Toets deel 2 Data-analyse en retrieval
Vrijdag 30 Juni 2017: 11.00-13.00

Opgave 1: Naive Bayes voor tekstclassificatie (16 punten)

Gegeven is de volgende collectie filmrecensies met bijbehorende beoordeling:

recensieID	woorden in recensie	Oordeel
r1	`goed script briljante acteurs`	Positief
r2	`mooie beelden geweldige soundtrack`	Positief
r3	`ingenieus script mooie kostuums`	Positief
r4	`matig script slechte dialogen`	Negatief
r5	`slechte acteurs matig camerawerk mooie soundtrack`	Negatief

Het Vocabulaire bestaat uit:

1. goed
2. script
3. briljante
4. acteurs
5. mooie
6. beelden
7. geweldige
8. soundtrack
9. ingenieus
10. kostuums
11. matig
12. slechte
13. dialogen
14. camerawerk

$|V|$ is dus 14.

(a) (8 pnt) Schat de kansen $P(\texttt{mooie}|\text{Negatief})$ en $P(\texttt{mooie}|\text{Positief})$ volgens het multinomiale Naive Bayes model. Gebruik hierbij Laplace smoothing.

ANTWOORD:

$$\hat{P}(\texttt{mooie}|\text{Negatief}) = \frac{1+1}{10+14} = \frac{1}{12}$$

$$\hat{P}(\texttt{mooie}|\text{Positief}) = \frac{2+1}{12+14} = \frac{3}{26}$$

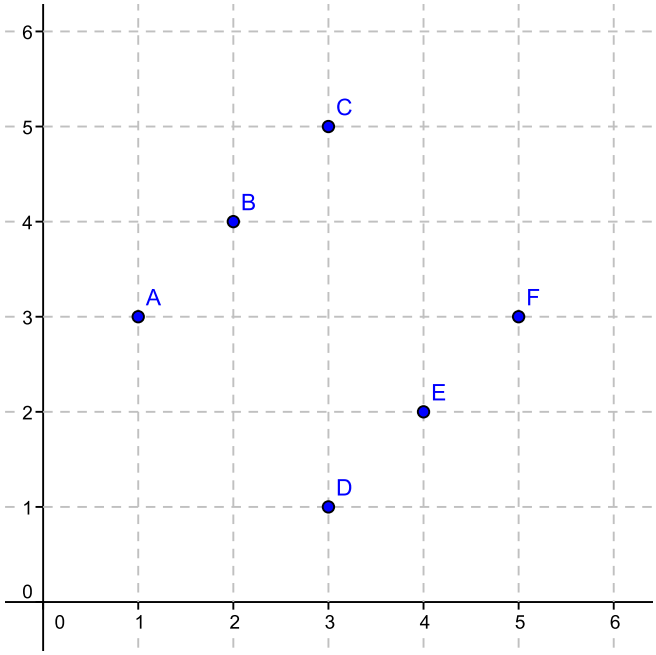
(b) (8 pnt) Schat de kansen $P(\texttt{mooie} = 1|\text{Positief})$ en $P(\texttt{mooie} = 1|\text{Negatief})$ volgens het Bernoulli Naive Bayes model. Gebruik wederom Laplace smoothing.

ANTWOORD:

$$\hat{P}(\texttt{mooie} = 1|\text{Positief}) = \frac{2+1}{3+2} = \frac{3}{5}$$
$$\hat{P}(\texttt{mooie} = 1|\text{Negatief}) = \frac{1+1}{2+2} = \frac{2}{4}$$

Opgave 2: Clustering (24 punten)

(a) (12 pnt) Beschouw de datapunten A, B, C, D, E en F zoals hieronder weergegeven.



Voer het K-means clustering algoritme uit met $K = 2$, en startend met $C_1 = \{A, C, D\}$, en $C_2 = \{B, E, F\}$. Wijs een punt in geval van een "onbeslist" toe aan cluster C_1. Geef tenslotte de RSS van de aldus verkregen clustering.

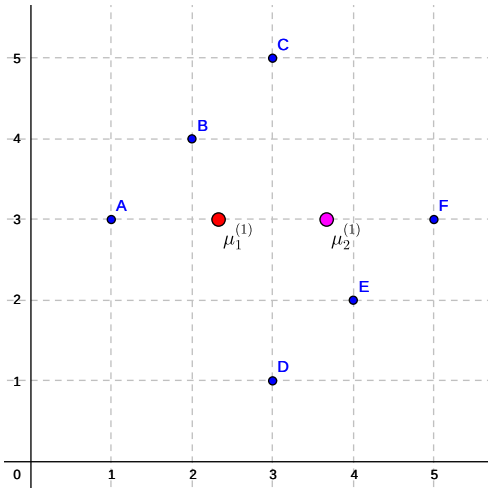
ANTWOORD:
Iteraties van K-means:

Iteratie	C_1	μ_1	C_2	μ_2
1	{A,C,D}	$(2\frac{1}{3}, 3)$	{B,E,F}	$(3\frac{2}{3}, 3)$
2	{A,B,C,D}	$(2\frac{1}{4}, 3\frac{1}{4})$	{E,F}	$(4\frac{1}{2}, 2\frac{1}{2})$
3	{A,B,C}	$(2, 4)$	{D,E,F}	$(4, 2)$
4	{A,B,C}	$(2, 4)$	{D,E,F}	$(4, 2)$

Bijvoorbeeld:

$$\mu_1^{(1)} = \frac{1}{3}\left(\begin{pmatrix}1\\3\end{pmatrix} + \begin{pmatrix}3\\5\end{pmatrix} + \begin{pmatrix}3\\1\end{pmatrix}\right) = \begin{pmatrix}2\frac{1}{3}\\3\end{pmatrix}$$

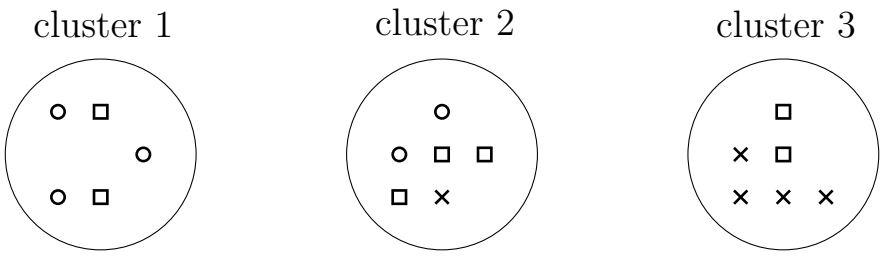
Iteratie 1 geeft de volgende cluster-gemiddelden:



Toekennen aan meest nabije gemiddelde geeft: $C_1^{(2)} = \{A, B, C, D\}$ en $C_2^{(2)} = \{E, F\}$.
Etc.

$$\text{RSS} = 2 + 0 + 2 + 2 + 0 + 2 = 8$$

(b) (12 pnt) Gegeven is de onderstaande clustering van objecten met klasse vierkantje, cirkeltje of kruisje.



Bereken de Rand-Index van deze clustering.

ANTWOORD:

	zelfde cluster	verschillende clusters
zelfde klasse	TP=15	FN
verschillende klasses	FP	TN=70

TP(cluster 1) $= \binom{3}{2} + \binom{2}{2} = 3 + 1 = 4$
TP(cluster 2) $= \binom{3}{2} + \binom{2}{2} = 3 + 1 = 4$
TP(cluster 3) $= \binom{4}{2} + \binom{2}{2} = 6 + 1 = 7$
TP = 4+4+7 = 15
TN(cluster 1, cluster 2) $= 3 \times 3 + 3 \times 1 + 2 \times 2 + 2 \times 1 = 18$
TN(cluster 1, cluster 3) $= 3 \times 4 + 3 \times 2 + 2 \times 4 = 26$
TN(cluster 2, cluster 3) $= 3 \times 4 + 2 \times 4 + 2 \times 2 + 1 \times 2 = 26$
TN = 18+26+26=70
Totaal aantal paren: $\binom{17}{2} = \frac{17 \times 16}{2} = 136$.

$$\text{RI} = \frac{15 + 70}{136} = \frac{5}{8} = 0.625$$

Opgave 3: Gemengde Vragen (18 punten)

(a) (6 pnt) We willen voorspellen welk bedrag personen per jaar aan pizza's uitgeven. We denken dat dit bedrag zowel van leeftijd als van inkomen afhangt, en dat het deel van het inkomen dat aan pizza's wordt uitgegeven kleiner wordt naarmate de leeftijd vordert. Welke predictorvariabelen moeten we in ons regressiemodel opnemen om dit veronderstelde gedrag te kunnen modelleren?

ANTWOORD:

Definieer predictorvariabelen:

- x_1: leeftijd
- x_2: inkomen
- x_3: $x_1 \times x_2$

We hebben dan de regressievergelijking:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

Beetje herschikken geeft:

$$\hat{y} = b_0 + b_1 x_1 + (b_2 + b_3 x_1) x_2$$

We zien dat de invloed van inkomen op pizza-uitgaven nu afhangt van de waarde van leeftijd, zoals gewenst. Als b_3 negatief is zal het deel van het inkomen dat aan pizza's wordt uitgegeven dalen naarmate de leeftijd toeneemt.

(b) (6 pnt) Eén op de tienduizend mensen heeft een bepaalde ziekte en hiervoor bestaat een test die 99% betrouwbaar is. Dit betekent dat de test bij 99% van de personen die aan deze ziekte lijden een positieve uitslag geeft. Andersom geeft de test bij 99% van de personen die niet lijden aan deze ziekte een negatieve uitslag. U test positief bij deze test. Wat is de kans dat u daadwerkelijk de ziekte heeft?

ANTWOORD:
Klassieke toepassing van de regel van Bayes. $P(Z) = 0.0001$, $P(+|Z) = 0.99$, $P(+|\bar{Z}) = 0.01$.

$$P(Z|+) = \frac{P(+|Z)P(Z)}{P(+|Z)P(Z) + P(+|\bar{Z})P(\bar{Z})} = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \approx 0.0098$$

Dus pak 'm beet 1%.

(c) (6 pnt) Beschouw een toevalsexperiment waarin twee zuivere dobbelstenen worden geworpen. De uitkomst van de eerste en tweede worp wordt met X_1 respectievelijk X_2 aangeduid, en de som van de twee uitkomsten noteren we als Y, ofwel $Y = X_1 + X_2$. Stel dat we alleen X_1 waarnemen, en de waarde van Y willen voorspellen. Geef de waarden van β_0 en β_1 in de regressievergelijking

$$E(Y \mid X_1) = \beta_0 + \beta_1 X_1$$

ANTWOORD
Formeel:

$$E(X_1 + X_2 \mid X_1) = E(X_1 \mid X_1) + E(X_2 \mid X_1) = X_1 + E(X_2) = X_1 + 3.5$$

Dus

$$E(Y \mid X_1) = 3.5 + X_1$$

Dus $\beta_0 = 3.5$, en $\beta_1 = 1$.

Opgave 4: Logistische Regressie (30 punten)

We analyseren data van de politie van New York[1], waarin gegevens zijn verzameld van personen die staande zijn gehouden. Soms wordt iemand die staande is gehouden gefouilleerd. Dat is de responsvariabele in ons logistische regressie-model (1 = gefouilleerd, 0 = niet gefouilleerd). Als predictorvariabelen gebruiken we `Sex` en `Race`. Race kan één van de volgende zes waarden aannemen: Black, Black Hispanic, White Hispanic, White, Asian, Native American. In het model wordt de waarde Black als "baseline" gebruikt. Analyse met R geeft de volgende resultaten:

```
   Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.30649    0.04022  -7.620 2.53e-14
SexMale          1.23150    0.04046  30.437  < 2e-16
RaceBlackHisp   -0.05281    0.04377  -1.206   0.2276
RaceWhiteHisp   -0.24887    0.02574  -9.670  < 2e-16
RaceWhite       -0.58176    0.03108 -18.721  < 2e-16
RaceAsian       -0.50935    0.04526 -11.254  < 2e-16
RaceNativeAm    -0.32016    0.15169  -2.111   0.0348
```

Beantwoord de volgende vragen:

[1] Zie: https://www.nyclu.org/en/stop-and-frisk-data

(a) (6 pnt) Wat is de geschatte kans dat een Aziatische man die staande is gehouden wordt gefouilleerd? (rond voor je berekeningen de coëfficiënten af op 2 decimalen)
ANTWOORD:

$$\frac{e^{-0.31+1.23-0.51}}{1+e^{-0.31+1.23-0.51}} = 0.601$$

Dus ongeveer 60%.

(b) (6 pnt) Welke van onderstaande uitspraken zijn in overeenstemming met het model? (er kunnen er 0 of meer goed zijn)

(i) Een man heeft een grotere kans om te worden gefouilleerd dan een vrouw van hetzelfde ras. (ANTWOORD: Juist, want coëfficiënt van `SexMale` is positief.)

(ii) Een zwarte persoon heeft een grotere kans om te worden gefouilleerd dan een witte persoon van hetzelfde geslacht. (ANTWOORD: Juist, `RaceBlack` heeft de facto coëfficiënt 0 (baseline) terwijl `RaceWhite` een negatieve coëfficiënt heeft)

(iii) Een witte persoon heeft een grotere kans om te worden gefouilleerd dan een aziatische persoon van hetzelfde geslacht. (ANTWOORD: Onjuist, de coëfficiënt van `RaceWhite` is negatiever dan van `RaceAsian`)

(iv) Een zwarte vrouw heeft een kleinere kans om gefouilleerd te worden dan een witte man. (ANTWOORD: Juist, $0.58176 - 1.23150 < 0$)

(c) (4 pnt) Is de coëfficiënt van `RaceNativeAm` significant bij significantieniveau $\alpha = 0.05$?
ANTWOORD:
Ja, de p-waarde is 0.0348. Dat is lager dan $\alpha = 0.05$, dus significant.

We passen de classificatieregel toe op de training set zelf, en krijgen dan de onderstaande confusion matrix (rijen: voorspelde klasse, kolommen: werkelijke klasse):

	0	1
0	1,731	1,075
1	13,291	28,514

(d) (8 pnt) Geef de accuracy, recall, precision en F_1 score van het model.

ANTWOORD:

$$\text{Accuracy} = \frac{1,731 + 28,514}{44,611} = 0.678$$
$$\text{Recall} = \frac{28,514}{28,514 + 1,075} = 0.964$$
$$\text{Precision} = \frac{28,514}{28,514 + 13,291} = 0.682$$
$$F_1 = \frac{2 \times 0.964 \times 0.682}{0.964 + 0.682} = 0.799$$

In een alternatief model gebruiken we alleen leeftijd als predictor. Dit levert het volgende resultaat op:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.243807   0.026176   47.52   <2e-16 ***
age         -0.019920   0.000841  -23.68   <2e-16 ***
```

(e) (6 pnt) Bij welke leeftijd is de kans precies 50% dat je wordt gefouilleerd? (rond voor je berekeningen de coëfficiënten af op 2 decimalen)
ANTWOORD:
Bij 62 jaar:

$$\begin{aligned} 1.24 - 0.02x &= 0 \\ 0.02x &= 1.24 \\ x &= \frac{1.24}{0.02} = 62 \end{aligned}$$

Opgave 5: Ordinale Classificatie (12 punten)

We analyseren een verzameling dagboekjes die worden bijgehouden door jonge diabetespatiëntjes. Ten behoeve van een "robotmaatje" voor de kinderen willen we graag de gemoedstoestand van de kinderen afleiden uit hun dagboekjes. De gemoedstoestand wordt gecodeerd als: negatief ($y = 1$), neutraal ($y = 2$), of positief ($y = 3$). Na de nodige voorbewerking van de tekst, en selectie van features vinden we het volgende ordinale logistische regressie model:

coëfficiënt van	schatting
`leuk`	1.63
`niet leuk`	−1.79
`goed`	1.39
`lekker`	1.31
`lol`	0.96
`niet zo`	−0.95
`saai`	−0.78
`gezellig`	0.75
`ziek`	−0.68
`helemaal goed`	0.64

threshold	schatting
t_1	0.25
t_2	0.48

Hierbij zijn de feature-waarden tellingen van woorden (unigrams), of directe opeenvolgingen van 2 woorden (bigrams) in de tekst. Het stukje tekst

```
lekker helemaal goed
```

bevat bijvoorbeeld de unigrams {`lekker`, `helemaal`, `goed`}, en de bigrams {`lekker helemaal`, `helemaal goed`} allemaal één keer.

We roepen in herinnering dat

$$\hat{P}(y \leq j \mid \mathbf{x}) = \Lambda(\hat{t}_j - \hat{\beta}^\top \mathbf{x}), \qquad j \in \{1, 2\},$$

waarbij Λ de cumulatieve logistische kansdichtheidsfunctie is.

(a) (6 pnt) Wat is volgens dit model de kans dat de tekst

```
vandaag helemaal niet leuk
```

een negatieve gemoedstoestand uitdrukt?
Merk op dat alle features de waarde nul hebben behalve `niet leuk` en `leuk` die allebei de waarde 1 hebben.

$$P(y = 1) = P(y \leq 1) = \frac{e^{0.25-1.63+1.79}}{1 + e^{0.25-1.63+1.79}} = 0.601$$

De kans is dus 60.1%.

(b) (6 pnt) We stellen vast dat de geschatte coëfficiënt van `gezellig` positief is. Uit dit enkele feit kunnen we concluderen dat wanneer het aantal voorkomens van `gezellig` stijgt (en er verder niets verandert), dan (kruis de juiste antwoorden, één per rij, in onderstaande tabel aan):

	daalt	stijgt	kan beide
$\hat{P}(y = 1)$			
$\hat{P}(y = 2)$			
$\hat{P}(y = 3)$			

ANTWOORD:
De correct ingevulde tabel is:

	daalt	stijgt	kan beide
$\hat{P}(y = 1)$	X		
$\hat{P}(y = 2)$			X
$\hat{P}(y = 3)$		X	