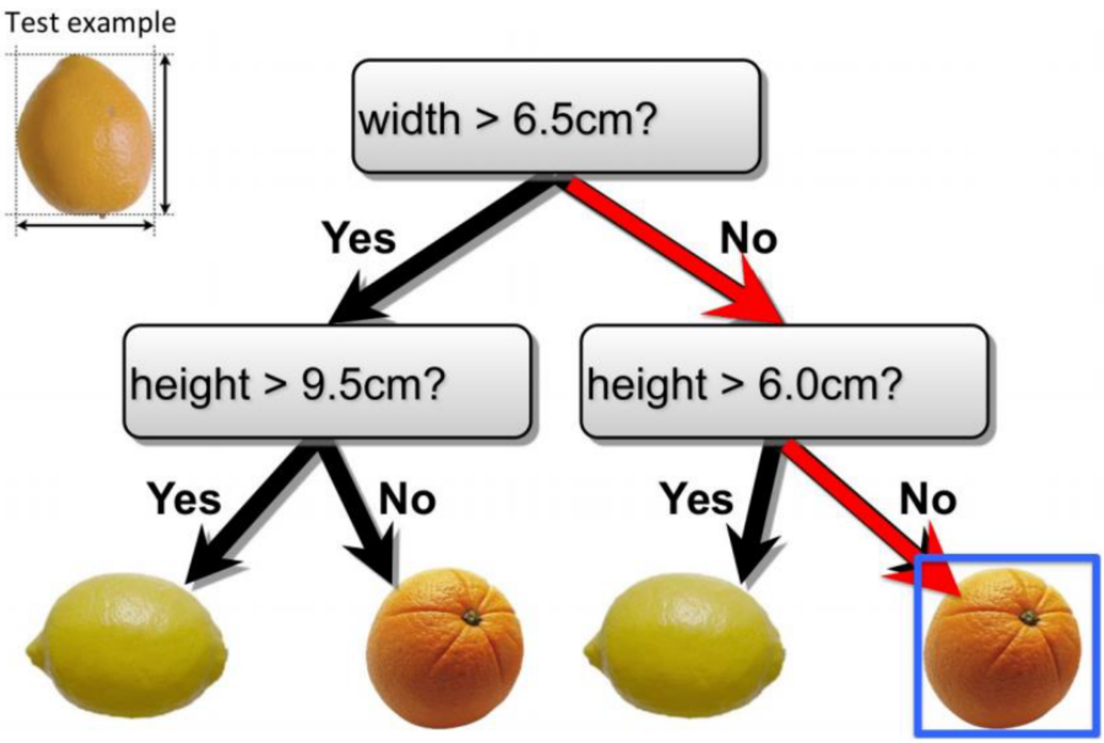
Decision Trees

Decision Trees

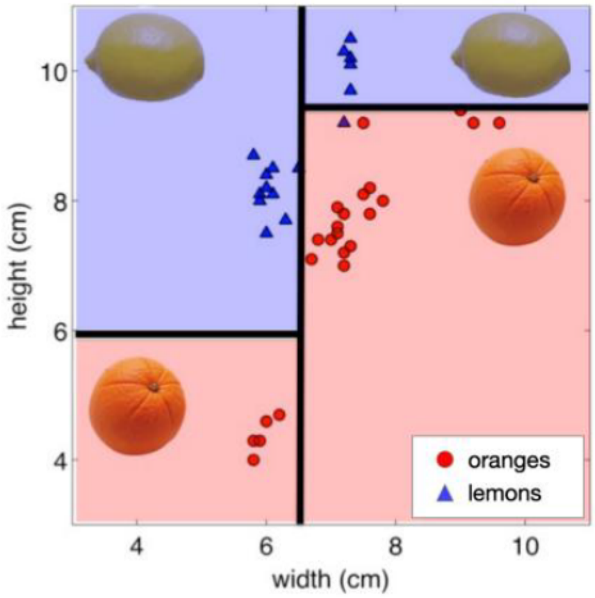
Introduction

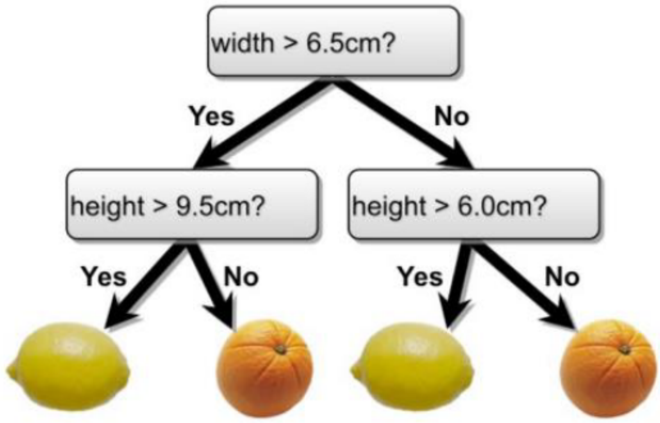
- Make predictions by splitting on features according to a tree structure.



- Internal nodes test a feature
- Branching is determined by the feature value
- Leaf nodes are outputs (predictions)

上面的决策树中，从 root 到 leaf 共有 4 条路径，它们定义了输入空间的 region R_m，通过 split 时的判断，它们将空间划为 4 部分。





决策树分为两种：

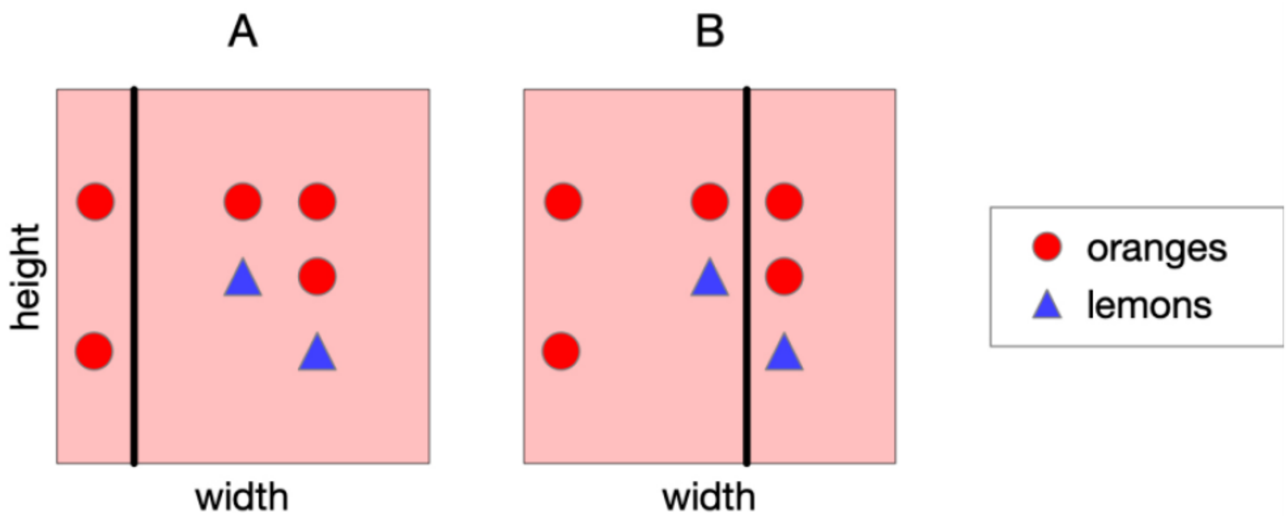
- Classification tree
    - discrete output – leaf 的值是训练集中常见的 y 值。
- Regression tree
    - continuous output – leaf 的值是训练集中 y 的均值（特征空间中某部分的均值）。

Decision Trees

决策树是一个 universal function approximators。

要建立一个决策树，我们可以使用 greedy heuristic：即对于训练集中的 features，我们一个一个试，找到一个能最大程度减少 loss 的用来 split。

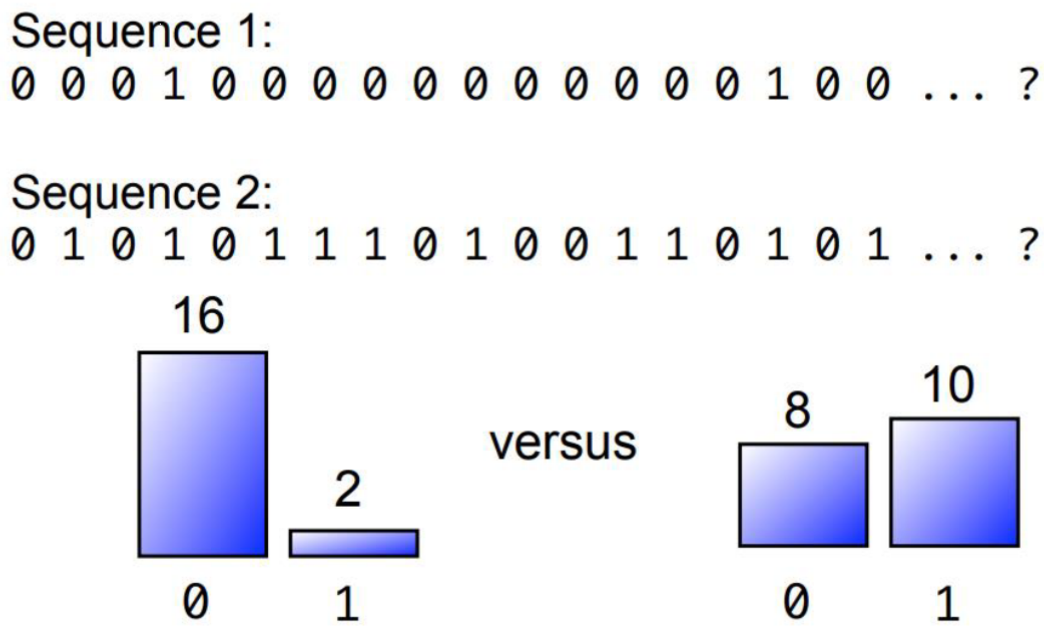
但是现在我们面临一个问题，对于下面这个输入空间，那种划分方式更好？又怎么衡量？



Quantifying Uncertainty

entropy：对于一个离散的随机变量，我们对其可能的 outcomes 的预测有多么没把握 **(uncertainty)**。

举一个硬币的例子：假如我们抛两组硬币，两组使用不同的硬币（硬币不一定公平），得到下面的结果（用 0 和 1 代表正反面）：



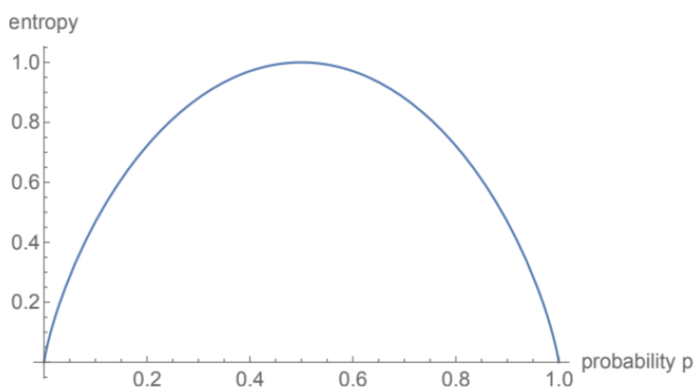
现在我们求一下两个硬币抛出 0 的 entropy：

$$-p\log_2(p) - (1-p)\log_2(1-p)$$

$$-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx \frac{1}{2} \qquad -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0.99$$

注意：**entropy** 越小，我们的把握 **(certainty)** 就越大。

这很容易理解：如果有一个人干了 100 件事，其中 90 件是坏事，我们就可以有把握的说他是坏人；但如果他干了 60 件坏事，40 件好事，那我们就没那么有把握了。



entropy 的单位是 bits。

Information Entropy

More generally, the entropy of a discrete random variable Y is given by

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y)$$

- High Entropy:
  - 变量都有着类似的分布(概率相近)
  - 柱状图呈扁平状(flat),差不多一样高
  - 从中抽取的值具备较低的可预测性(predictable)
- Low Entropy:
  - 变量集中分布在某些区域(概率相差大)
  - 柱状图中柱子间高度差距大
  - 从中抽取的值具备较高的可预测性

Conditional Entropy

接下来看下面这个例子:

- Example: $X = \{\text{Raining, Not raining}\}$, $Y = \{\text{Cloudy, Not cloudy}\}$

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

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

上面我们求出了整个数据的 information entropy(信息熵)。

信息熵只考虑一个随机变量的所有可能取值,即所有可能发生事件所带来的信息量的期望。

而条件熵会考虑两个变量,为 X 给定条件下,Y 的条件概率分布的熵对 X 的数学期望。

Specific Conditional Entropy

- What is the entropy of cloudiness Y, **given that it is raining**?

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

上面是求在下雨的时候，云量的 entropy。其中 $p(y|x)$ 代表在 x 发生的情况下 y 发生的概率。而 $p(y \mid x) = \frac{p(x,y)}{p(x)}$，假如 x 代表下雨，y 代表 cloudy，那么 x 和 y 同时发生的概率 $p(x,y)$ 是 $\frac{24}{100}$，而 x 发生的概率是 $\frac{24}{100} + \frac{1}{100} = \frac{25}{100}$。因此得到上面的结果。

这里把条件 X 固定了，对于所有 X 的条件熵在下面。

Expected conditional entropy

现在，我们可以推出条件熵的通式：

The expected conditional entropy:

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in X} p(x)H(Y|X=x) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x)
\end{aligned}
$$

- What is the entropy of cloudiness, given the knowledge of whether or not it is raining?

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

上面是整个数据的条件熵。其中“是否下雨”是 X，因此有两种情况，即两个 p(x)，分别为 $\frac{24}{100} + \frac{1}{100} = \frac{1}{4}$，和 $\frac{25}{100} + \frac{50}{100} = \frac{3}{4}$。

Properties

- 熵永远非负
- Chain rule: $H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$，即条件熵+当条件的变量的信息熵=整体的信息熵

  If X and Y independent, then X does not affect our uncertainty about Y: $H(Y|X) = H(Y)$
- But knowing Y makes our knowledge of Y certain: $H(Y|Y) = 0$

  By knowing X, we can only decrease uncertainty about Y: $H(Y|X) \leq H(Y)$

Information Gain

熵：表示随机变量的不确定性。

条件熵：在一个条件下，随机变量的不确定性。

信息增益：熵 - 条件熵。表示在一个条件下，信息不确定性减少的程度。

- This is called the information gain $IG(Y|X)$ in Y due to X, or the mutual information of Y and X

$$IG(Y|X) = H(Y) - H(Y|X)$$

- If X is completely uninformative about Y: $IG(Y|X) = 0$
- If X is completely informative about Y: $IG(Y|X) = H(Y)$

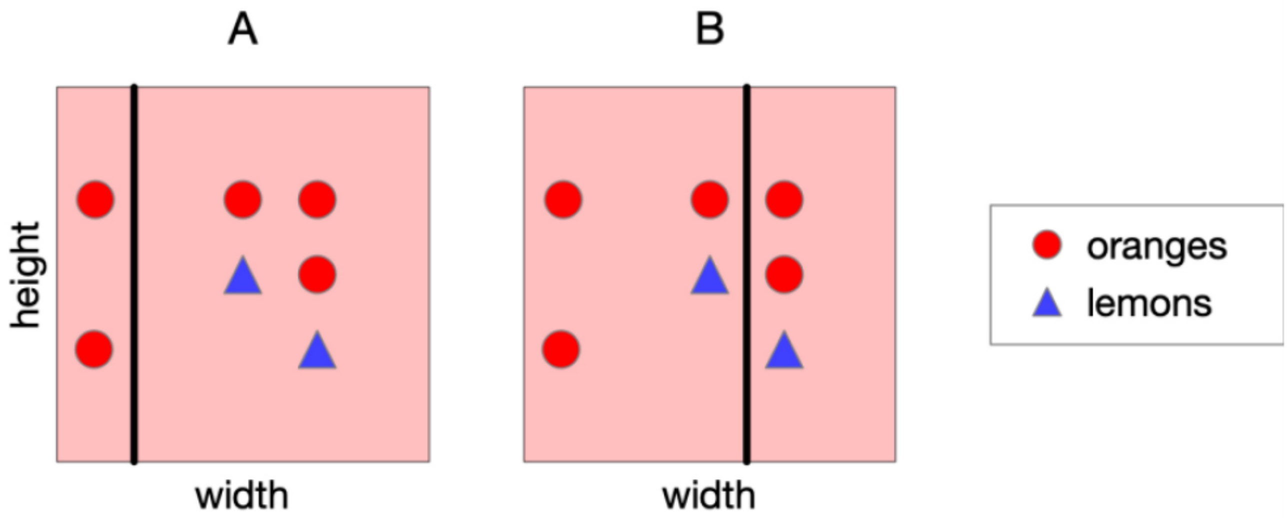
举个例子：Y（明天下雨）是一个随机变量，Y 的信息熵可以算出来， X（明天阴天）也是随机变量，在阴天情况下下雨的条件熵我们也知道。

Y 的熵减去 X 条件下 Y 的熵（条件熵），就是信息增益。具体解释：原本明天下雨（Y）的信息熵是 2，条件熵是 0.01（因为如果知道明天是阴天，那么下雨的概率很大，信息量少），这样相减后信息增益为 1.99。在获得阴天这个信息后，下雨信息不确定性减少了 1.99，不确定减少了很多，所以信息增益大。也就是说，明天阴天这个信息对明天下午这一推断来说非常重要。

所以在特征选择的时候常常用信息增益，如果 IG（信息增益）大的话那么这个特征对于分类来说很关键，决策树就是这样来找特征的。

Decision Trees

现在，我们回到之前的问题，下面两种划分方式哪种好。



我们分别求它们的 IG。

首先，我们要确定 Y 和 X。Y 一般是预测的结果，因此这里的 Y 为 oranges 和 lemons；对应的 X 则是 Y 被分在了左边还是右边。

然后我们找出它们的相关信息，计算 IG。

这里以 B 为例：

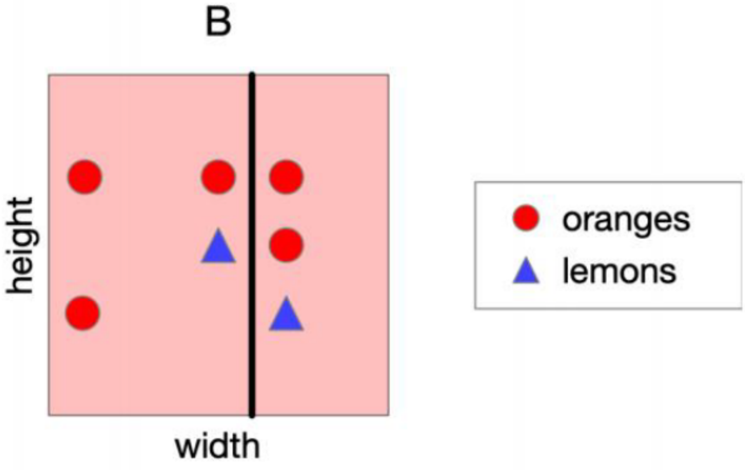
	orange	lemon
left	$\frac{3}{7}$	$\frac{1}{7}$
right	$\frac{2}{7}$	$\frac{1}{7}$

$$IG(Y|X) = H(Y) - H(Y|X)$$

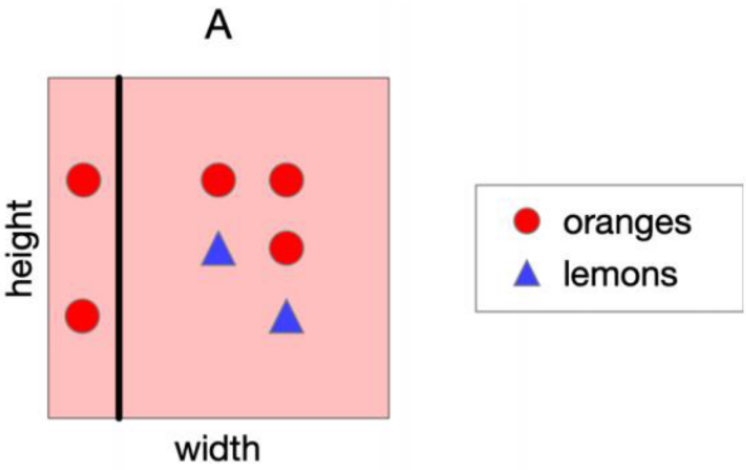
$$H(Y) = -\sum p(y)\log_2 p(y)$$
$$= -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log\frac{2}{7}$$
$$\approx 0.86$$

$$H(Y|X) = \sum p(x) H(Y|X=x)$$
$$= -\sum_{X}\sum_{Y} p(x,y)\log_2 p(y|x)$$
$$= \frac{3}{7}\cdot\log_2\left[\frac{\frac{3}{7}}{\frac{4}{7}}\right] + \cdots + \frac{1}{7}\cdot\log_2\left[\frac{\frac{1}{7}}{\frac{3}{7}}\right]$$
$$\approx 0.857$$

结果：



- Root entropy of class outcome: $H(Y) = -\frac{2}{7}\log_2(\frac{2}{7}) - \frac{5}{7}\log_2(\frac{5}{7}) \approx 0.86$
- Leaf conditional entropy of class outcome: $H(Y|left) \approx 0.81$, $H(Y|right) \approx 0.92$
- $IG(split) \approx 0.86 - (\frac{4}{7}\cdot 0.81 + \frac{3}{7}\cdot 0.92) \approx 0.006$

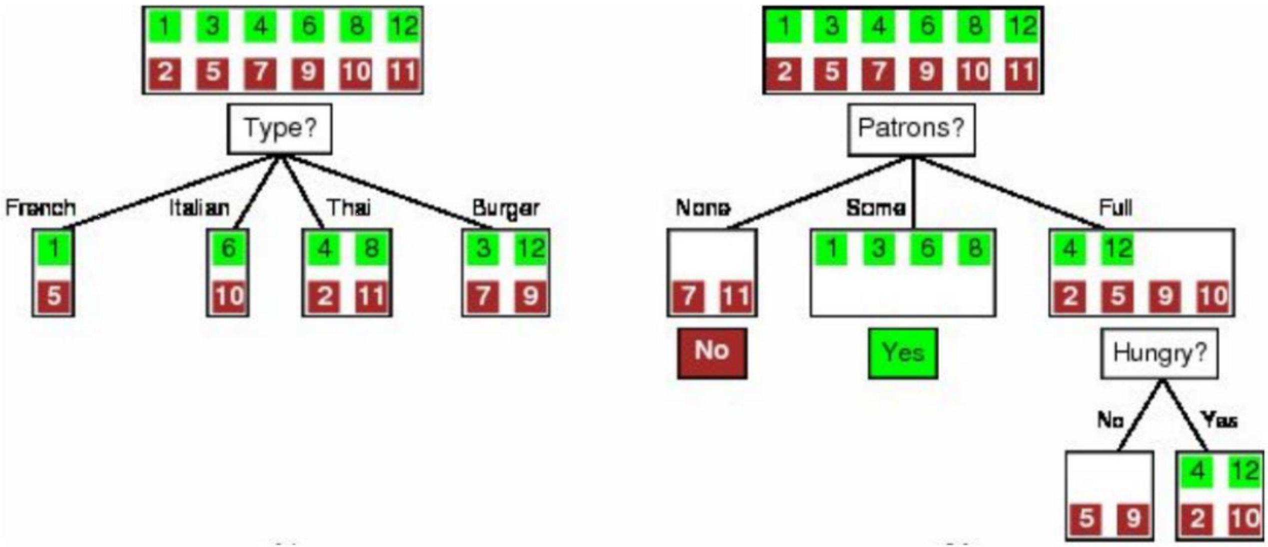


- Root entropy of class outcome: $H(Y) = -\frac{2}{7}\log_2(\frac{2}{7}) - \frac{5}{7}\log_2(\frac{5}{7}) \approx 0.86$
- Leaf conditional entropy of class outcome: $H(Y|left) = 0$, $H(Y|right) \approx 0.97$
- $IG(split) \approx 0.86 - (\frac{2}{7}\cdot 0 + \frac{5}{7}\cdot 0.97) \approx 0.17!!$

Decision Tree Construction Algorithm

构建决策树的算法：

- Simple, greedy, recursive approach, builds up tree node-by-node
  1. pick a feature to split at a non-terminal node
  2. split examples into groups based on feature value
  3. for each group:
     - if no examples – return majority from parent
     - else if all examples in same class – return class
     - else loop to step 1
- Terminates when all leaves contain only examples in the same class or are empty.



$$IG(Y) = H(Y) - H(Y|X)$$

$$IG(type) = 1 - \left[\frac{2}{12}H(Y|\text{Fr.}) + \frac{2}{12}H(Y|\text{It.}) + \frac{4}{12}H(Y|\text{Thai}) + \frac{4}{12}H(Y|\text{Bur.})\right] = 0$$

$$IG(Patrons) = 1 - \left[\frac{2}{12}H(0,1) + \frac{4}{12}H(1,0) + \frac{6}{12}H(\frac{2}{6},\frac{4}{6})\right] \approx 0.541$$

Comparison to some other classifiers

Advantages of decision trees over KNNs and neural nets

- Simple to deal with discrete features, missing values, and poorly scaled data
- Fast at test time
- More interpretable

Advantages of KNNs over decision trees

- Few hyperparameters
- Can incorporate interesting distance measures (e.g. shape contexts)

Advantages of neural nets over decision trees

- Able to handle attributes/features that interact in very complex ways (e.g. pixels)