BI-VZD + BI-ZNS - Rozhodovací stromy: algoritmus konstrukce stromů, hyperparametry. Náhodné lesy.

Vastl Martin

May 27, 2019

Rozhodovací strom je druh stromu, který obsahuje ve svých vrcholech rozhodovací pravidla a v listech předpovězenou hodnotu. Může se jednat jak o klasifikační stromy nebo o regresní stromy.

Cílem je vytvořit strom zadané hloubky k, který správně přiřadí hodnotu Y co nejvíce řádkům tabulky na základě p příznaků $X_0, X_1, \ldots, X_{p-1}$. Konstrukce ideálního stromu je NP-úplný problém, a proto je nutné postupovat jiným způsobem. Pro konstrukci se proto využívá hladového algoritmu ID3. Tento algoritmus vybírá jeden ze zatím nepoužitých příznaků, který rozdělí data na 2 části tak, že vzniklé rozdělení maximalizuje vybrané kritérium.

1 Algoritmus konstrukce stromu

K tomu, abychom byli schopni konstruovat strom, potřebujeme nějak kvantifikovat míru neuspořádanosti stromu. K tomu použijeme entropii

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

Při konstrukci chceme vybrat příznak, který rozdělením dat nejvíce sníží neuspořádanost. Toto snížení určuje informační zisk.

$$\mathrm{IG}(D, X_i) = H(D) - t_0 H(D_0) - t_1 H(D_1), \tag{2}$$

kde D_0 a D_1 jsou podmnožiny dat D a t_i je podíl počtu prvků v D_i a D, neboli $t_i = \frac{\#D_i}{\#D}$. Takto vybereme příznak, kterým získáme největší IG. Tento postup opakujeme rekurzivně na rozdělené množiny a zastavíme se v moment, kdy dosáhneme nějakého kritéria např. hloubky stromu, minimální počet dat v množině nebo minimální nutná hodnota informačního zisku, které jsou také hyperparametrem modelu.

Namísto entropie lze využít i Gini index, jehož výhodou je menší výpočetní náročnost.

$$\mathrm{GI}(D) = 1 - \sum_{i=0}^{k-1} p_i^2 \tag{3}$$

V případě, kdy již máme strom sestavený, rozhodování probíhá v případě klasifikačních stromů tak, že se postupuje z kořene do listu dle podmínek ve vrcholech. V moment, kdy se dostaneme do vrcholu, je výsledná kategorie ta kategorie která v tomto vrcholu převládá. Pokud se jedná o regresní strom, pak je výsledek průměr ze všech datových bodů, které do tohoto listu spadají.

2 Náhodné lesy

Základní myšlenka spočívá v tom, že namísto jednoho modelu (např. rozhodovacího stromu) použijeme více modelů a jejich predikce nějakým způsobem zkombinujeme do finálního rozhodnutí.

Pro jednoduchost předpokládejme, že máme binární klasifikační problém, tj. rozhodujeme jestli $Y = 0$ nebo $Y = 1$.

1. Ze vstupního trénovacího datasetu D vytvoříme n datasetů $D_1, \ldots, D_n$ stejně velkých jako D pomocí metody bootstrap, neboli (středoškolsky) pomocí výběru s opakováním.

2. Na každém datasetu D_i naučíme rozhodovací strom tak, jak jsme si předvedli u rozhodovacích stromů. Může být málo hluboký, klidně hloubky (parametr max_depth) dva nebo tři (používá se i hloubka jedna). Označme tyto stromy $T_1, \ldots, T_n$.

3. Každý datový bod (tj. řádek z tabulky s daty D) proženeme všemi stromy $T_1, \ldots, T_n$ a od každého z nich si uložíme rozhodnutí $Y_1, \ldots, Y_n$.

4. Všechny tyto stromy $T_1, \ldots, T_n$ tvoří náhodný les a jeho finální rozhodnutí o hodnotě Y je dané většinovým rozhodnutím stromů, je-li např. v množině $\{Y_1, \ldots, Y_n\}$ více jedniček než nul, je predikce náhodného lesa $Y = 1$.