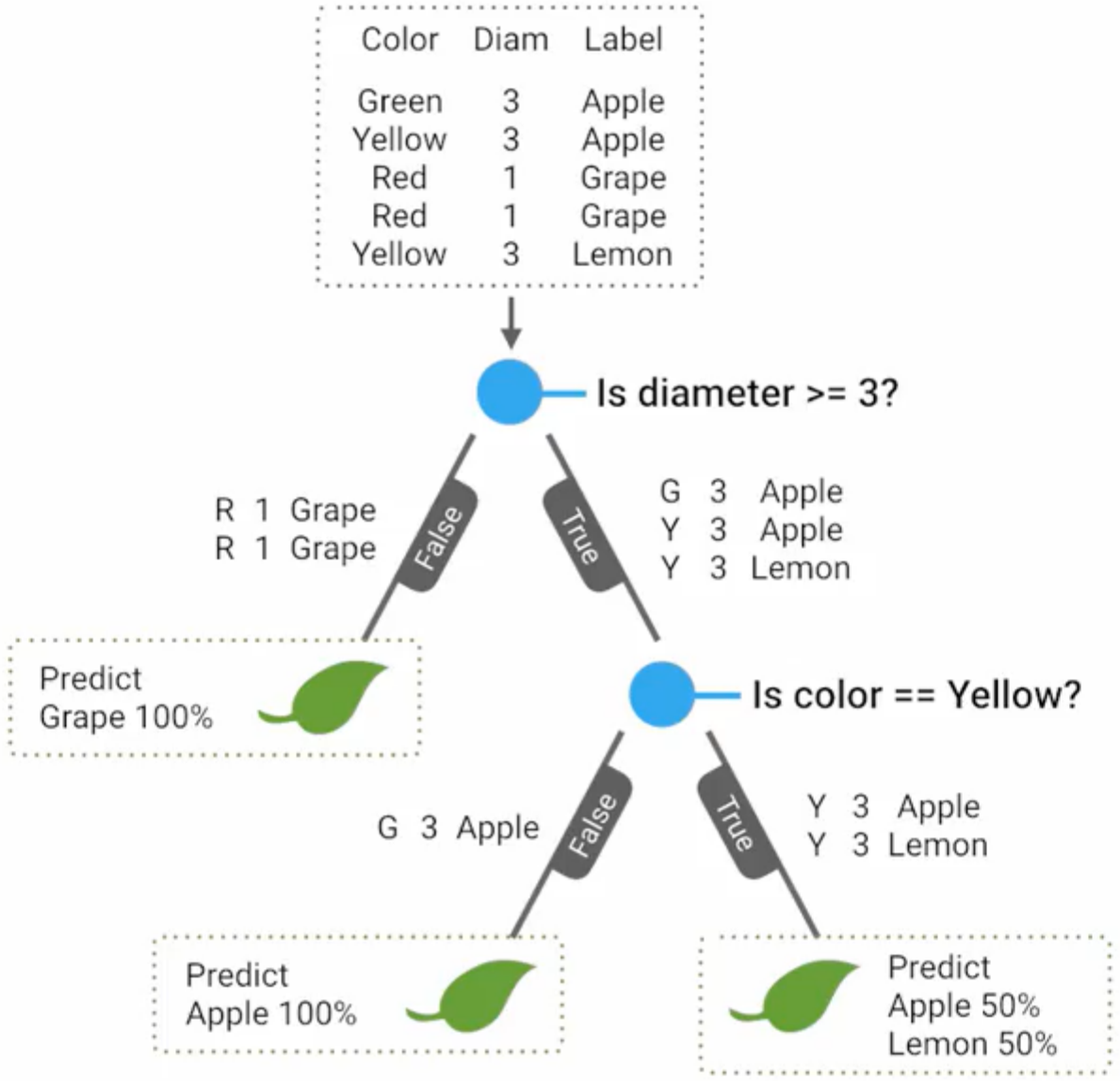
Color	Diameter	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

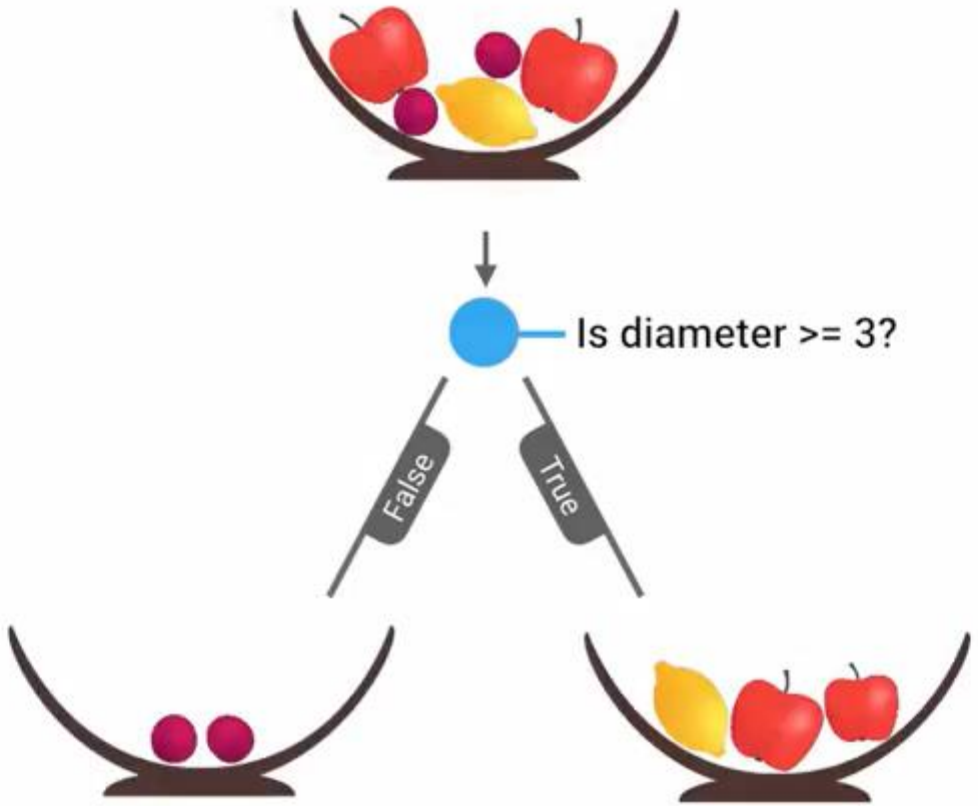
A categorical attribute

Training Data

Color	Diameter	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

A numeric attribute


Color Diam Label
Green 3 Apple
Yellow 3 Apple
Red 1 Grape
Red 1 Grape
Yellow 3 Lemon
Is diameter >= 3?
False
True
R 1 Grape
R 1 Grape
G 3 Apple
Y 3 Apple
Y 3 Lemon
Predict
Grape 100%
Is color == Yellow?
False
True
G 3 Apple
Y 3 Apple
Y 3 Lemon
Predict
Apple 100%
Predict
Apple 50%
Lemon 50%


Is diameter >= 3?
False
True

Possible questions

Is the color green?

Is the diameter >=3?

Is the color yellow?

TRUE

Green 3 Apple

FALSE

Yellow 3 Apple

Yellow 3 Lemon

Green	3	Apple
Yellow	3	Apple
Yellow	3	Lemon

Possible questions

- Is the color green?
- Is the diameter >=3?
- Is the color yellow?

TRUE

FALSE

信息增益(Information Gain)

- 定义5.2 (信息增益):特征A对训练数据集D的信息增益，$g(D,A)$，定义为集合D的经验熵$H(D)$与特征A给定条件下D的经验条件熵$H(D|A)$之差，即

$$g(D,A) = H(D) - H(D|A)$$

- (Information gain)表示得知特征X的信息而使得类Y的信息的不确定性减少的程度。
- 一般地，熵$H(Y)$与条件熵$H(Y|X)$之差称为互信息（mutual information)
- 决策树学习中的信息增益等价于训练数据集中类与特征的互信息。

信息增益(Information Gain)

- 输入：训练数据集D和特征A;
- 输出：特征A对训练数据集D的信息增益$g(D,A)$

1、计算数据集D的经验熵$H(D)$

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}$$

2、计算特征A对数据集D的经验条件熵$H(D|A)$

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

3、计算信息增益

$$g(D,A)=H(D)-H(D|A)$$

信息增益(Information Gain)

- 输入：父节点数据集D_p和对其的划分得到的两个子数据集 D_{left}， D_{right};
- 输出：信息增益$IG(D_p)$

计算信息增益

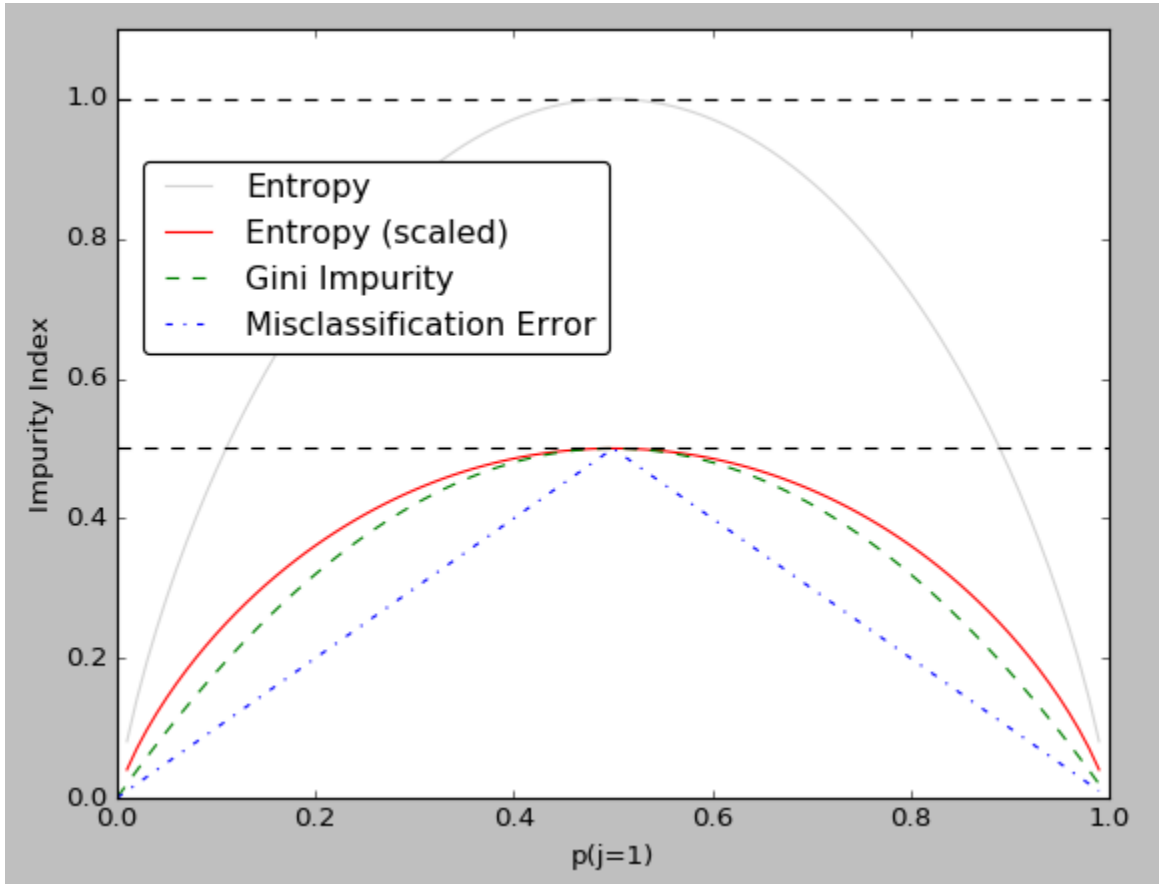
$$IG(D_p) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

- 可能的不纯度指数

1. Entropy: $\sum_{j=1}^{K} p_j \ln \frac{1}{p_j}$.
2. Misclassification rate: $1 - \max_j p_j$.
3. Gini index: $\sum_{j=1}^{K} p_j(1-p_j) = 1 - \sum_{j=1}^{K} p_j^2$.

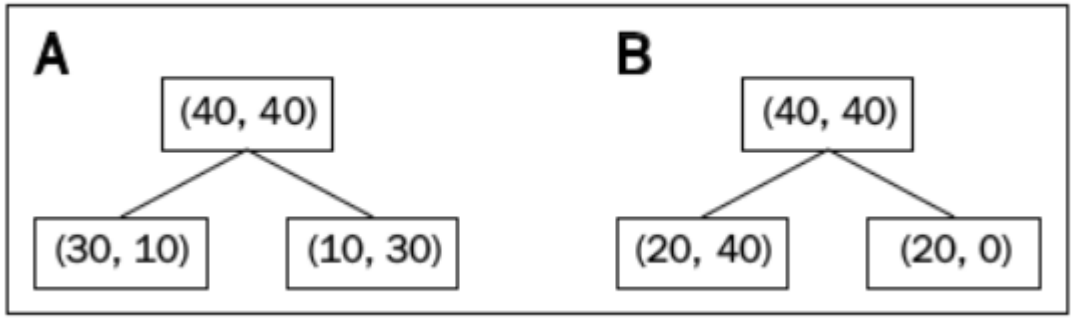
不纯度指数(Impurity Index)

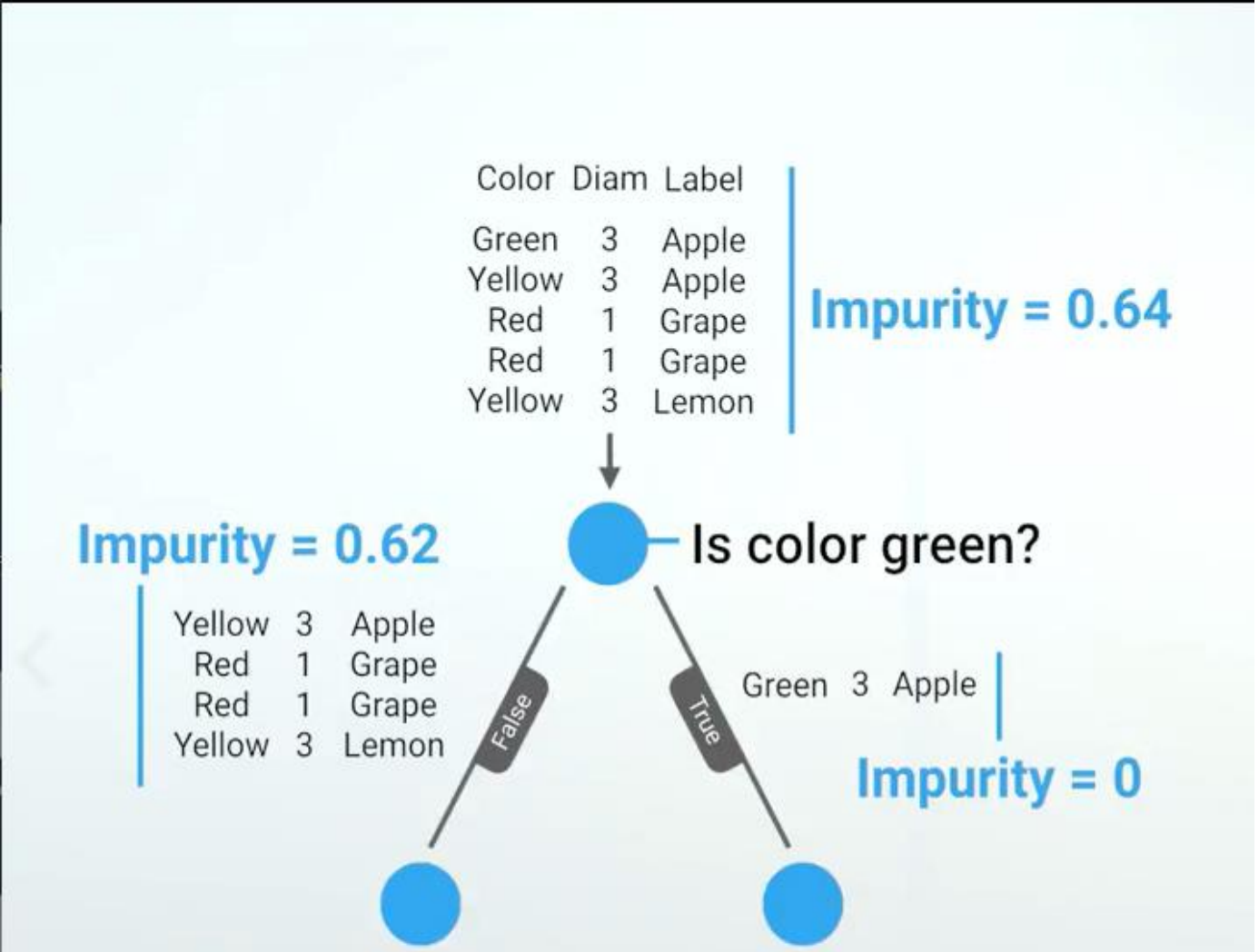
1. Entropy: $\sum_{j=1}^{K} p_j \ln \frac{1}{p_j}$.
2. Misclassification rate: $1 - \max_j p_j$.
3. Gini index: $\sum_{j=1}^{K} p_j(1-p_j) = 1 - \sum_{j=1}^{K} p_j^2$.

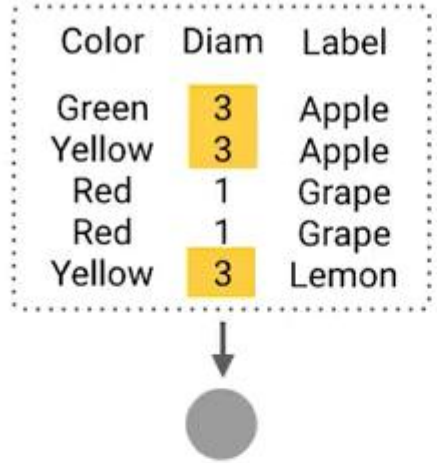


不纯度指数(Impurity Index)

1. Entropy: $\sum_{j=1}^{K} p_j \ln \frac{1}{p_j}$.
2. Misclassification rate: $1-\max_j p_j$.
3. Gini index: $\sum_{j=1}^{K} p_j(1-p_j) = 1-\sum_{j=1}^{K} p_j^2$.



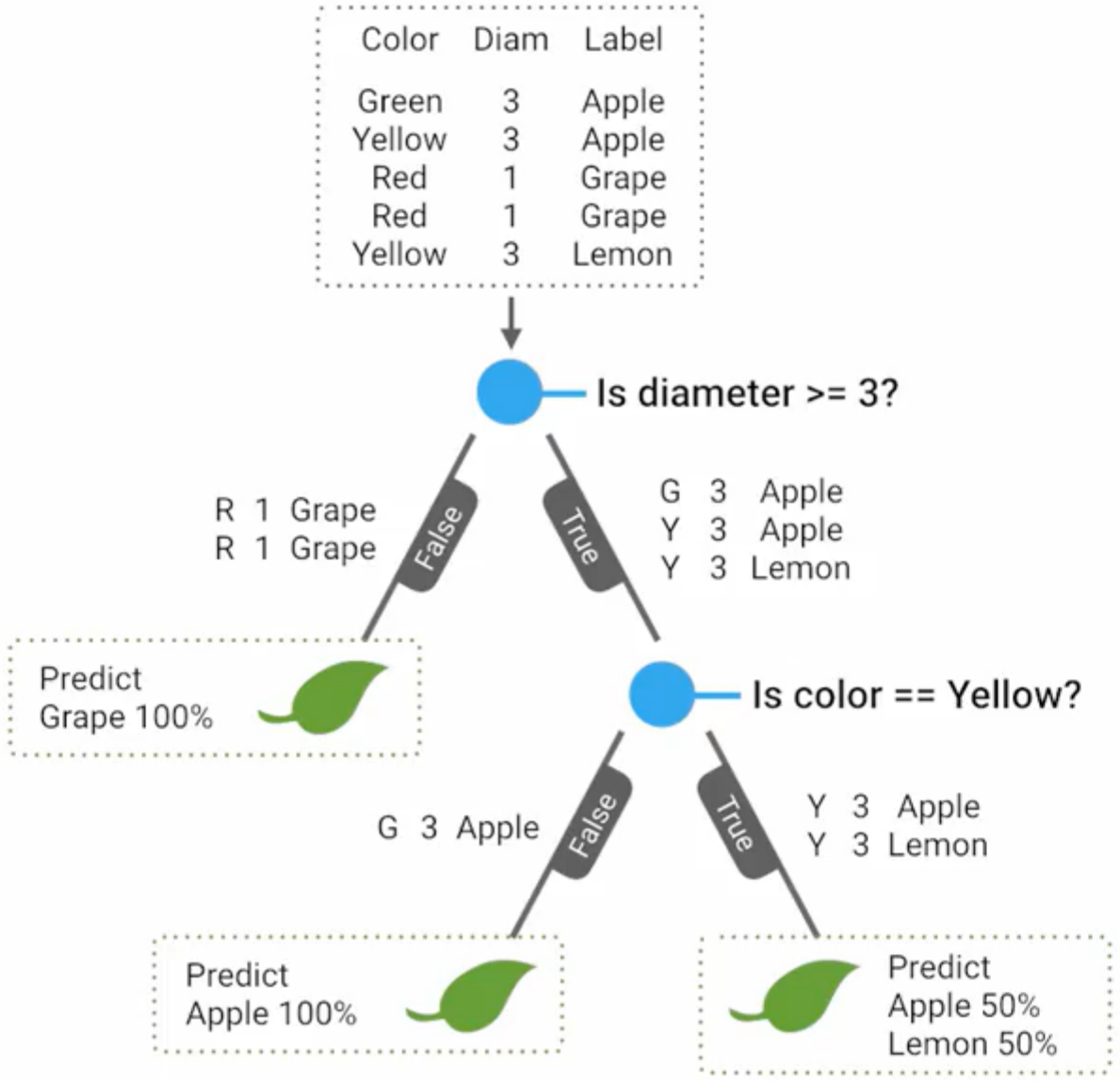




Information Gain

Question	Gain
Color == Green?	0.14
Diameter >= 3?	0.37
Color == Yellow?	0.17
Color == Red?	0.37
Diameter >=1?	0

Here there is a tie. We'll pick the first.


Color Diam Label
Green 3 Apple
Yellow 3 Apple
Red 1 Grape
Red 1 Grape
Yellow 3 Lemon
Is diameter >= 3?
False
True
R 1 Grape
R 1 Grape
G 3 Apple
Y 3 Apple
Y 3 Lemon
Predict
Grape 100%
Is color == Yellow?
False
True
G 3 Apple
Y 3 Apple
Y 3 Lemon
Predict
Apple 100%
Predict
Apple 50%
Lemon 50%

谢谢各位同学！