5. Aufgabenblatt
- Auswertung -

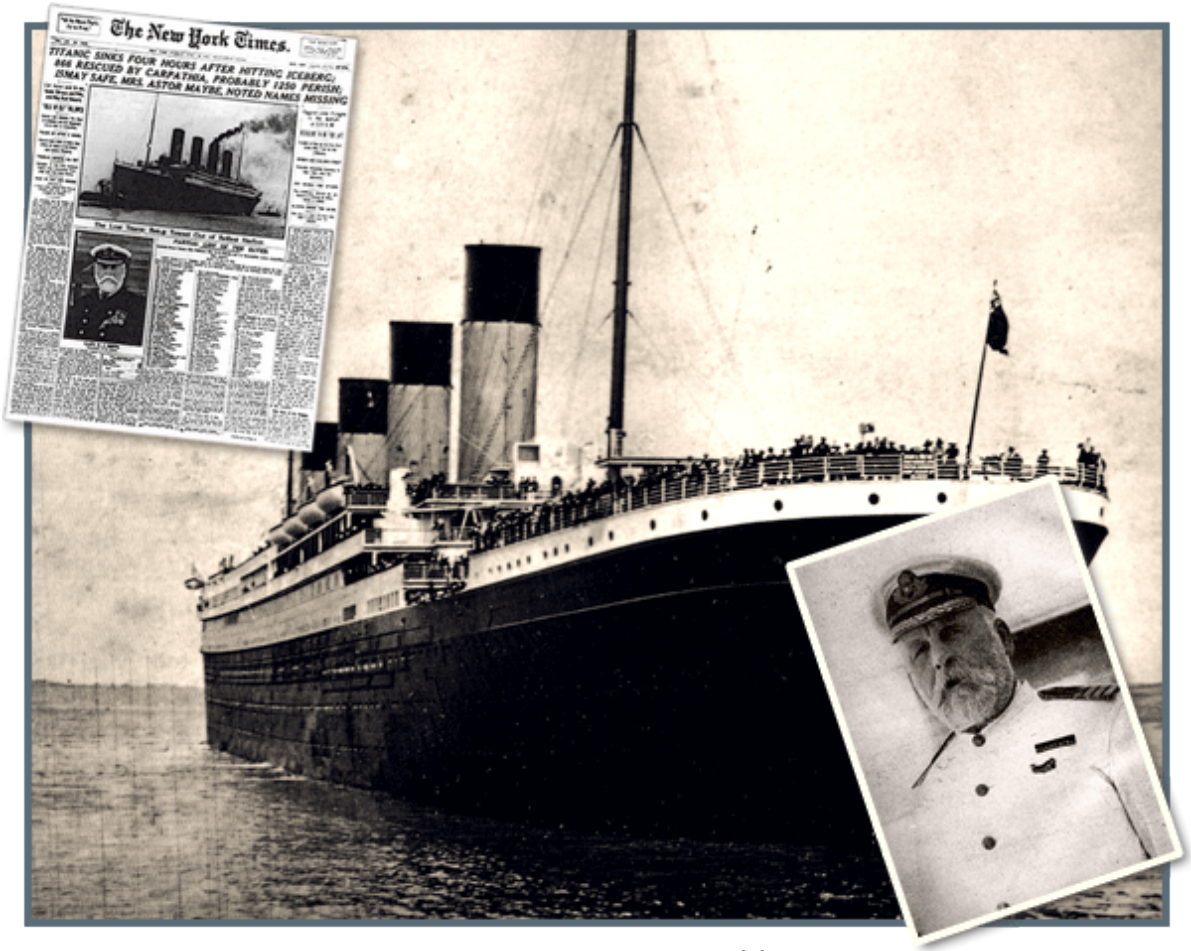
Patrick Schäfer

Berlin, 12. Februar 2017

patrick.schaefer@hu-berlin.de

Daten (Kreuzfahrt)

- Ein Veranstalter möchte für die Planung zukünftiger Kreuzfahrten vorhersagen, welche seiner Gäste Spaß haben werden
- Gegeben sind die folgenden Daten der Gäste einer Kreuzfahrt:
  - ID, LABEL, PCLASS, NAME, SEX, AGE, FARE, EMBARKED
    - LABEL: Der Gast hatte Spaß (1) oder nicht (0)
    - PLCASS: Ticket-Klasse
    - Name: Name des Gasts
    - Sex: Geschlecht
    - FARE: Ticket-Kosten
    - AGE: Alter des Gasts, falls bekannt
    - EMBARKED: Eingeschifft im Hafen



From: http://www.titanicfacts.net

Facts

- Dataset: Titanic Passagiere
  - 492 Überlebende
  - *„Women and children first“*
    - 20% der Männer
    - 75% der Frauen
    - 109 Kinder, davon haben 56 überlebt
  - Nach Klassen:
    - 61% First Class (alle bis auf 1 Kind)
    - 42% Second Class (alle Kinder)
    - 24% Third Class (alle bis auf 52 Kinder)

Anmerkung

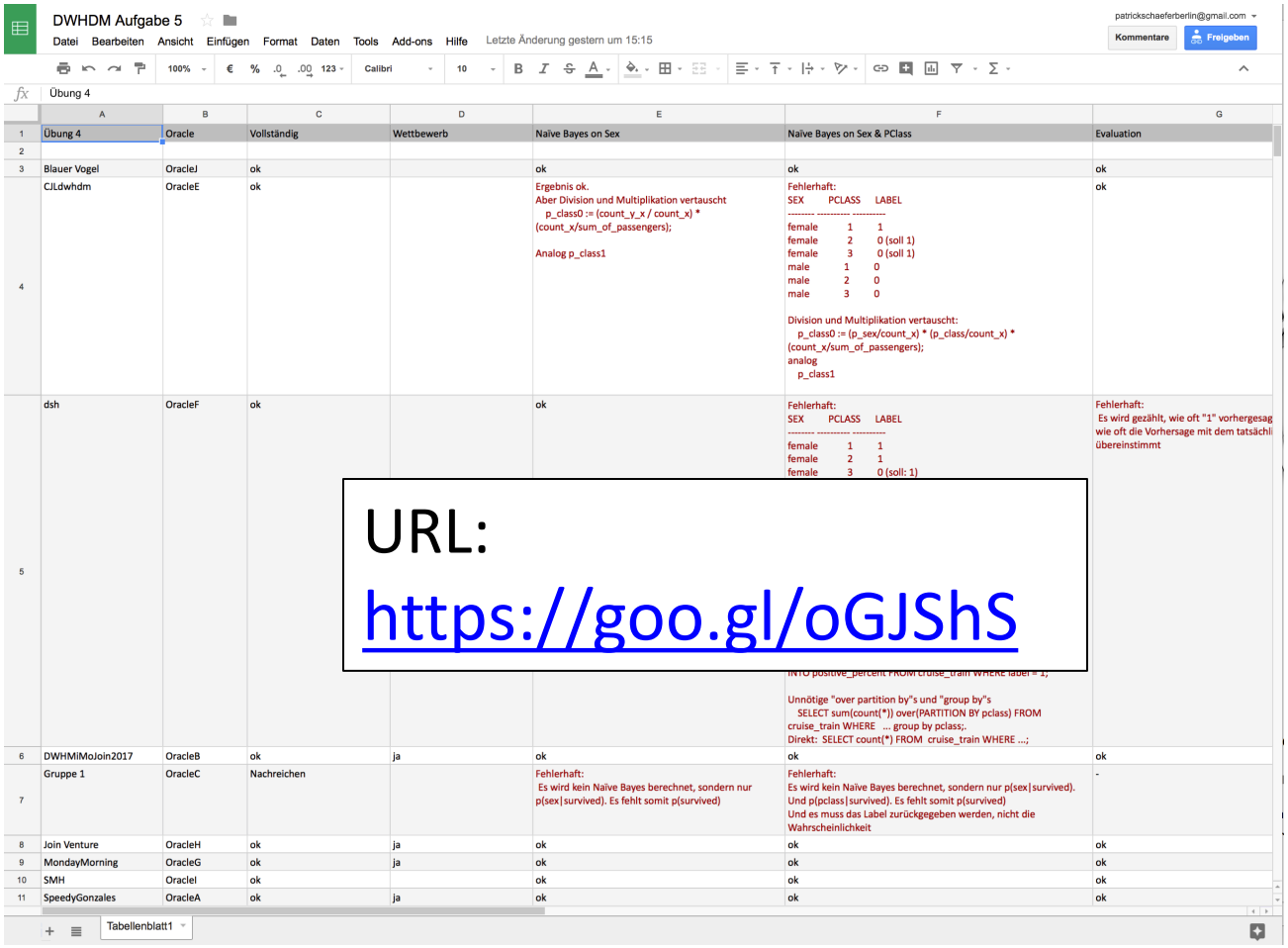
- Naïve Bayes:
  1. $p(Fun|Sex) \propto p(Sex|Fun) \cdot p(Fun)$
  2. $p(\neg Fun|Sex) \propto p(Sex|\neg Fun) \cdot p(\neg Fun)$
  3. $p(Fun|Sex, PClass) \propto p(Sex|Fun) \cdot p(PClass|Fun) \cdot p(Fun)$
  4. $p(\neg Fun|Sex, PClass) \propto p(Sex|\neg Fun) \cdot p(PClass|\neg Fun) \cdot p(\neg Fun)$

- Was wir also brauchen:
  1. $p(Fun)$
  2. $p(\neg Fun)$
  3. $p(Sex|Fun)$
  4. $p(Sex|\neg Fun)$
  5. $p(PClass|Fun)$
  6. $p(PClass|\neg Fun)$

12 Abgaben

Gruppe
Blauer Vogel
CJLdwhdm
dsh
DWHMiMoJoin2017
Gruppe 1
JoinVenture
MondayMorning
SMH
SpeedyGonzales
TeamHA9876
wwgds
GundDan

Anmerkungen im Detail



DWHDM Aufgabe 5

Datei Bearbeiten Ansicht Einfügen Format Daten Tools Add-ons Hilfe Letzte Änderung gestern um 15:15

patrickschaeferberlin@gmail.com

Kommentare Freigeben

Übung 4

	A	B	C	D	E	F	G
1	Übung 4	Oracle	Vollständig	Wettbewerb	Naïve Bayes on Sex	Naïve Bayes on Sex & PClass	Evaluation
2							
3	Blauer Vogel	OracleJ	ok		ok	ok	ok
4	CJLdwhdm	OracleE	ok		Ergebnis ok. Aber Division und Multiplikation vertauscht p_class0 := (count_y_x / count_x) * (count_x/sum_of_passengers); Analog p_class1	Fehlerhaft: SEX PCLASS LABEL -------- ---------- ---------- female 1 1 female 2 0 (soll 1) female 3 0 (soll 1) male 1 0 male 2 0 male 3 0 Division und Multiplikation vertauscht: p_class0 := (p_sex/count_x) * (p_class/count_x) * (count_x/sum_of_passengers); analog p_class1	ok
5	dsh	OracleF	ok		ok	Fehlerhaft: SEX PCLASS LABEL -------- ---------- ---------- female 1 1 female 2 1 female 3 0 (soll: 1) Unnötige "over partition by"s und "group by"s SELECT sum(count(*)) over(PARTITION BY pclass) FROM cruise_train WHERE ... group by pclass;. Direkt: SELECT count(*) FROM cruise_train WHERE ...;	Fehlerhaft: Es wird gezählt, wie oft "1" vorhergesag wie oft die Vorhersage mit dem tatsächl übereinstimmt
6	DWHMiMoJoin2017	OracleB	ok	ja	ok	ok	ok
7	Gruppe 1	OracleC	Nachreichen		Fehlerhaft: Es wird kein Naïve Bayes berechnet, sondern nur p(sex|survived). Es fehlt somit p(survived)	Fehlerhaft: Es wird kein Naïve Bayes berechnet, sondern nur p(sex|survived). Und p(pclass|survived). Es fehlt somit p(survived) Und es muss das Label zurückgegeben werden, nicht die Wahrscheinlichkeit	-
8	Join Venture	OracleH	ok	ja	ok	ok	ok
9	MondayMorning	OracleG	ok	ja	ok	ok	ok
10	SMH	OracleI	ok		ok	ok	ok
11	SpeedyGonzales	OracleA	ok	ja	ok	ok	ok

Tabellenblatt1

URL:

https://goo.gl/oGJShS

Präsentation

- Montags (Gruppe 1)
  https://dudle.inf.tu-dresden.de/dwhdm_mo_ue5/

- Mittwochs (Gruppe 2)
  https://dudle.inf.tu-dresden.de/dwhdm_mi_ue5/

Lösung für Naïve-Bayes

- $p(Fun)$ und $p(\neg Fun)$

```
SELECT label, count(*) / sum(count(*)) over () as perc
  FROM cruise_train
 GROUP BY label
```

- $p(Sex|Fun)$ und $p(Sex|\neg Fun)$

```
SELECT label, sex, count(*)/sum(count(*)) over (PARTITION BY label) as perc
  FROM cruise_train
 GROUP BY label, sex
```

- $p(Fun|Sex) \propto p(Sex|Fun) \cdot p(Fun)$ und $p(\neg Fun|Sex) \propto \cdots$

```
SELECT p_fun.label, (p_fun.perc * p_sex_given_fun.perc) as perc
  FROM p_fun JOIN p_sex_given_fun ON m_sex = p_sex_given_fun.sex
 WHERE p_sex_given_fun.label = p_fun.label
```

- Naïve Bayes Classifier (Maximum):

```
SELECT p.label INTO prediction
  FROM … // p(Fun|Sex)
 WHERE p.perc = (SELECT max(perc) from posterior);
```

Wettbewerb: 4 von 12 Abgaben

Gruppe	Wettbewerb
Blauer Vogel	
CJLdwhdm	
dsh	
DWHMiMoJoin2017	
Gruppe 1	
JoinVenture	
MondayMorning	
SMH	
SpeedyGonzales	
TeamHA9876	
wwgds	
GrundDan	

Ansätze

Gruppe	Ansatz
DWHMiMoJoin2017	Naïve Bayes auf Sex, PClass, Fare und Age?
JoinVenture	Naïve Bayes auf Sex und Pclass?
MondayMorning	Decision Tree auf Fare, Sex, PClass, Age
SpeedyGonzales	Naïve Bayes auf Sex und PClass?

Genauigkeiten

Gruppe	Train	Test
DWHMiMoJoin2017	79,5%	
JoinVenture	78,3% (?)	
MondayMorning	85,7%	
SpeedyGonzales	78,3% (?)	

Genauigkeiten

Gruppe	Train	Test
DWHMiMoJoin2017	79,5%	77,1%
JoinVenture	78,3% (?)	77,5%
MondayMorning	85,7%	**85,1%**
SpeedyGonzales	78,3% (?)	77,5%

Punkte (Blatt 5)

1.	MondayMorning	5
2.	JoinVenture	3
	SpeedyGonzales	3
3.	DWHMiMoJoin2017	2

Gesamtpunkte (Blatt 1+2+4+5)

1.	MondayMorning:	16
2.	DWHMiMoJoin2017:	8.5
3.	Blauer Vogel:	7
4.	SpeedyGonzales:	5.5
5.	JoinVenture:	5
	CJLdwhdm:	5
6.	SMH:	2.5
7.	TeamHA9876:	2
	Gruppe 1:	2
	Dsh:	2
	Wwgds:	2

Klausurtermin

- Die Klausur findet am 23.02.18 statt:
  - Ort: RUD 25, 3.001
  - Einlass: 9:00 Uhr
  - Beginn: 9:30 Uhr
  - Dauer: 120 Minuten

Lehrevaluation

- Kommentare
  - Lockere Atmosphäre
  - Wettbewerb
  - Schön auf Fragen eingegangen und darum gekümmert, dass die Frage wirklich geklärt ist :)
  - sehr nett
  - schnelle Antwort auf E-Mails
  - Motiviert
  - Übungsbeispiele

  - Die Besprechung der vergangenen Übungsaufgabe war zumindest an den ersten Terminen etwas langwierig
  - Inhaltlich wäre eine längerfristige, in Gruppen zu bearbeitende Fallstudie, die die einzelnen Schwerpunkte abdeckt (vom Import aus heterogenen Datenquellen (ETL), über Schemaentwurf bis hin zu Data-Mining, alles in Hinblick auf einen "Anforderungskatalog") eine interessante Alternative zu den einzelnen Übungsaufgaben.
  - Falls zum Vorrechnen angemeldet aber Abgabe zu schlecht vlt. Hinweis an die betroffenen Studenten schreiben, s. d. diese Bescheid wissen und ggf. die Abgabe überarbeiten/sich noch einmal angucken können
  - Generelle Unterstützung von Moodle wäre super

Fragen?