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

Opgave 1: Naive Bayes voor tekstclassificatie (16 punten)

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)

$$\hat{P}(\texttt{mooie}|\text{Negatief}) = \frac{1+1}{10+14} = \frac{1}{12}$$
$$\hat{P}(\texttt{mooie}|\text{Positief}) = \frac{2+1}{12+14} = \frac{3}{26}$$

(b)

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

Opgave 2: Clustering (24 punten)

(a) 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}	$(4, 2)$

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

(b)

$$\text{RI} = \frac{15+70}{40+96} = \frac{5}{8} = 0.625$$

Opgave 3: Gemengde Vragen (18 punten)

(a) Definieer predictorvariabelen:

- X_1: leeftijd
- X_2: inkomen
- X_3: $X_1 \times X_2$

(b)

$$\frac{99}{10,098} = 0.0098$$

(c)

$$E(Y \mid X_1) = 3\tfrac{1}{2} + X_1$$

Dus $\beta_0 = 3\frac{1}{2}$, en $\beta_1 = 1$.

Opgave 4: Logistische Regressie (30 punten)

(a)
$$\frac{e^{-0.31+1.23-0.51}}{1+e^{-0.31+1.23-0.51}} = 0.601$$
Dus ongeveer 60%.

(b) (i),(ii), en (iv)

(c) Ja, de p-waarde is 0.0348. Dat is lager dan $\alpha = 0.05$, dus significant.

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

(e) 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)

(a) Merk op dat alle features de waarde nul hebben behalve `niet leuk` en `leuk` die allebei de waarde 1 hebben.
$$P(y=1) = \frac{e^{0.25-1.63+1.79}}{1+e^{0.25-1.63+1.79}} = 0.601$$
De kans is dus 60.1%.

(b) De correct ingevulde tabel is:

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