0 Rozhodovací stromy

Na vstup předpokládáme tabulku s N záznamy a p příznaky $X_0, X_1, \ldots, X_{p-1}$. Cílem je vytvořit rozhodovací strom který přiřadí co nejvíce vstupům co nejpřesnější hodnoty Y. Exhaustive search takového optimálního stromu je NP-úplný problém, existují však algoritmy nabízející kompromis. Pro klasifikaci ID3 a pro regresi CART nabízejí suboptimální ale dosažitelné řešení.

0.1 Algoritmus ID3

Algoritmus ID3 je hladový algoritmus pro konstrukci rozhodovacího stromu. V každém vrcholu hledá podle zadaného kritéria nejlepší test (rozhodovací pravidlo) nepoužitých příznaků, které nejlépe rozdělí data na dvě podmnožiny, které maximalizují vybrané kritérium.

Algoritmus začíná s celým datasetem a rekurzivně se zanořuje do synů, dokud nenastane zastavovací kritérium (typicky max hloubka, malý počet záznamů v listech, ...).

0.2 Kritéria pro větvení

0.2.1 Míra neuspořádanosti

Na binární množině kde p_0 a p_1 označují poměry 0 a 1, definujeme míru neuspořádanosti jako funkci p_0 splňující:

1. nezápornost na $[0, 1]$,
2. nulovost pro $p_0 = 1$ nebo $p_0 = 0$,
3. maximum nabývá v $p_0 = \frac{1}{2}$,
4. je rostoucí na $\left[0, \frac{1}{2}\right]$ a klesající na $\left[\frac{1}{2}, 1\right]$

0.2.2 Entropie

Definici míry neuspořádanosti například splňuje entropie.

$$H(\mathcal{D}) = -p_0 \log p_0 - (1 - p_0) \log(1 - p_0)$$

Entropii lze definovat i pro nebinární hodnoty.

$$H(\mathcal{D}) = -\sum_{i=0}^{k-1} p_i \log p_i$$

Entropie (pojem z teorie informace) používá dvojkový logaritmus a pracuje s jednotkou bit (Claude Shannon).

0.2.3 Gini index

Místo entropie lze použít gini index (gini impurity), které má podobné vlastnosti.

$$GI(\mathcal{D}) = \sum_{i=0}^{k-1} p_i(1 - p_i)$$

0.2.4 Informační zisk

Příznak pro rozdělení se volí podle informačního zisku

$$IG(\mathcal{D}) = H(\mathcal{D}) - t_0 H(\mathcal{D}_0) - t_1 H(\mathcal{D}_1)$$

kde $\mathcal{D}_0, \mathcal{D}_1$ jsou příslušné podmnožiny $\mathcal{D}$ a t_0, t_1 poměry počtu 0 a 1 v Y.

0.3 Použití pro klasifikaci a regresi

0.3.1 Klasifikace

V případě klasifikace strom rozhoduje většinovým hlasováním v příslušném listu, do kterého záznam přísluší. Poměr výsledného prvku v listu určuje jistotu modelu při takové volbě.

0.3.2 Regrese

U regrese se rozhoduje podle průměru v příslušném listu.

Konstrukce stromu v regresní úloze probíhá podobně jako v klasifikační. Místo minimalizace míry neuspořádanosti, algoritmus dělí data tak, aby byly hodnoty v listech co nejblíže ke střední hodnotě.

Pro odhad odchylky od střední hodnoty se používá MSE (mean squared error), případně MAE (mean absolute error).

$$\mathrm{MSE}(\mathbf{Y}) = \frac{1}{N} \sum_{i=0}^{N-1} (Y_i - \bar{Y})^2 \qquad \mathrm{MAE}(\mathbf{Y}) = \frac{1}{N} \sum_{i=0}^{N-1} |Y_i - \bar{Y}|$$

Hladovému algoritmu konstrukce stromu, ve kterém se minimalizuje MSE, se říká CART (classification and regression trees).

Jako alternativu informačního zisku (ID3), používá CART

$$\mathrm{MSE}(\mathcal{D}) - t_0 \,\mathrm{MSE}(\mathcal{D}_0) - t_1 \,\mathrm{MSE}(\mathcal{D}_1)$$

0.4 Hyperparametry

Rozhodovací stromy mají často nízký bias a vysokou varianci. Jako hyperparametry rozhodovacího stromu můžeme volit různá ukončovací pravidla, kterými zamezit přeučení.

1. dělící kritérium
   (a) gini, entropy (klasifikace)
   (b) squared error, absolute error (u regrese)
2. max hloubka, od které algoritmus už dále nedělí
3. minimální počet dat v množině před, resp. po, dělení
4. minimální nutná hodnota informačního zisku

0.5 Shrnutí

Rozhodovací stromy jsou nenáročné na přípravu, jsou dobře interpretovatelné, jednoduché a rychlé. Jsou však nerobustní a je snadné je přeučit.