Data Mining – Hints

Prof. Dr. Jürgen Cleve / Prof. Dr. Uwe Lämmel

Table of Contents

Below are some selected solutions for the exercises in the Data Mining book.
If you require further help with the solutions, please let us know.

1 Introduction

2 Foundations

Task 2.1

- Postcodes are not numbers, and they are generally not subject to any order. In this respect, only the nominal type remains here.
- The same applies to ISBNs.
- Fuel consumption and vehicle performance are numbers that can be calculated. Metric.
- House numbers are at least subject to an order. Ordinal.

Task 2.3

- King = Maximum.
- Rook = Manhattan.
- If you only count the moves of the rook, then it's Hamming.

3 Application classes

4 Knowledge representation

5 Classification

Task 5.1

Of course, we can solve the problem using Hamming distance. However, we should only use this as a last resort.
Numbers are ideal. Do we have a chance of converting all attributes into numbers? YES:

- The binary attributes *Alternative*, *Fri/Sat*, *Hungry*, *Reservation* can simply be converted to 0/1.
  (Additional question: Does it matter whether we use 0/1 or 1/0?)
- *Guests* is ordinal. So *none = 0*, *some = 0.5*, *full = 1* would work.
- However, *type* is nominal, so only binary coding remains here.

Now we can apply kNN using Euclidean distance.

Task 5.2

Analog.

Task 5.3

Similarly. However, we need to convert the non-numeric attributes into numbers and normalize the metric attributes.

Task 5.4

If we want to use ID3, we first need to convert the metric attributes into intervals.

Task 5.5

Table 1 contains 12 data records with the following attributes:

- Alternative: Is there another suitable restaurant nearby? {yes, no}
- Fri/Sat: Is it Friday or Saturday? {yes, no}
- Hungry: Am I hungry? {yes, no}
- Guests: How many people are in the restaurant? {none, a few, full}
- Reservation: Have I made a reservation? {yes, no}
- Type: What kind of restaurant is it? {French, Chinese, Italian, Burger}
- Waiting time: How long does the restaurant estimate the wait will be? {0–10, 10–30, 30–60, >60}
- Waiting (target attribute): Am I waiting? {yes, no}

Generate a decision tree and classify the last 3 data records in the table.

Table 1: Restaurant example data

Alternative	Fri/Sat	Hungry	Guests	Reservation	Type	Waiting time	Waiting
yes	no	yes	a few	yes	French	0–10	yes
yes	no	yes	full	no	Chinese	30–60	no
no	no	no	a few	no	Burger	0–10	yes
yes	yes	yes	full	no	Chinese	10–30	yes
yes	yes	no	full	yes	French	>60	no
no	no	yes	a few	yes	Italian	0–10	yes
no	no	no	none	no	Burger	0–10	no
no	no	yes	a few	yes	Chinese	0–10	yes
no	yes	no	full	no	Burger	>60	no
yes	yes	yes	full	yes	Italian	10–30	no
no	no	no	none	no	Chinese	0–10	no
yes	yes	yes	full	no	Burger	30–60	yes
yes	no	yes	a few	no	French	30–60	
yes	yes	yes	full	yes	Chinese	10–30	
no	no	no	none	no	Burger	0–10	

Entropy of Table 1?

- $p(\text{Waiting yes}) = \frac{6}{12}$
- $p(\text{Waiting no}) = \frac{6}{12}$

Entropy: $I(\text{Table}) = I(\text{Waiting}) = -\frac{6}{12} * \log_2(\frac{6}{12}) - \frac{6}{12} * \log_2(\frac{6}{12}) = 1$
Now we calculate the **Information gain** for all attributes.
We choose the attribute **Alternative**, which has 2 values: **yes, no**. Now we count, how often we wait:

Value	waiting	not waiting
yes	3	3
no	3	3

We get:

$$I(\text{Alt\_yes}) = -\frac{3}{6} * \log_2(\frac{3}{6}) - \frac{3}{6} * \log_2(\frac{3}{6}) = 1$$

$$I(\text{Alt\_no}) = -\frac{3}{6} * \log_2(\frac{3}{6}) - \frac{3}{6} * \log_2(\frac{3}{6}) = 1$$

So:

$$G(\text{Alternative}) = \sum_{i=1}^{n} \frac{|E_i|}{|E|} I(E_i) = \frac{6}{12} * 1 + \frac{6}{12} * 1 = 1$$

Analog:

Fr/Sa	0.98
Hungry	0.804
Guests	0.459
Reservation	0.98
Type	1
Waiting time	0.79

We get:

- gain(Alternative) $= 1 - 1 = 0$
- gain(Fr/Sa) $= 1 - 0.98 = 0.02$
- gain(Hungry) $= 1 - 0.804 = 0.196$
- gain(Guests) $= 1 - 0.459 = 0.541$
- gain(Reservation) $= 1 - 0.98 = 0.02$
- gain(Type) $= 1 - 1 = 0$
- gain(Waiting time) $= 1 - 0.79 = 0.21$

We choose **"Guests"** as root attribute. We continue the process recursively for each subtree. First, we consider the subtree for **Guests** = **full**. Only the following records remain in consideration:

Alt.	Fr/Sa	Hung.	Guests	Reserv.	Type	Waiting time	Waiting
yes	no	yes	full	no	Chin.	30–60	no
yes	yes	yes	full	no	Chin.	10–30	yes
yes	yes	no	full	yes	Franz.	>60	no
no	yes	no	full	no	Burger	>60	no
yes	yes	yes	full	yes	Ital.	10–30	no
yes	yes	yes	full	no	Burger	30–60	yes

We get:

- gain(Alt) $= 0.92 - 0.81 = 0.11$
- gain(Fr/Sa) $= 0.92 - 0.81 = 0.11$
- gain(Hungry) $= 0.92 - 0.67 = 0.25$
- gain(Reservation) $= 0.92 - 0.67 = 0.25$
- gain(Type) $= 0.92 - 0.67 = 0.25$
- gain(Waiting time) $= 0.92 - 0.67 = 0.25$

The other subtrees for **Guests** directly return a class.

Task 5.6 und 5.7

We can also use the ID3 here without any preprocessing.

Task 5.9

Attribute	r	o
Occupation = e	$\frac{2}{5}$	$\frac{2}{4}$
Occupation = f	$\frac{3}{5}$	$\frac{2}{4}$
Fam = m	$\frac{4}{5}$	$\frac{2}{4}$
Fam = s	$\frac{1}{5}$	$\frac{2}{4}$
Children = j	$\frac{1}{5}$	$\frac{3}{4}$
Children = n	$\frac{4}{5}$	$\frac{1}{4}$
Debts = j	$\frac{3}{5}$	$\frac{4}{4}$
Debts = n	$\frac{2}{5}$	$\frac{0}{4}$

- $P(M|\{a,l,j,n\} = \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{9} = \frac{4}{5 \cdot 5 \cdot 5 \cdot 9}$
- $P(E|\{a,l,j,n\} = \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{3}{4} \cdot \frac{0}{4} \cdot \frac{4}{9} = 0$

Prediction: Renter.

With Laplace correction:

Attribute	r	o
Occupation = e	$\frac{3}{7}$	$\frac{3}{6}$
Occupation = f	$\frac{4}{7}$	$\frac{3}{6}$
Fam = m	$\frac{5}{7}$	$\frac{3}{6}$
Fam = s	$\frac{2}{7}$	$\frac{3}{6}$
Children = j	$\frac{2}{7}$	$\frac{4}{6}$
Children = n	$\frac{5}{7}$	$\frac{2}{6}$
Debts = j	$\frac{4}{7}$	$\frac{5}{6}$
Debts = n	$\frac{3}{7}$	$\frac{1}{6}$

- $P(M|\{a,l,j,n\} = \frac{3}{7} \cdot \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{3}{7} \cdot \frac{5}{9} = \frac{20}{49 \cdot 49} = \frac{20}{2401} = 0.0083$
- $P(E|\{a,l,j,n\} = \frac{3}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{1}{6} \cdot \frac{4}{9} = \frac{2}{81} = 0.0247$

Prediction: *Owner* with 74.77% (*Miete*: 25.23%) .

6 Clustering

Task 6.3

Before we can apply k-means, we need to preprocess the data:

1. We can remove the *ID column*.
2. We convert *Moisture* to 0/1.
3. *Acidity* is ordinal. We can convert the data to 0 / 0.5 / 1.
4. *Class* is binary and can be converted to 0/1.
5. We should normalize *Temp*. Interval $[0, 1]$.

7 Association analysis

Task 7.1

- $supp(Saft \rightarrow Cola) = \frac{4}{6}, \quad conf(Saft \rightarrow Cola) = \frac{4}{5}$

Task 7.2

We abbreviate the items using their numbers: 123, 124, 134, 234, 235, 345, 357, 245, 367. Potential 4 FIS candidates are: 1234, 1235, 1245, 1345, 2345, 2357, 3457, 3567. However, most of them are eliminated through pruning:

- 1235 (because of 125)

- 1245 (because of 125)
- 1345 (because of 135)
- 2345 (because of 245)
- 2357 (because of 237)
- 3457 (because of 347)
- 3567 (because of 567)

That leaves only 1234 as a 4 FIS candidate.
There can be no 5 FIS. For that to happen, there would have to be at least 5 FIS_4 candidates.

8 Data preparation

Task 8.3

- **No.** can be removed.
- The ordinal attributes must be converted to numbers.
- **Age** should be normalized.

9 Evaluation