Wskazówki do ćwiczenia 4.

WSI - ćwiczenia 2022Z

Jakub Łyskawa, Stanisław Pawlak

November 24, 2022

Algorytm ID3

Indukcja drzew decyzyjnych — ID3

Algorithm 1: ID3

Input: Y: zbiór klas, D: zbiór atrybutów wejściowych, $U \neq \emptyset$: zbiór par uczących

1 **if** $\forall_{\{x_i,y_i\}\in U}\ y_i == y$ **then**
2 **return** *Liść zawierający klasę* y
3 **if** $|D| == 0$ **then**
4 **return** *Liść zawierający najczęstszą klasę w* U
5 $d = \arg\max_{d\in D} InfGain(d, U)$
6 $U_j = \{x_i, y_i\} \in U \ : \ x_i[d] = d_j,$ gdzie d_j - j-ta wartość atrybutu d
7 **return** *Drzewo z korzeniem* d *oraz krawędziami* $d_j, j = 1, 2, \ldots$ *prowadzącymi do drzew: ID3(*Y*,* $D - \{d\}$*,* U_1*), ID3(*Y*,* $D - \{d\}$*,* U_2*),...*

Ograniczanie głębokości drzewa

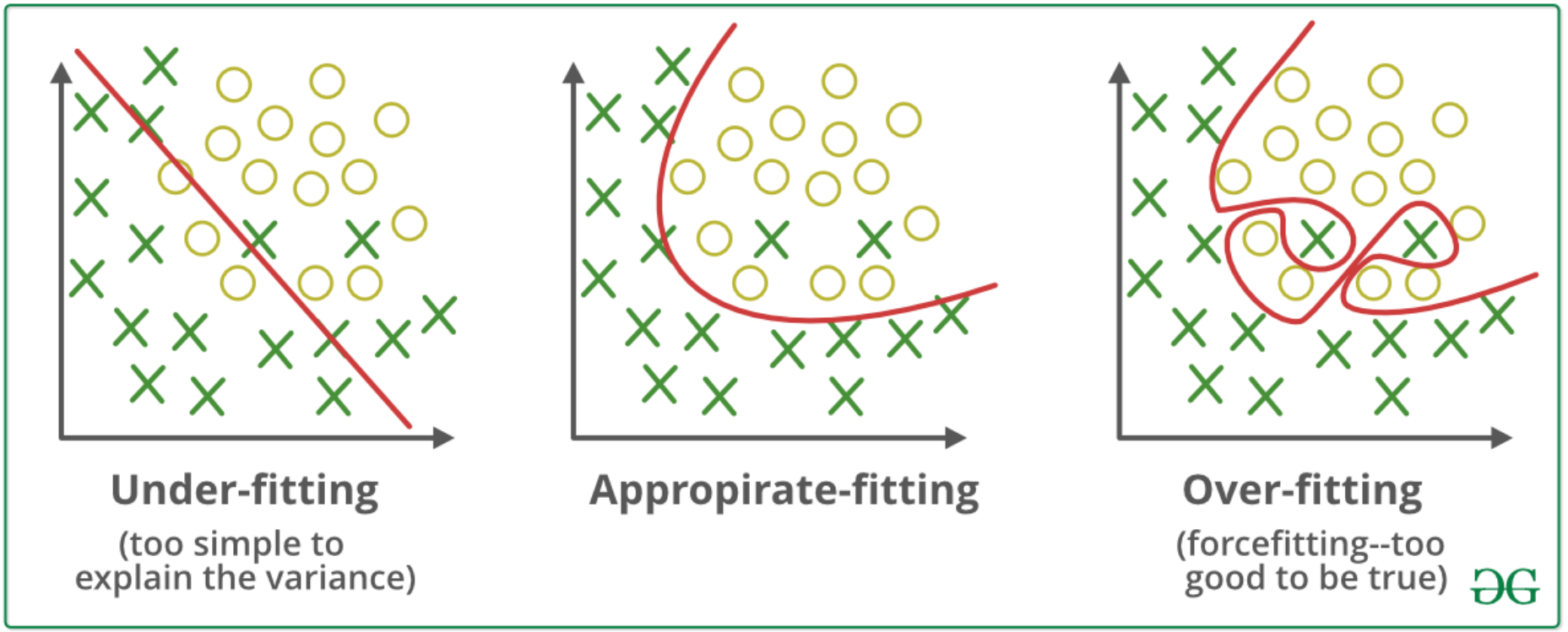
Analogicznie jak przy braku atrybutów do podziału, tworzony jest liść zawierający najczęstszą klasę w pozostałym zbiorze.

Ograniczanie głębokości drzewa

Analogicznie jak przy braku atrybutów do podziału, tworzony jest liść zawierający najczęstszą klasę w pozostałym zbiorze.

Ale dlaczego?

Przeuczanie



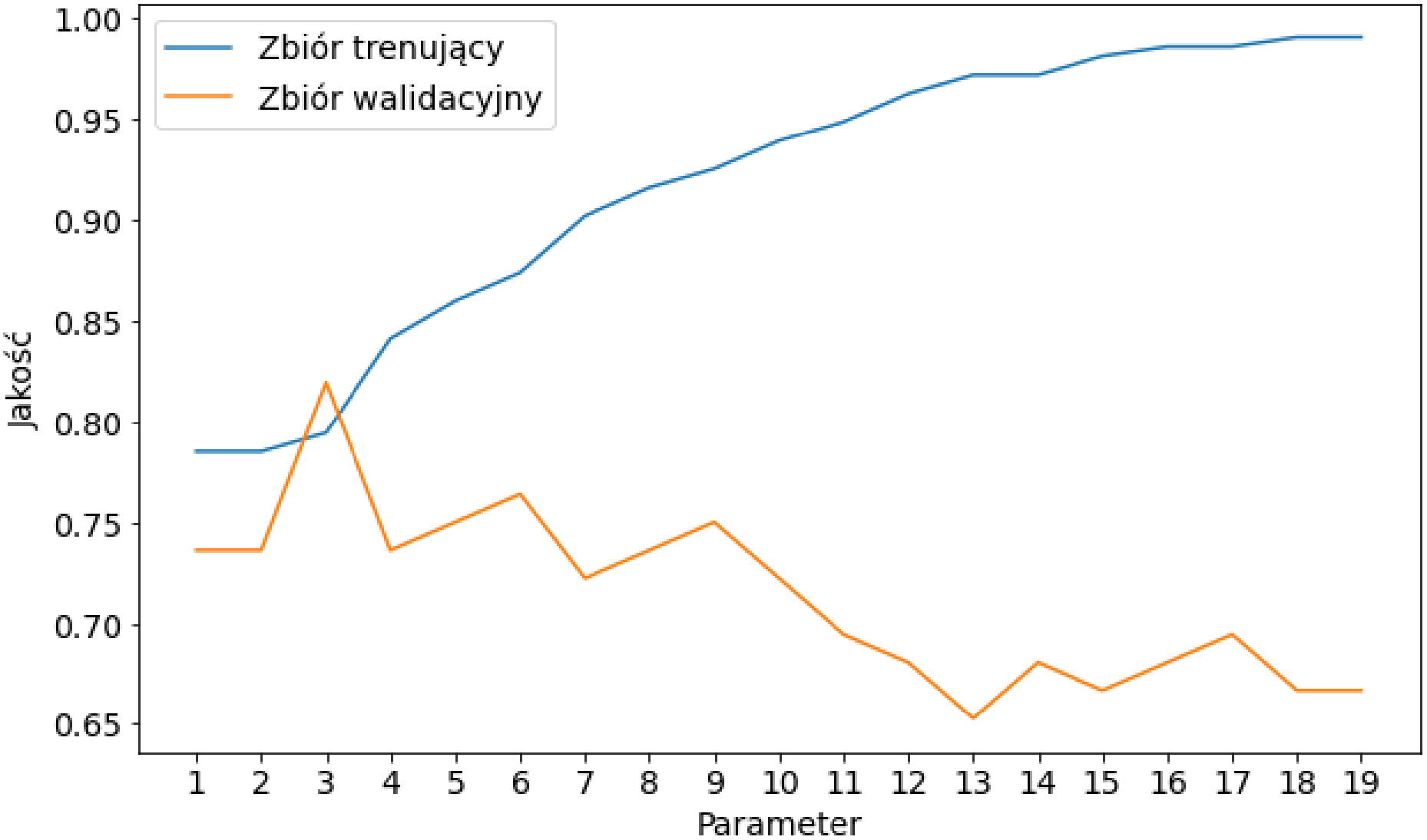
Źródło: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Co można zrobić?

Ograniczyć możliwości dopasowania

Co można zrobić?

Ograniczyć możliwości dopasowania



Ale...

Używając zbioru walidacyjnego do wyboru najlepszej wartości hiperparametru, dokonuje się dopasowania do zbioru walidacyjnego

Ale...

Używając zbioru walidacyjnego do wyboru najlepszej wartości hiperparametru, dokonuje się dopasowania do zbioru walidacyjnego

Potrzebny jest trzeci zbiór do oceny jakości wybranego modelu

Na co jeszcze zwracać uwagę?

- Brakujące wartości?

Na co jeszcze zwracać uwagę?

- Brakujące wartości?
- Co zrobić, jeżeli w danych pojawi się wartość, której nie było w zbiorze trenującym?

Dodatkowe źródła

- Visual intro to ML:
  `http://www.r2d3.us/visual-intro-to-machine-learning-part-1/`
- scikit-learn train-test split function
  (`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`)