

Toets deel 2 Data-analyse en retrieval

Vrijdag 30 Juni 2017: 11.00-13.00

Algemene aanwijzingen

1. Het is toegestaan een aan beide zijden beschreven A4 met aantekeningen te raadplegen.
2. Het is toegestaan een (grafische) rekenmachine te gebruiken.
3. Geef bij berekeningen niet alleen het eindresultaat, maar laat ook de belangrijke tussenstappen zien.

Opgave 1: Naive Bayes voor tekstclassificatie (16 punten)

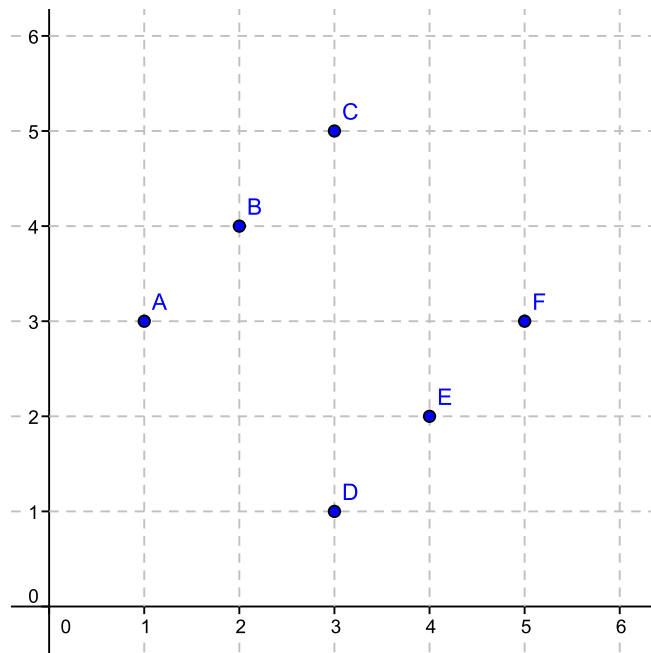
Gegeven is de volgende collectie filmrecensies met bijbehorende beoordeling:

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

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

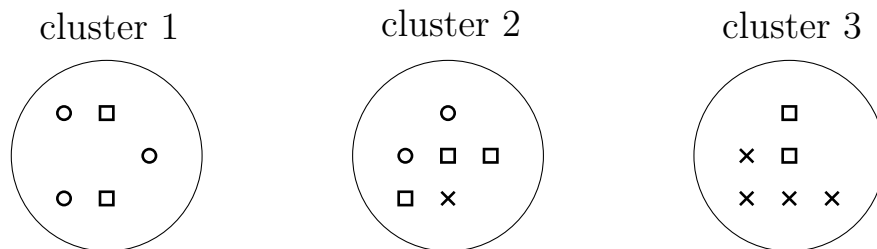
Opgave 2: Clustering (24 punten)

(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.

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



Bereken de Rand-Index van deze clustering.

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?

- (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?
- (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

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

Opgave 4: Logistische Regressie (30 punten)

We analyseren data van de politie van New York¹, 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:

- (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)

¹Zie: <https://www.nyclu.org/en/stop-and-frisk-data>

(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.
- (ii) Een zwarte persoon heeft een grotere kans om te worden gefouilleerd dan een witte persoon van hetzelfde geslacht.
- (iii) Een witte persoon heeft een grotere kans om te worden gefouilleerd dan een aziatische persoon van hetzelfde geslacht.
- (iv) Een zwarte vrouw heeft een kleinere kans om gefouilleerd te worden dan een witte man.

(c) (4 pnt) Is de coëfficiënt van `RaceNativeAm` significant bij significantieniveau $\alpha = 0.05$?

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.

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)

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

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

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?

(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
$\hat{P}(y = 1)$			
$\hat{P}(y = 2)$			
$\hat{P}(y = 3)$			