Übungsaufgabe Entropie im Kurs DAML

1. Sie Sie möchten anhand von Merkmalen von Kreditnehmern vorhersagen, ob ein Kredit ausfallen wird oder nicht. Untersuchen Sie, ob das Merkmal „Saldo ≥ 50.000 Euro“ ein geeigneter Startpunkt für den Aufbau eines Entscheidungsbaums ist.

Gegeben, bereits ausgezählt, sind :Anzahl Kunden, Anzahl bekannter Kreditausfälle , Anzahl der Rückzahler 

Berechnen Sie den Informationsgewinn (IG), der sich ergibt, wenn die Grundgesamtheit anhand des Merkmals „Saldo < 50.000 Euro“ und „Saldo ≥ 50.000 Euro“ aufgeteilt wird.

Dokumentieren Sie jeden Zwischenschritt Ihrer Berechnung.

Runden Sie alle Ergebnisse auf zwei Nachkommastellen.

Erklären Sie abschließend kurz, warum ein hoher Informationsgewinn ein gutes Kriterium für die Auswahl eines Merkmals im Entscheidungsbaum ist.

Hinweise:

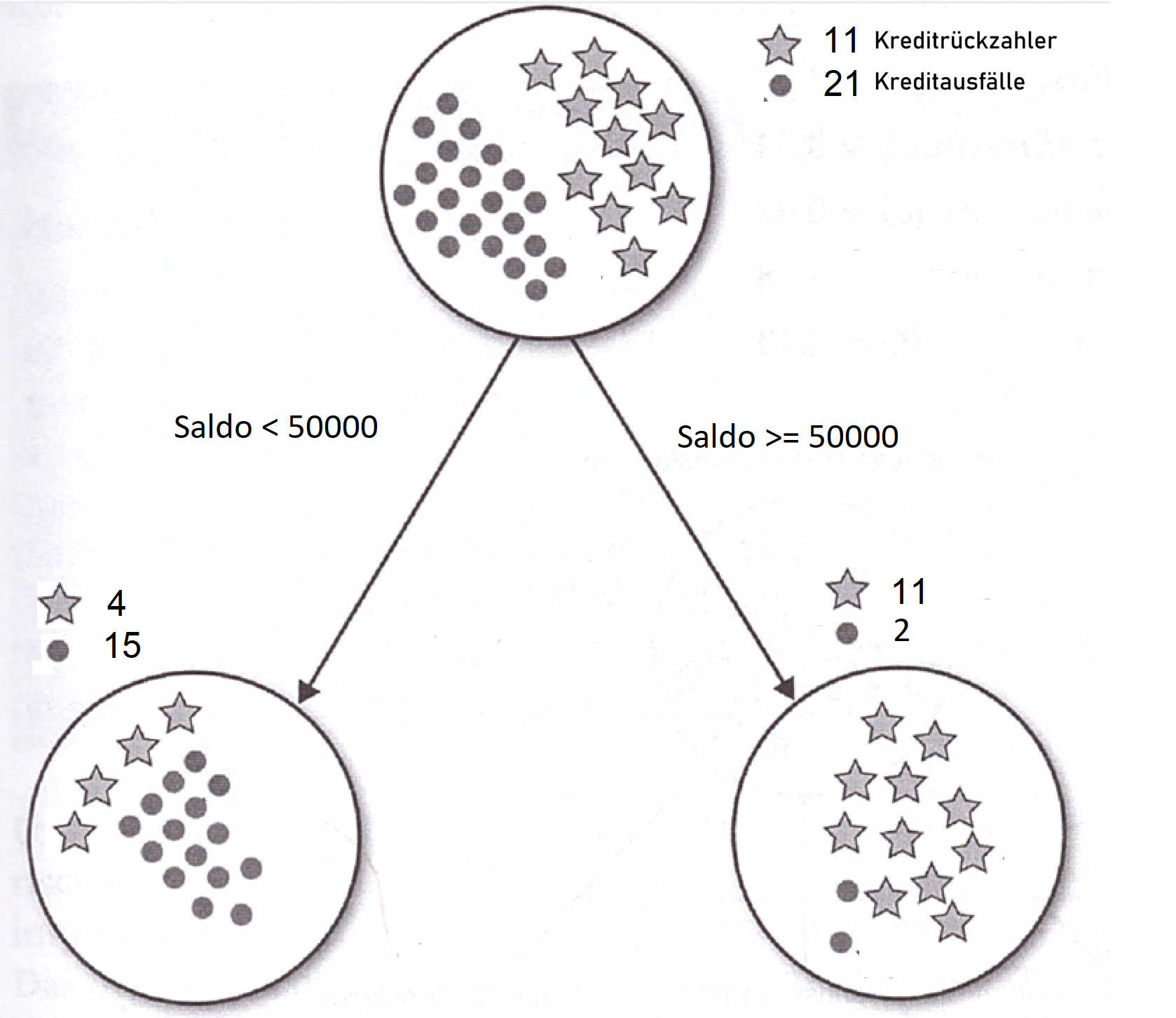
Entropie einer Menge = - p1 \* log2 (p1) – p2 \* log2 (p2)

mit p1 und p2 als die aus den Häufigkeiten je Klasse gewonnen Wahrscheinlichkeiten. Entnehmen Sie log2 aus der folgenden Tabelle.

Informationsgewinn = Entropie (Grundgesamtheit) – **(** (p1 \* Entropie (c1) + (p2 \* Entropie (c2) **)**

mit p1, p2 als Anteil der Teilmenge c1 und c2 in Bezug auf die Mächtigkeit der Elternmenge.

Erreichbare Punktzahl: 16



Notieren Sie Ihre Antwort auf einem neuen Blatt.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X** | **LOG Basis 2** |  | **X** | **LOG Basis 2** |  | **X** | **LOG Basis 2** |
| 0,01 | -6,64 |  | 0,34 | -1,56 |  | 0,67 | -0,58 |
| 0,02 | -5,64 |  | 0,35 | -1,51 |  | 0,68 | -0,56 |
| 0,03 | -5,06 |  | 0,36 | -1,47 |  | 0,69 | -0,54 |
| 0,04 | -4,64 |  | 0,37 | -1,43 |  | 0,70 | -0,51 |
| 0,05 | -4,32 |  | 0,38 | -1,40 |  | 0,71 | -0,49 |
| 0,06 | -4,06 |  | 0,39 | -1,36 |  | 0,72 | -0,47 |
| 0,07 | -3,84 |  | 0,40 | -1,32 |  | 0,73 | -0,45 |
| 0,08 | -3,64 |  | 0,41 | -1,29 |  | 0,74 | -0,43 |
| 0,09 | -3,47 |  | 0,42 | -1,25 |  | 0,75 | -0,42 |
| 0,10 | -3,32 |  | 0,43 | -1,22 |  | 0,76 | -0,40 |
| 0,11 | -3,18 |  | 0,44 | -1,18 |  | 0,77 | -0,38 |
| 0,12 | -3,06 |  | 0,45 | -1,15 |  | 0,78 | -0,36 |
| 0,13 | -2,94 |  | 0,46 | -1,12 |  | 0,79 | -0,34 |
| 0,14 | -2,84 |  | 0,47 | -1,09 |  | 0,80 | -0,32 |
| 0,15 | -2,74 |  | 0,48 | -1,06 |  | 0,81 | -0,30 |
| 0,16 | -2,64 |  | 0,49 | -1,03 |  | 0,82 | -0,29 |
| 0,17 | -2,56 |  | 0,50 | -1,00 |  | 0,83 | -0,27 |
| 0,18 | -2,47 |  | 0,51 | -0,97 |  | 0,84 | -0,25 |
| 0,19 | -2,40 |  | 0,52 | -0,94 |  | 0,85 | -0,23 |
| 0,20 | -2,32 |  | 0,53 | -0,92 |  | 0,86 | -0,22 |
| 0,21 | -2,25 |  | 0,54 | -0,89 |  | 0,87 | -0,20 |
| 0,22 | -2,18 |  | 0,55 | -0,86 |  | 0,88 | -0,18 |
| 0,23 | -2,12 |  | 0,56 | -0,84 |  | 0,89 | -0,17 |
| 0,24 | -2,06 |  | 0,57 | -0,81 |  | 0,90 | -0,15 |
| 0,25 | -2,00 |  | 0,58 | -0,79 |  | 0,91 | -0,14 |
| 0,26 | -1,94 |  | 0,59 | -0,76 |  | 0,92 | -0,12 |
| 0,27 | -1,89 |  | 0,60 | -0,74 |  | 0,93 | -0,10 |
| 0,28 | -1,84 |  | 0,61 | -0,71 |  | 0,94 | -0,09 |
| 0,29 | -1,79 |  | 0,62 | -0,69 |  | 0,95 | -0,07 |
| 0,30 | -1,74 |  | 0,63 | -0,67 |  | 0,96 | -0,06 |
| 0,31 | -1,69 |  | 0,64 | -0,64 |  | 0,97 | -0,04 |
| 0,32 | -1,64 |  | 0,65 | -0,62 |  | 0,98 | -0,03 |
| 0,33 | -1,60 |  | 0,66 | -0,60 |  | 0,99 | -0,01 |

Elternmenge:

Gesamtanzahl 32

21 Kreise, 11 Sterne -> p(Kreise)=21/32=0,66 p(Sterne)=11/32=0,34

Entropie (Elternmenge) = - (0,66\*(-0,60) + 0,34\*(-1,56))

= -(-0,40 -0,53) = 0,93

Teilmenge links:

Anzahl 19

15 Kreise, 4 Sterne -> p(Kreise)=15/19=0,79, p(Sterne)=4/19=0,21

Entropie (Teilmenge 1) = - (0,79\*-0,34 + 0,21\* -2,25)

=- (-0,27 - 0,47) = 0,74

Teilmenge rechts:

Anzahl 13

2 Kreise, 11 Sterne -> p(Kreise)=2/13=0,15, p(Sterne)=11/13=0,85

Entropie (Teilmenge 2) = - (0,15\*-2,74 + 0,85\* -0,23)

= - (-0,41 – 0,20) = 0,61

Wahrscheinlichkeiten der Aufteilung in Teilmenge 1 und 2:

P1 = 19/32 = 0,59

P2 = 13/32 = 0,41

Informationsgewinn der Teilung:

IG= 0,93 – (0,59 \* 0,74 + 0,41 \* 0,61)

= 0,93 – (0,44 + 0,25)

= 0,93 – 0,69 = 0,24

Die Teilung der Grundmenge nach Merkmal Saldo bei 50000 führt also zu einem IG von 0,24. Nun kann der IG für andere Teilungen des Saldos, z.B. bei 40000, berechnet werden. Für einen Knoten wird unter allen verfübaren Variablen und Teilungen der größte Informationsgewinn berechnet und zur Teilung eines Knoten verwendet. Diese Verfahren wird Top-Down angewendet bis keine Daten mehr aufteilbar sind.