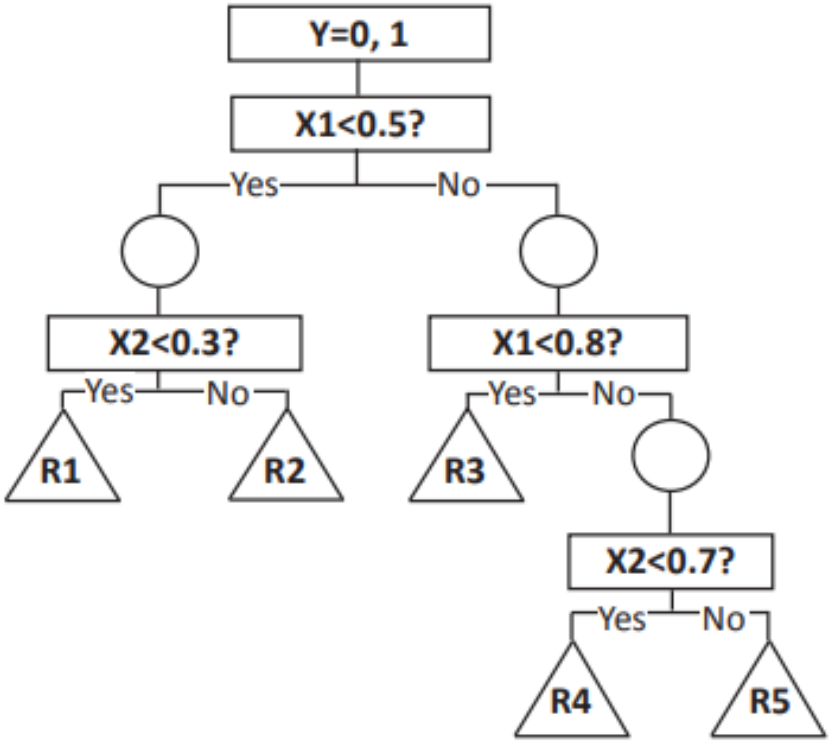
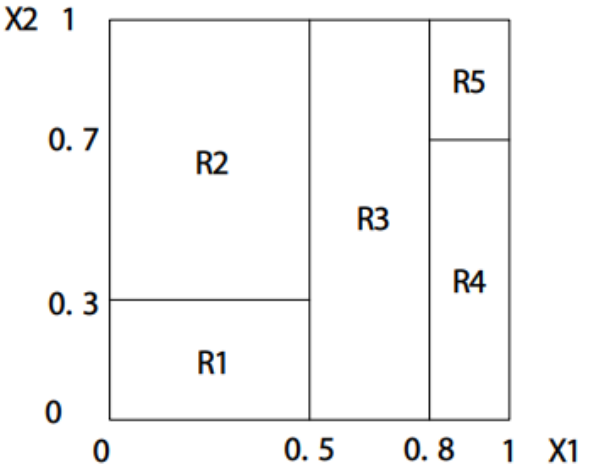
Деревья решений

Что такое дерево решений





- Узлы:
  - Внутренние- правило разделения (ветвления)
  - Листья- конечное значение
- Критерий разделения
- Критерий остановки
- Критерий постобрезки (не обязательно)

Критерий разделения

- В каждом узле m разбиение множества всех объектов, соответствующих данному узлу($\widetilde{X}_M$) на множества $X_{m,l}$ и $X_{m,r}$ выбирается таким образом, чтобы максимизировать функционал $H(\widetilde{X}_M) - \frac{|X_{m,l}|}{|\widetilde{X}_m|} H(X_{m,l}) - \frac{|X_{m,r}|}{|\widetilde{X}_m|} H(X_{m,r})$, где H критерий информативности.
- Для регрессионных задач мы можем использовать сумму квадратов $\frac{1}{|X|} \sum_{y_i \in X} (y_i - \bar{y})^2$, где $\bar{y}$ среднее значение
- Для классификационных задач:
  - Критерий Джини $H(X) = \sum_{k=1}^{K} p_k (1 - p_k)$, где p_k это доля объектов класса k $\frac{|\{y_i : y_i \in X \wedge y_i = k\}|}{|X|}$
  - Энтропия Шеннона $H(X) = -\sum_{k=1}^{K} p * \log_2 p$, где $0 \log_2 0$ принимается равным нулю

Критерии остановки

- Достижение максимальной глубины дерева
- Достижение максимального количества узлов в дереве
- Достижении приростом функционала качества минимального значения
- Прекращение ветвления узлов, когда все объекты в узле принадлежат одному классу
- Прочее

Алгоритмы построения

- Взято из https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

Methods	CART	C4.5	CHAID	QUEST
Measure used to select input variable	Gini index; Twoing criteria	Entropy info-gain	Chi-square	Chi-square for categorical variables; J-way ANOVA for continuous/ordinal variables
Pruning	Pre-pruning using a single-pass algorithm	Pre-pruning using a single-pass algorithm	Pre-pruning using Chi-square test for independence	Post-pruning
Dependent variable	Categorical/ Continuous	Categorical/ Continuous	Categorical	Categorical
Input variables	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous	Categorical/ Continuous
Split at each node	Binary; Split on linear combinations	Multiple	Multiple	Binary; Split on linear combinations