Paradigmen moderner Softwareentwicklung
A7: Decision Trees

Hendrik Thorun
MIM, FH Lübeck

Hamburg, 2. Juli 2016

Aufgabe

Sie finden als nächsten Eintrag in Moodle eine Datei Data.gif mit Trainingsdaten und Testdaten!

Read and understand the new link in loop about decision trees with the play golf example.[http://www.saedsayad.com/decision_tree.htm Golf Beispiel]

Try to play with the dataset and try to create more then one decision tree!

Remark: I do not expect to draw and calculate everything. This would be more then a dozen pages (can be given to you at the end of the course!

Anmerkung: Bei der Bearbeitung der Aufgabe habe ich mich stak an den Ausführungen des „Play Golf"-Beispiels[1] orientiert (auch was die Farbgebung in meinen Tabellen angeht).

[1] http://www.saedsayad.com/decision_tree.htm

Trainingsdaten

Für die Initialisierung des Decision Trees wurden folgende Trainingsdaten verwendet (Abbildung 1)

Quelle: http://moodle.oncampus.de/pluginfile.php/625704/mod_resource/content/1/DATA.gif

ID	Age	Income	Student	Credit_rating	Buys_computer
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31...40	High	No	Fair	Yes
4	>40	Low	Yes	Excellent	No
5	>40	Medium	No	Excellent	No
6	>40	Medium	No	Fair	Yes
7	31...40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Low	Yes	Fair	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	31...40	Medium	No	Excellent	Yes
13	31...40	High	Yes	Fair	Yes
14	>40	Medium	Yes	Fair	Yes

Abbildung 1: Die verwendeten Trainingsdaten

Berechnung der Entropien

Zuerst wird die Entropie des einzelnen `buys_computer`-Attributs berechnet (Abbildung 2):

Buys_computer	
Yes	No
9	5

Abbildung 2: Entropie von Buys_computer.

$$Entropy(Buys\_computer) = Entropy(5, 9) = Entropy(0.36, 0.64)$$
$$= -(0.36 * log2(0.36)) - (0.64 * log2(0.64))$$
$$= 0.94$$

Anschließend folgt die Berechnung der Entropien über jeweils zwei Attribute (Abbildungen 3 bis 6):

		Buys_computer		
		Yes	No	
Age	<=30	2	3	5
	31...40	4	0	4
	>40	3	2	5
				14

Abbildung 3: Entropie von Buys_computer und Age.

$$E(Buys\_computer, Age) = P(<= 30)*E(2,3)+P(31\ldots40)*E(4,0)+P(>40)*E(3,2))$$
$$= (\tfrac{5}{14}) * 0.97 + (\tfrac{4}{14}) * 0.0 + (\tfrac{5}{14}) * 0.97$$
$$= 0.693$$

		Buys_computer		
		Yes	No	
Income	High	2	2	4
	Medium	4	2	6
	Low	3	1	4
				14

Abbildung 4: Entropie von Buys_computer und Income.

$$E(Buys\_computer, Income) =$$
$$P(High) * E(2,2) + P(Medium) * E(4,2) + P(Low) * E(3,1)$$
$$= (\tfrac{4}{14}) * 1 + (\tfrac{6}{14}) * 0.918 + (\tfrac{4}{14}) * 0.811$$
$$= 0.911$$

		Buys_computer		
		Yes	No	
Student	Yes	6	1	7
	No	3	4	7
				14

Abbildung 5: Entropie von Buys_computer und Student.

$$E(Buys\_computer, Student) = P(Yes) * E(6,1) + P(No) * E(3,4)$$
$$= (\tfrac{7}{14}) * 0.592 + (\tfrac{7}{14}) * 0.985$$
$$= 0.789$$

		Buys_computer		
		Yes	No	
Credit_ rating	Fair	6	2	8
	Excellent	3	3	6
				14

Abbildung 6: Entropie von Buys_computer und Credit_rating.

$$\begin{aligned} E(Buys\_computer, Credit\_rating) &= P(Fair) * E(6,2) + P(Excellent) * E(3,3) \\ &= (\tfrac{8}{14}) * 0.811 + (\tfrac{6}{14}) * 1 \\ &= 0.892 \end{aligned}$$

Berechnung des Information Gain

Mithilfe dieser Entropien kann jetzt der jeweilige Information Gain berechnet werden:

$$\begin{aligned} G(Buys\_computer, Age) &= E(Buys\_computer) - E(Buys\_computer, Age) \\ &= 0.94 - 0.693 \\ &= 0.247 \end{aligned}$$

$$\begin{aligned} G(Buys\_computer, Income) &= E(Buys\_computer) - E(Buys\_computer, Income) \\ &= 0.94 - 0.911 \\ &= 0.029 \end{aligned}$$

$$\begin{aligned} G(Buys\_computer, Student) &= E(Buys\_computer) - E(Buys\_computer, Student) \\ &= 0.94 - 0.789 \\ &= 0.151 \end{aligned}$$

$$\begin{aligned} &G(Buys\_computer, Credit\_rating) = \\ &E(Buys\_computer) - E(Buys\_computer, Credit\_rating) \\ &= 0.94 - 0.892 \\ &= 0.048 \end{aligned}$$

Das Attribut mit dem höchsten Information Gain wird jetzt als „decision node“ verwendet. In diesem Fall ist dies das Attribut `Age`.

Der Branch mit einer Entropie von 0 wird als „leaf node“ genutzt (Abbildung 7):

ID	Age	Income	Student	Credit_rating	Buys_computer
3	31...40	High	No	Fair	Yes
7	31...40	Low	Yes	Excellent	Yes
12	31...40	Medium	No	Excellent	Yes
13	31...40	High	Yes	Fair	Yes

Abbildung 7: Leaf node.

Branches mit einer Entropie von 0 oder mehr werden weiter aufgeteilt. Dies ist notwendig in den Fällen `Age <= 31` (Abbildung 8) sowie `Age > 40` (Abbildung 9):

ID	Age	Income	Student	Credit_rating	Buys_computer
1	<=30	High	**No**	Fair	No
2	<=30	High	**No**	Excellent	No
8	<=30	Medium	**No**	Fair	No
9	<=30	Low	**Yes**	Fair	Yes
11	<=30	Medium	**Yes**	Excellent	Yes

Abbildung 8: Splitting für Age <= 30.

ID	Age	Income	Student	Credit_rating	Buys_computer
4	>40	Low	Yes	**Excellent**	No
5	>40	Medium	No	**Excellent**	No
6	>40	Medium	No	**Fair**	Yes
10	>40	Low	Yes	**Fair**	Yes
14	>40	Medium	Yes	**Fair**	Yes

Abbildung 9: Splitting für Age >40.

Decision Tree

Aus diesen Berechnungen und Tabellen bildet sich dann folgender Decision Tree (Abbildung 10):

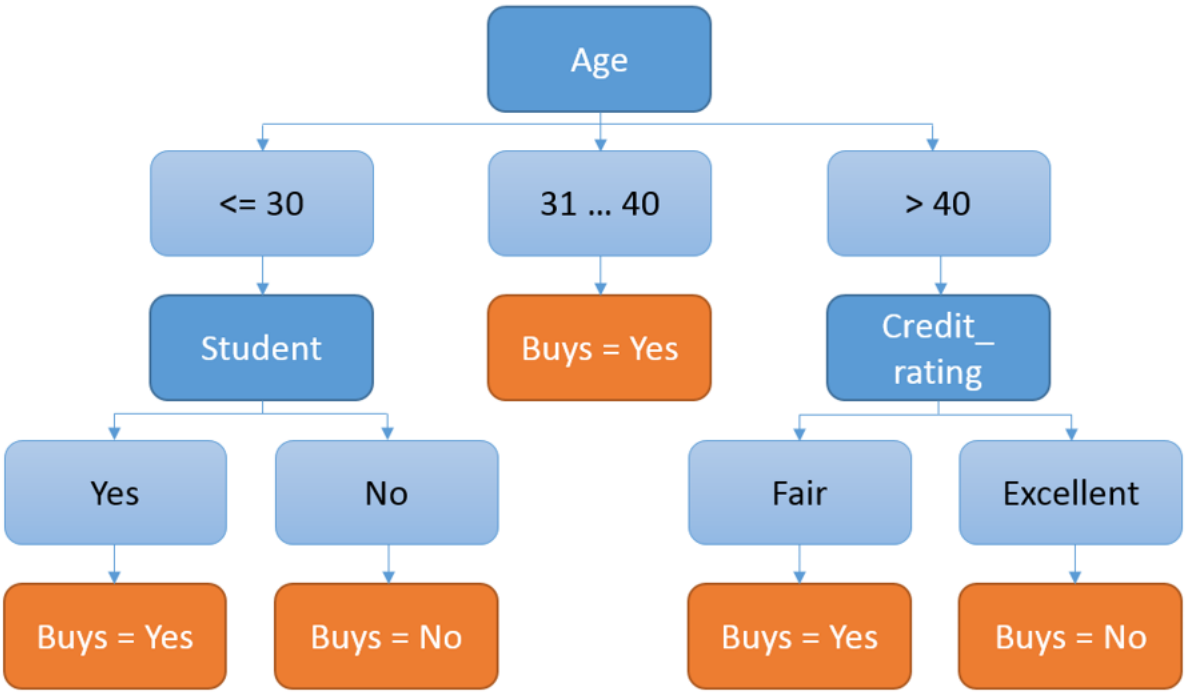

Abbildung 10: Der entwickelte Decision Tree.

Der Decision Tree wurde anschließend in Python implementiert. Der Code kann auf GitHub eingesehen werden:

https://github.com/thr0n/decision-tree

Test

Zum Test des Decision Trees werden die zu Trainingsdaten gehörenden Testdaten aus Moodle verwendet (Abbildung 11)

Quelle: http://moodle.oncampus.de/pluginfile.php/625704/mod_resource/content/1/DATA.gif

ID	Age	Income	Student	Credit_rating	Buys_computer
15	<=30	Medium	No	Excellent	No
16	<=30	Low	No	Fair	No
17	<=30	Low	No	Excellent	No
18	31...40	Low	Yes	Fair	Yes
19	>40	Medium	Yes	Excellent	Yes
20	31...40	High	No	Excellent	Yes

Abbildung 11: Die verwendeten Testdaten.

Beim Test der Anwendung stellte sich heraus, dass die Testfälle 15 bis 18 und 20 korrekt zugeordnet werden konnten. Lediglich der Testfall 19 wird zu `False` bzw. „No“ evaluiert. Grund hierfür ist die Tatsache, dass die Testdaten den Schluss zulassen, dass bei einem potentiellen Kunden, der älter als 40 ist, ausschließlich das `Credit_rating` Attribut betrachtet werden muss (vgl. Abbildung 9). Tatsächlich müssten in diesem Zweig aber weitere Attribute hinzugezogen werden (Student oder Income), um das gewünschte Ergebnis für `Buys_computer` zu erreichen.