Naïve Bayes Classifier
Data Warehousing and Data Mining

Patrick Schäfer
Berlin, 22. Januar 2018
patrick.schaefer@hu-berlin.de

Vorlesung: https://hu.berlin/vl_dwhdm17
Übung: https://hu.berlin/ue_dwhdm17

Example

- We are given a list of donors
- What is the probability of someone being a donor at the age above 50?
  $p(Donor|Age>50)$
- What is the probability of someone being a donor at the age above 50 and with >50k salary?
  $p(Donor|Age>50, Salary > 50k)$

Name	Age	Salary	Donor?
Nancy	21	37k	N
Jim	27	41k	N
Allen	43	61k	Y
Jane	38	55k	N
Steve	44	30k	N
Peter	51	56k	Y
Sayani	53	70k	Y
Lata	56	74k	Y
Mary	59	25k	N
Vitor	61	68k	Y
Dale	63	51k	Y

Bayes Theorem

- Bayes' theorem for conditional probabilities:

$$\mathrm{p}(c|f_1, \dots, fn) = \frac{p(f_1, \dots, f_n|c) \cdot p(c)}{p(f_1, \dots, f_n)} \propto p(f_1, \dots, f_n|c) \cdot p(c)$$

is proportional to

constant wrt. the class c

- The a-priori probability $p(f)$ of every feature f
  - How many entries from T have f?
- The a-priori probability $p(c)$ of every class $c \in C$
  - How many entries in T are of class c?
- The conditional probabilities $p(f|c)$ for feature f being true in class c
  - Proportion of entries in c with feature f among all entries in c

Naïve Bayes

- For some feature combinations $f_1, \ldots, f_n$ there may not be a single instance
- "Naïve": thus, we assume statistical independence:

$$\begin{aligned} p(c|f_1, \ldots, f_n) &\propto p(f_1, \ldots, f_n|c) \cdot p(c) \\ &\propto p(f_1|c) \cdot \ \ldots \ \cdot p(f_n|c) \cdot p(c) \\ &\propto p(c) \cdot \prod_{i=1}^{n} p(f_i|c) \end{aligned}$$

- Naïve Bayes Classification:
  - pick the class c with the maximum conditional probability $p(c|f_1, \ldots, f_n)$

Example

$$p(Donor/Age{>}50) \propto p(Age > 50|Donor) \cdot p(Donor)$$

Name	Age	Salary	Donor?
Nancy	21	37k	N
Jim	27	41k	N
Allen	43	61k	Y
Jane	38	55k	N
Steve	44	30k	N
Peter	51	56k	Y
Sayani	53	70k	Y
Lata	56	74k	Y
Mary	59	25k	N
Vitor	61	68k	Y
Dale	63	51k	Y

$$P(Donor) = \frac{6}{11}$$

$$P(Age > 50|Donor) = \frac{5}{6}$$

$$P(Donor \mid Age > 50)$$
$$\propto \frac{6}{11} \cdot \frac{5}{6} = \frac{5}{11}$$

$$P(\neg Donor \mid Age > 50)$$
$$\propto \frac{5}{11} \cdot \frac{1}{5} = \frac{1}{11}$$

Example

$$p(Donor|Age > 50, Salary > 50k) \propto \cdots ?$$

Name	Age	Salary	Donor?
Nancy	21	37k	N
Jim	27	41k	N
Allen	43	61k	Y
Jane	38	55k	N
Steve	44	30k	N
Peter	51	56k	Y
Sayani	53	70k	Y
Lata	56	74k	Y
Mary	59	25k	N
Vitor	61	68k	Y
Dale	63	51k	Y

$$P(Donor) = \frac{6}{11}$$

$$P(Age > 50|Donor) = \frac{5}{6}$$

$$P(Salary > 50k|Donor) = \frac{6}{6}$$

$$P(Donor \mid A > 50, S > 50k)$$
$$\propto \frac{6}{11} \cdot 1 \cdot \frac{5}{6} = \frac{5}{11}$$

$$P(\neg Donor \mid A > 50, S > 50k)$$
$$\propto \frac{5}{11} \cdot \frac{1}{5} \cdot \frac{1}{5} = \frac{1}{55}$$

(Conditional) Probabilities in SQL

- $p(x)$:

```
SELECT x, count(*) / sum(count(*)) over () as percent
  FROM table
GROUP BY x
```

- $p(x|c)$: …?

- See: Oracle Analytic functions:
  - https://docs.oracle.com/cd/E11882_01/server.112/e41084/functions004.htm