Toets deel 2 Data-analyse en retrieval
Vrijdag 1 Juli 2016: 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 campingrecensies met bijbehorende beoordeling:

recensieID	woorden in recensie	Oordeel
r1	`goed onderhouden mooie ligging`	Positief
r2	`mooie speeltuin geweldig sanitair`	Positief
r3	`sanitair verwaarloosd slecht schoongemaakt`	Negatief
r4	`speeltuin oud vies slecht onderhouden`	Negatief

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

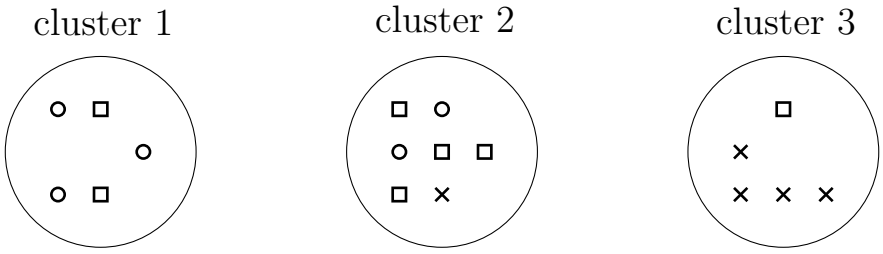
(b) (8 punten) Schat de kansen $P(\texttt{mooie} = 1|\text{Positief})$ en $P(\texttt{mooie} = 1|\text{Negatief})$ volgens het Bernoulli Naive Bayes model. Gebruik wederom Laplace smoothing.

Opgave 2: Clustering (24 punten)

(a) (8 punten) Leg uit hoe je de recall van een document retrieval systeem kunt verbeteren door gebruik te maken van clustering.

(b) (8 punten) We voeren het K-means clustering algoritme twee keer uit op dezelfde dataset, één keer met $K = 2$, en één keer met $K = 3$. We beginnen beide malen vanuit een *willekeurige* toewijzing van punten aan clusters. Produceert de uitvoering van het algoritme met $K = 3$ *altijd* een betere oplossing dan met $K = 2$? Met beter wordt hier bedoeld: een lagere residual sum of squares (RSS). Zo ja, leg uit waarom. Zo nee, geef een voorbeeld waarin dit niet het geval is.

(c) (8 punten) 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 punten) Stel dat we de verkoopprijs (in euro's) van huizen willen voorspellen. Naast de verkoopprijs beschikken we over de volgende gegevens: de perceeloppervlakte (in vierkante meters), en of het huis al dan niet op een aantrekkelijke locatie staat. We willen een model met de volgende eigenschap: huizen op een aantrekkelijke locatie hebben mogelijk een andere prijs per extra vierkante meter perceeloppervlakte dan huizen die niet op een aantrekkelijke locatie staan. Welke predictorvariabelen moeten we in ons regressiemodel opnemen?

(b) (6 punten) Neem aan dat de lengte van volwassen Nederlandse mannen en vrouwen normaal verdeeld is met gemiddelde respectievelijk 182 cm en 168 cm. Verder is gegeven dat de verhouding man-vrouw in de populatie 50-50 is. Iemand selecteert willekeurig een persoon uit de populatie en vertelt mij dat deze persoon een lengte heeft van 175 cm. Als ik de kans op een foute classificatie wil minimaliseren moet ik voorspellen dat deze persoon (kies één van onderstaande opties):

(A) Een man is.

(B) Een vrouw is.

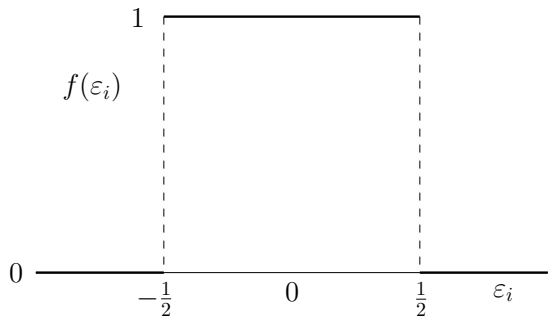
(C) Het is om het even, de kans is op beide even groot.

(D) Er is niet voldoende informatie gegeven om hierover een uitspraak te kunnen doen.

(c) (6 punten) Beschouw het binaire classificatiemodel

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim U\left(-\tfrac{1}{2}, \tfrac{1}{2}\right),$$

ofwel ε_i heeft een uniforme verdeling op het interval $[-\frac{1}{2}, \frac{1}{2}]$. In een plaatje:



Hierbij is y_i^* een latente variabele, we observeren alleen of y_i^* groter is dan nul:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Geef de formule voor $P(y_i = 1 \mid x_i)$ volgens dit model.

Opgave 4: Logistische Regressie (24 punten)

Kunnen we voorspellen welke programma's fouten bevatten? Thomas Zimmermann en collega's (Predicting Defects for Eclipse, Third International Workshop on Predictor Models in Software Engineering, IEEE Computer Society 2007) hebben een onderzoek hiernaar uitgevoerd op de code base van de Eclipse programmeeromgeving (een van de grootste open-source projecten). We analyseren data van Eclipse 3.0 packages; dit zijn er 661 in totaal. We proberen te voorspellen of van een package één of meer defecten zijn gerapporteerd binnen 6 maanden na de release, of dat er geen enkel defect is gerapporteerd. Deze mogelijkheden worden als respectievelijk `defect=1` en `defect=0` gecodeerd. We modelleren dit probleem met logistische regressie. We gebruiken een eenvoudig model met als enige predictorvariabele het totaal aantal regels code in het package gedeeld door 100. Deze predictor wordt aangeduid met `TLOC` (Total Lines Of Code). We schatten het model met maximum likelihood.

Dit levert het volgende resultaat op (zie het extract van de R output hieronder):

```
Coefficients:
                Estimate   Std. Error
(Intercept)    -1.050383   0.123470
TLOC            0.064049   0.007252
```

(a) (6 punten) Wat is de geschatte kans op een fout in een package met 1000 regels code? (Let op de definitie van `TLOC`!)

(b) (4 punten) Welke van onderstaande uitspraken is juist? (kies één antwoord)

(A) Als het aantal regels code met 100 toeneemt, dan neemt de kans dat er een defect wordt gerapporteerd met ongeveer 6.4 procentpunten toe.

(B) Als het aantal regels code met 100 toeneemt, dan neemt de kans dat er een defect wordt gerapporteerd toe, maar de grootte van de toename hangt af van de uitgangswaarde van `TLOC`.

(C) Als het aantal regels code met 100 toeneemt, dan kan de kans dat er een defect wordt gerapporteerd zowel toenemen als afnemen; dit komt doordat de intercept negatief is.

(D) Alle bovenstaande uitspraken zijn onjuist.

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

(d) (4 punten) Geef een eenvoudige classificatieregel om te voorspellen of van een package al dan niet een defect gerapporteerd zal worden. Ga er hierbij van uit dat je de klasse met de grootste kans gegeven het aantal regels code voorspelt.

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	290	154
1	58	159

(e) (6 punten) Geef de accuracy, recall, precision en F_1 score van het model.

Opgave 5: Ordinale Classificatie (18 punten)

Gegeven is een dataset met 2933 query-document paren en bijbehorende relevantiebeoordelingen. De relevantiebeoordeling kan zijn: laag (code: 1), middel (code: 2), of hoog (code: 3). De predictorvariabelen zijn Term Frequency (`TF`) (het aantal keren dat een query-term voorkomt in het document) en Document Length (het aantal woorden in het document) (`DL`). Beide variabelen zijn geschaald tussen 0 en 1, dat wil zeggen, hun waarden liggen in het interval $[0, 1]$. We passen het proportional odds logistische regressiemodel toe en vinden de volgende resultaten:

Coëfficiënt	Schatting
`TF`	0.7643
`DL`	−0.4590

Threshold	Schatting
t_1	1.3776
t_2	2.7270

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 punten) Wat is volgens dit model de kans dat een query-document paar de hoogste relevantie-klasse heeft wanneer beide predictorvariabelen de waarde 0.5 hebben?

(b) (6 punten) We stellen vast dat de geschatte coëfficiënt van `TF` positief is. Uit dit enkele feit kunnen we concluderen dat wanneer Term Frequency stijgt bij gelijkblijvende Document Length, dan (kruis de juiste antwoorden, één per rij, in onderstaande tabel aan):

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

(c) (6 punten) De verdeling van de klassen in de trainingset is als volgt:

Klasse	1	2	3
Aantal	2316	427	190

Hoeveel gevallen worden door het proportional odds logistische regressiemodel fout geclassificeerd op de training set? Neem hierbij aan dat we aan de klasse met de grootste kans gegeven de waarden van `TF` en `DL` toewijzen.

Kies één van onderstaande opties:

(A) 2316

(B) 190

(C) 617

(D) Er is onvoldoende informatie gegeven om hierover een uitspraak te kunnen doen.