Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Anna Beer, Florian Richter

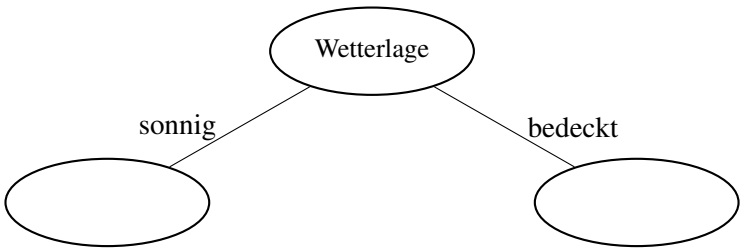
Knowledge Discovery in Databases
WS 2017/18

Übungsblatt XX: Zusatz zu Gini-Index

Folgendes Datenset könnte gegeben sein:

Tag	Wetterlage	Temperatur	Wind	Tennis?
1	sonnig	heiß	schwach	nein
2	sonnig	heiß	stark	nein
3	bedeckt	heiß	schwach	ja
4	bedeckt	kalt	stark	ja
5	bedeckt	kalt	schwach	ja
6	sonnig	kalt	schwach	ja
7	sonnig	heiß	schwach	nein
8	bedeckt	kalt	schwach	ja
9	bedeckt	kalt	stark	nein
10	sonnig	kalt	stark	ja

Angenommen, es wurde schon nach Wetterlage gesplittet. Dann haben wir



Aufgabe XX-1 Gini-Index

Für den linken Zweig benötigen wir nur die „sonnigen“ Tage: 1, 2, 6, 7, 10.

Angenommen, wir nutzen Temperatur als trennendes Attribut. Dann gibt es Teilbäume heiß und kalt. Für beide rechnen wir den Gini-Index aus:

$$gini(heiß|Wetterlage = sonnig) = 1 - p(nein)^2 - p(ja)^2 = 1 - 1^2 - 0^2 = 0$$

Denn bei sonnig heißem Wetter ist die Frage nach Tennis stets nein (Wahrscheinlichkeit 1). Analog:

$$gini(kalt|Wetterlage = sonnig) = 1 - 0^2 - 1^2 = 0$$

Es gibt 5 Tage, die sonnig sind. Davon sind 3 heiß und 2 kalt. Der Gini-Index für Temperatur ist damit:

$$gini(Temperatur|Wetterlage = sonnig) = 3/5 * 0 + 2/5 * 0$$

Da das Attribut mit dem kleinsten Gini-Index am besten trennt, wäre dies schon ein unschlagbarer Kandidat. Wir rechnen das andere Attribut trotzdem aus:

$$gini(schwacherWind|Wetterlage = sonnig) = 1 - p(nein)^2 - p(ja)^2 = 1 - (2/3)^2 - (1/3)^2 = 4/9$$

$$gini(starkerWind|Wetterlage = sonnig) = 1 - p(nein)^2 - p(ja)^2 = 1 - (1/2)^2 - (1/2)^2 = 1/2$$

$$gini(Wind|Wetterlage = sonnig) = 3/5 * 4/9 + 2/5 * 1/2 = 7/15$$

Wir wählen also wie zu erwarten das Attribut Temperatur zum splitten des linken Teilbaums. Bleibt noch der rechte Teilbaum:

$$gini(heiß|Wetterlage = bedeckt) = 1 - p(nein)^2 - p(ja)^2 = 1 - 0^2 - 1^2 = 0$$

$$gini(kalt|Wetterlage = bedeckt) = 1 - p(nein)^2 - p(ja)^2 = 1 - (1/4)^2 - (3/4)^2 = 6/16$$

$$gini(schwach|Wetterlage = bedeckt) = 1 - p(nein)^2 - p(ja)^2 = 1 - 0^2 - 1^2 = 0$$

$$gini(stark|Wetterlage = bedeckt) = 1 - p(nein)^2 - p(ja)^2 = 1 - (1/2)^2 - (1/2)^2 = 1/2$$

$$gini(Temperatur|Wetterlage = bedeckt) = 1/5 * 0 + 4/5 * 6/16 = 3/10$$

$$gini(Wind|Wetterlage = bedeckt) = 3/5 * 0 + 2/5 * 1/2 = 1/5 = 2/10$$

Wir wählen also Wind als rechtes Splitattribut aus.

Falls wir nun noch Entscheidungslabels als Blätter formulieren müssten (von links nach rechts):

- „sonnig, heiß“: Nein (100%)
- „sonnig, kalt“: Ja (100%)
- „bedeckt, schwach windig“: Ja (100%)
- „bedeckt, stark windig“: Ja (50%)