Rozhodovací stromy

Rozhodovací strom je binární zakořenený strom s rozhodovacím kriteriem c_t v každém vnitřním vrcholu.

Nejdříve začínáme se všemi daty v počátečním vrcholu a ten štěpíme. Chceme vždycky štěpit ten vrchol, jehož štěpením si nejvíce pomůžeme. Tzn. hledáme takový vrchol, pro který je $c_{t_l} + c_{t_r} - c_t$ největší

Regrese

Štěpící kriterum je

$$c_t = \sum_{i \in I_t} (t_i - \hat{t_t})^2, \text{ kde } \hat{t_t} = \frac{1}{|I_t|} \sum_i t_i$$

Klasifikace

Pro každý list vytvoříme distibuci $p_t(k)$ pravděopdobnost, že v listu t dostanu třídu k. To je přesně distribuce tříd všech dat, které spadli do listu.

Používáme dvě různá štěpící kritéria

$$c_{gini}(t) = |I_t| \sum_k p_t(k)(1 - p_t(k1))$$

nebo

$$c_{entropy}(t) = |I_t| \cdot H(p_t)$$