

Datascience: halfway meeting
Datamining

Karin Groothuis-Oudshoorn

7-10-2020

Paper and pencil exercises: Exercise 2.4

A	B	Number of Instances	
		Y	O
a1	b1	4	10
a2	b1	6	2
a3	b1	8	6
a1	b2	2	8
a2	b2	6	2

- ▶ Rule of Bayes: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
- ▶ $P(Y|a3, b2) = \frac{P(a3, b2|Y) \cdot P(Y)}{P(a3, b2)}$
- ▶ Unknown: $P(a3, b2|Y)$ and $P(a3, b2)$

Naive Bayes

A	B	Number of Instances	
		Y	O
a1	b1	4	10
a2	b1	6	2
a3	b1	8	6
a1	b2	2	8
a2	b2	6	2

- ▶ Naive Bayes: $P(a3, b2) = P(a3) \cdot P(b2)$ and $P(a3, b2|Y) = P(a3|Y) \cdot P(b2|Y)$
- ▶ $P(a3, b2) = P(a3) \cdot P(b2) = \frac{8+6}{54} \cdot \frac{2+8+6+2}{54}$
- ▶ $P(a3|Y) = \frac{8}{26}$

Paper and pencil exercises: Exercise 2.5

A	Y	O	label	number of errors
a1	6	18	O	6
a2	12	4	Y	4
a3	8	6	Y	6

- ▶ classification error rate for attribute A equals $\frac{16}{54} = 0.296$
- ▶ Gini index for attribute A: $GI(\text{node } m) = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$,
where \hat{p}_{mk} is the proportion of class k observations in node m
 - ▶ $GI(a1) : \frac{6}{24} \times \frac{18}{24} + \frac{18}{24} \times \frac{6}{24} = 0.375$
 - ▶ $GI(a2) : 2 \times \frac{12}{16} \times \frac{4}{16} = 0.375$
 - ▶ $GI(a3) : 2 \times \frac{8}{14} \times \frac{6}{14} = 0.490$
 - ▶ Overall Gini index for attribute A:
 $GI(A) = \frac{24}{54} \times 0.375 + \frac{16}{54} \times 0.375 + \frac{14}{54} \times 0.490 = 0.4048$

Exercise 2.5 (II)

B	Y	O	label	number of errors
$b1$	18	18	O	18
$b2$	8	10	Y	8

- ▶ classification error rate for attribute B equals $\frac{26}{54} = 0.481$
- ▶ Gini index for attribute B :
 - ▶ $GI(b1) : \frac{18}{36} \times \frac{18}{36} + \frac{18}{36} \times \frac{18}{36} = 0.5$
 - ▶ $GI(b2) : 2 \times \frac{8}{18} \times \frac{10}{18} = 0.4938$
 - ▶ Overall Gini index for attribute B :
 $GI(B) = \frac{36}{54} \times 0.5 + \frac{18}{54} \times 0.4938 = 0.498$

Projects

- ▶ PSCD (DPV / DM)
- ▶ MOCHA (primary DPV)
- ▶ COVID (DM)

DM project: PSCD

The dataset contains data of 4087 real surgeries of patients.

Aim:

- ▶ identify patterns in surgical case durations
- ▶ derive prediction models for the surgery time in order to decrease overtime at the TCT

Suggestions:

- ▶ Focus on a subset of the surgeries (the most prevalent, or only one type)
- ▶ Recode the surgeries into less categories
- ▶ Focus on patient characteristics or process characteristics (e.g. the physicians)

Challenges:

- ▶ a lot of missing data (multiple imputation with `mice()`)

DPV project: MOCHA

The dataset contains data of 2640 questionnaires.

Aim: What are the priorities of European citizens in assessing the quality of primary care for children in Europe?

Suggestions:

- ▶ Calculate best-worst counts (e.g. per country) for all attribute-items
- ▶ Look at differences between countries or other background characteristics

Challenges:

- ▶ merge the information in the separate datasets to get clear tables for a star schema.

COVID19

The dataset contains data of 375 patients.

Aim: Develop a prediction model for the mortality of COVID-19 patients based on the biomarkers that are available in the data

Suggestions:

- ▶ choose e.g. 3 appropriate machine learning models, train the models and compare their accuracy on one or more appropriate accuracy measures.
- ▶ validate all model steps: so including the feature selection

Challenges:

- ▶ extract the latest measurement per patient (or use all data, when you get a longitudinal dataset)

Obligatory items in your DM project

- ▶ Identify what kind of DM problem it is
- ▶ Well formulated research questions
- ▶ Design of a valid DM pipeline (including feature construction and selection)
- ▶ Comparison of different DM models
- ▶ Assessment of the performance of the constructed DM models in a sound way (with test/train set, Cross-validation or bootstrap)
- ▶ Critical reflection: strenghts and weaknesses of the methodology, results: place the results in the context of the problem