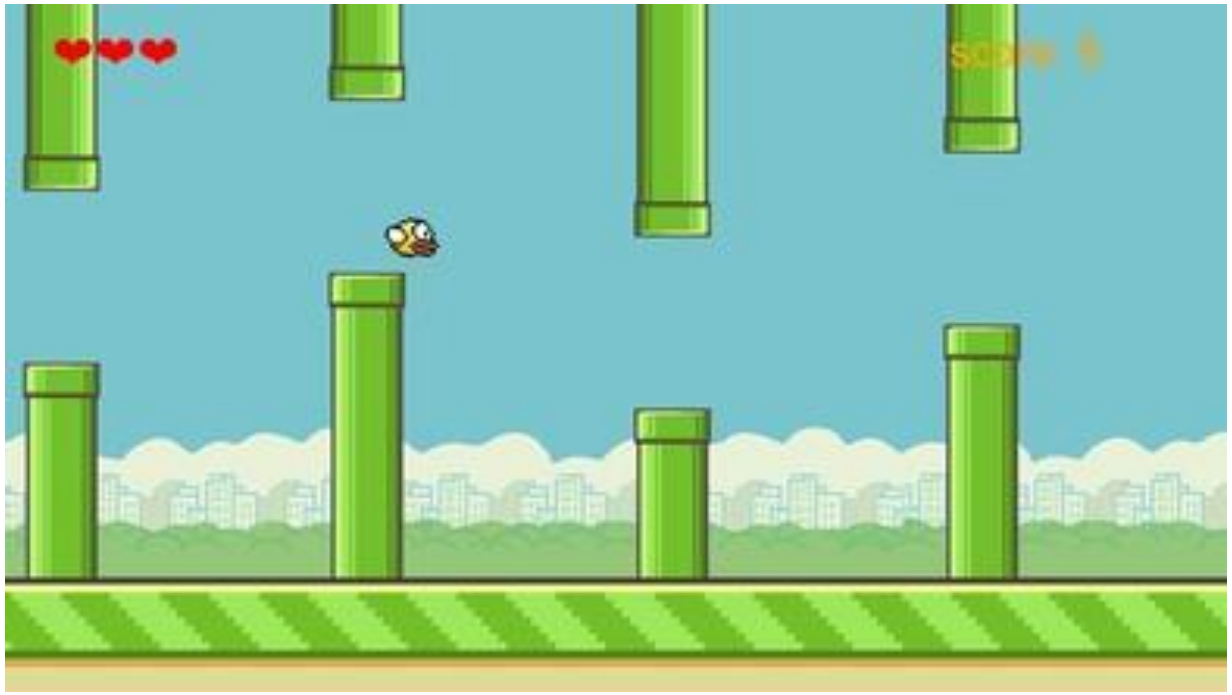
Decision Trees 决策树

CODEGIRLS2016

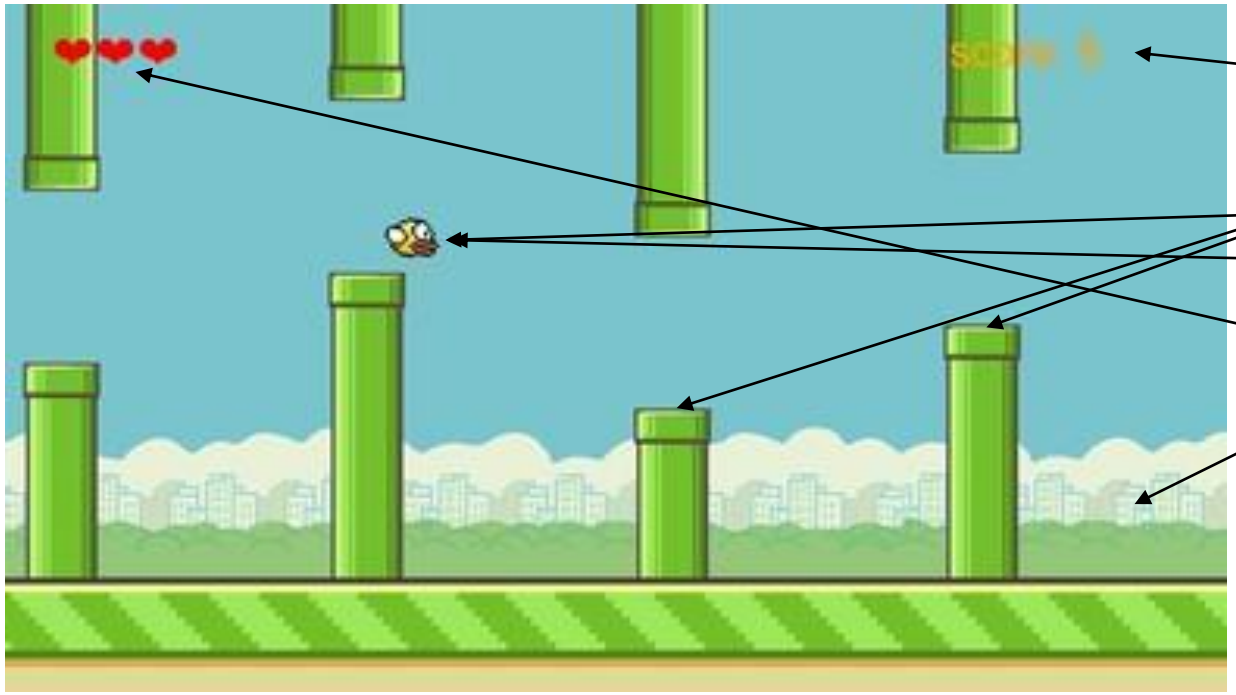
讨论



在游戏中怎样判断要不要做出一个选择？

- Flappy Bird 只需要按一个按钮，这给了我们多大的选择空间？
- 我们做出这个选择的目标是什么？
- 需要哪些信息知道我们要不要让小鸟跳？

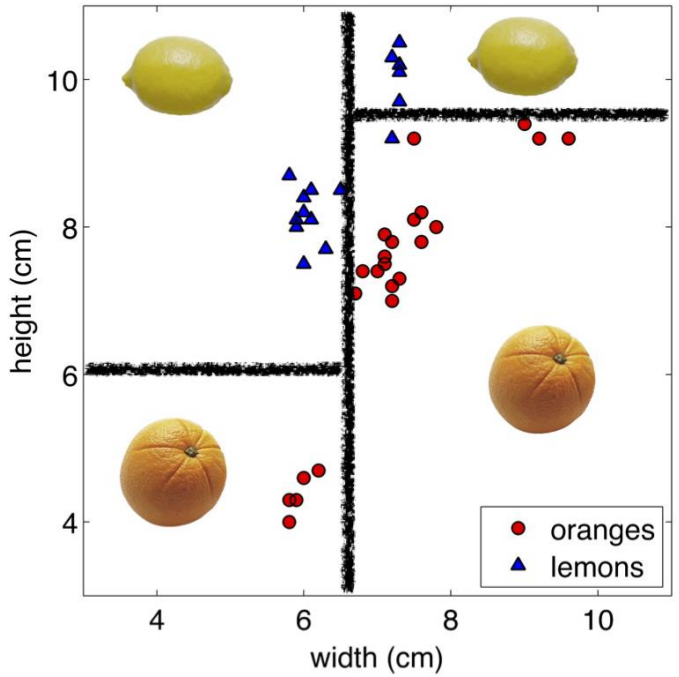
游戏细节



- 分数
- 柱子位置
- 小鸟位置
- 小鸟速度
- 房子
- 生命

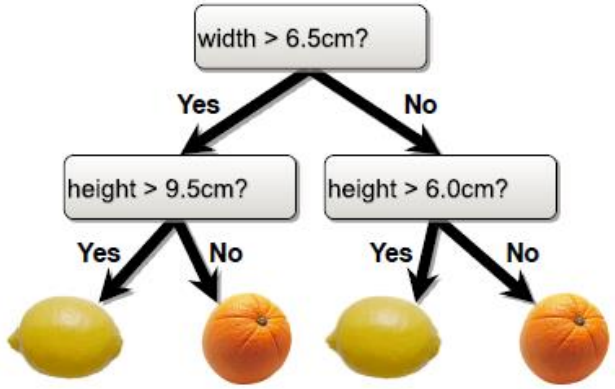
这些信息有哪些是可以帮助我们做出判断的？

分界线



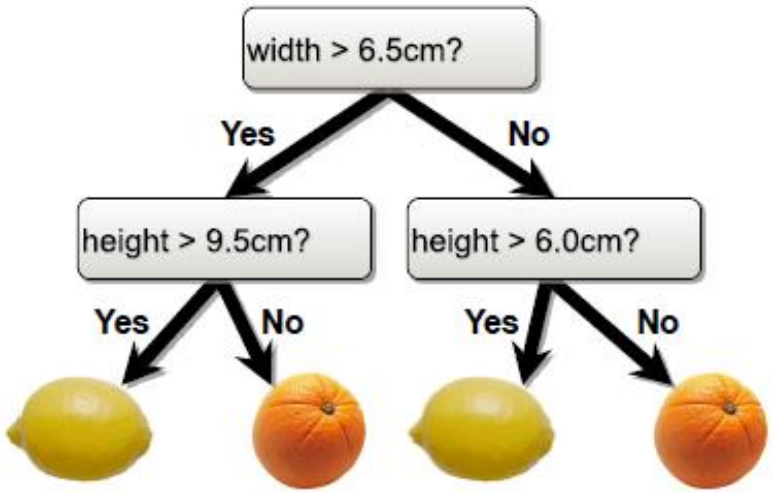
- 复杂的分界线可以划分成几段简单的分界线
- 先切分一个变量（宽度），再切分另一个变量（长度）

树状结构



- 每一次切分都是对数据的一次分割
- 每一个节点都对应一个属性测试
- 最底下的树叶是输出（类别判断）
- 决策树可以表达任何一个二进制的函数

Algorithm 基本算法



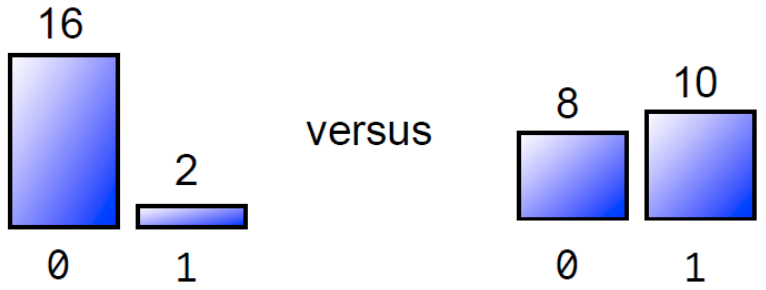
- 1. 在每一层选择一个属性
- 2. 先将上层选择固定
- 3. 通常每一层只限制一个变量/维度
- 4. 在底层产生输出/判断

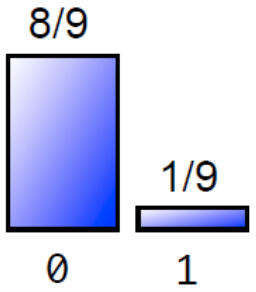
建立决策树

- 香农熵： $\mathrm{H}(\mathrm{x}) = -\sum_{\mathrm{x}\in\mathrm{X}} p(x)\log_2 p(x)$

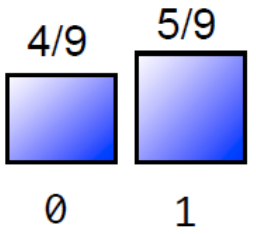
Sequence 1:
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 ... ?

Sequence 2:
0 1 0 1 0 1 1 1 0 1 0 0 1 1 0 1 0 1 ... ?





$$\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx \frac{1}{2}$$



$$\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0.99$$

	Cloudy	Not Cloudy
Raining	24/100	1/100
Not Raining	25/100	50/100

X = Cloud random variable
Y = Rain random variable

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) \\
&= -\frac{24}{100}\log_2\frac{24}{100} - \frac{1}{100}\log_2\frac{1}{100} - \frac{25}{100}\log_2\frac{25}{100} - \frac{50}{100}\log_2\frac{50}{100} \\
&\approx 1.56\text{bits}
\end{aligned}
$$

$$
\begin{aligned}
H(X|Y=y) &= \sum_{x\in X} p(x|y)\log_2 p(x|y) \\
&= -\frac{24}{25}\log_2\frac{24}{25} - \frac{1}{25}\log_2\frac{1}{25} \\
&\approx 0.24\text{bits}
\end{aligned}
$$

Information Gain (IG)

$$
\begin{aligned}
H(X|Y) &= \sum_{y \in Y} p(y) H(X|Y=y) \\
&= \frac{1}{4} H(\text{clouds}|\text{is raining})) + \frac{3}{4} H(\text{clouds}|\text{not raining}) \\
&\approx 0.75 \text{ bits}
\end{aligned}
$$

$$
\begin{aligned}
IG(X|Y) &= H(X) - H(X|Y) \\
&\approx 0.25 \text{ bits}
\end{aligned}
$$

- 当我们知道“是否下雨”（Y）的时候，我们会更加了解天空中有没有云（X）
- 最大的信息增可以让我们选择出最好的随机变量和阈值