Herkansing Toets deel 2 Data-analyse en retrieval
Vrijdag 12 Juli 2019: 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 (24 punten)

Gegeven is de volgende collectie artikelen met bijbehorend onderwerp (Economie of Sport):

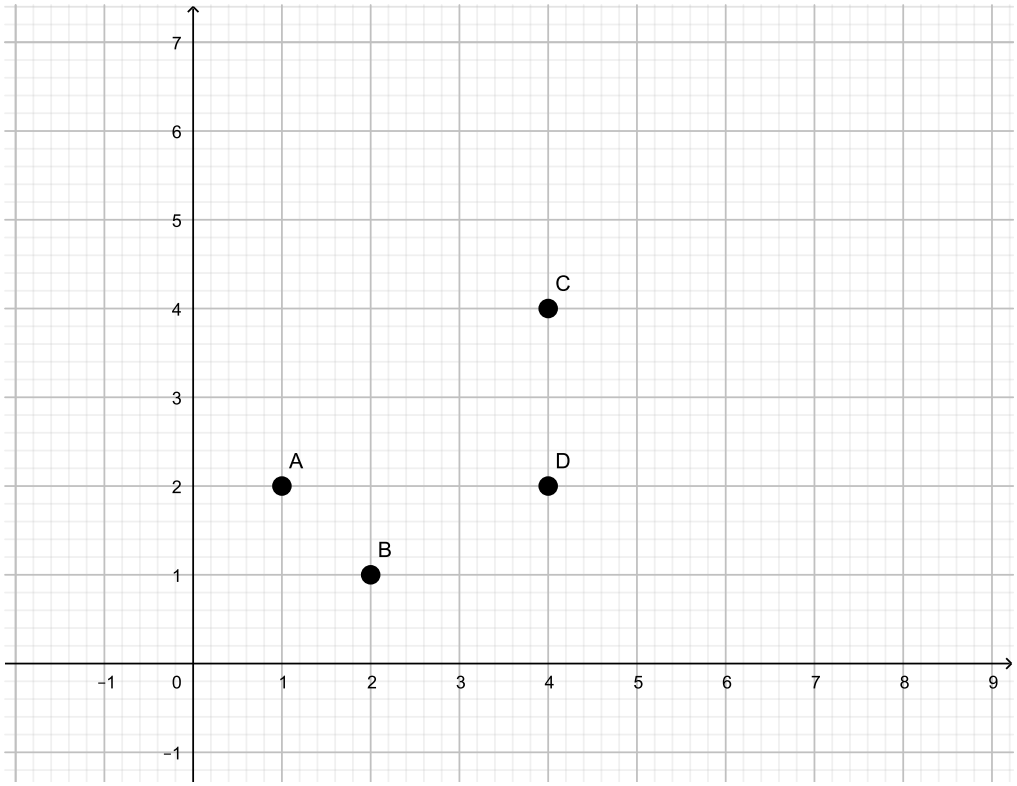
artikelID	woorden in artikel	Onderwerp
a1	`rente winst investeringen`	Economie
a2	`aandeel dividend winst`	Economie
a3	`doelpunt winst wedstrijd`	Sport
a4	`wedstrijd verlies speler bal`	Sport
a5	`winst verlies`	?

(a) (14 pnt) Gebruik a1 t/m a4 om de kansen te schatten die nodig zijn voor het classificeren van a5 volgens het multinomiale Naive Bayes model. Gebruik hierbij Laplace smoothing.

(b) (10 pnt) Bereken $P(\text{Economie} \mid \text{a5})$ en $P(\text{Sport} \mid \text{a5})$ op basis van het multinomiale Naive Bayes model. Aan welke klasse wordt a5 toegewezen?

Opgave 2: Clustering (20 punten)

(a) (10 pnt) Beschouw de datapunten A, B, C, en D zoals hieronder weergegeven:



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

(b) (10 pnt) Stel we hebben 12 observaties, waarvan 4 van klasse A, 6 van klasse B en 2 van klasse C. Het klasse-label wordt gebruikt als "gouden standaard" om de kwaliteit van een clustering te beoordelen.

Wat is de Rand-Index als we 12 clusters maken met ieder slechts één observatie?

Opgave 3: Gemengde Vragen (16 punten)

(a) (6 pnt) 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: de verkoopprijs hangt af van de perceeloppervlakte, en voor een huis dat op een aantrekkelijke locatie staat is de prijs per vierkante meter perceeloppervlakte groter dan voor een huis dat niet op een aantrekkelijke locatie staat. Welke predictorvariabelen moeten we in ons regressiemodel opnemen om deze eigenschap te kunnen modelleren?

(b) (6 pnt) In het college hebben we het principe van maximum likelihood schatten behandeld. In het algemeen maximaliseren we de likelihoodfunctie:

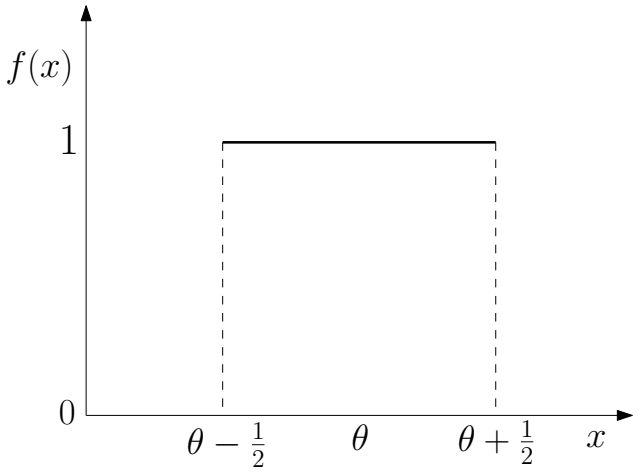
$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

naar θ, waarbij θ de onbekende parameter is die we willen schatten, en f de kans(dichtheids)-functie van x is. Er zijn in totaal n waarnemingen, en x_i is de i-de waarneming van x.

Laat nu x een toevalsvariabele zijn met kansdichtheidsfunctie:

$$f(x; \theta) = \begin{cases} 1 & \text{als } \theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2} \\ 0 & \text{anders} \end{cases}$$

Ofwel, x heeft een uniforme verdeling op het interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, met $\theta \in \mathbb{R}$. In een plaatje:



De parameter θ is onbekend, en die willen we graag schatten uit een steekproef $x_1, x_2, \ldots, x_n$ ter grootte n.

Beargumenteer dat *iedere* waarde in het interval

$$[\max(x_1, \ldots, x_n) - \tfrac{1}{2}, \min(x_1, \ldots, x_n) + \tfrac{1}{2}]$$

een maximum likelihood schatter is van θ. Je hoeft geen formeel bewijs te geven, een informeel argument volstaat.

(c) (4 pnt) In een ordinaal logistisch regressie-model geeft de responsvariabele aan hoe relevant een document is voor een query op een schaal van 1 (volstrekt irrelevant) tot 5 (hoogst relevant). De feature x_1 is gedefinieerd als "het aantal querytermen dat in het document voorkomt", en heeft een *positieve* coëfficiënt in het geschatte model. Uit het enkele feit dat de coëfficiënt van x_1 positief is, volgt dat (volgens het model) als de waarde van x_1 toeneemt, en de andere feature-waarden gelijk blijven, (kies het beste antwoord)

1. de kans op klasse 1 afneemt.
2. de kans op klasse 5 afneemt.
3. de kans op alle klassen afneemt.
4. de kans op alle klassen toeneemt.
5. de kans op de lagere klassen (1 en 2) afneemt, en de kans op de hogere klassen (4 en 5) toeneemt.

Opgave 4: Logistische Regressie (24 punten)

We analyseren gegevens van professionele dartswedstrijden om te voorspellen welke speler een wedstrijd gaat winnen. Alleen wedstrijden van het type "best of x legs" zijn meegenomen. In het model hangt de kans dat de beginnende speler a wint van speler b af van het verschil in gemiddelde en checkoutpercentage tussen de twee spelers, en een constante β_0:

$$P(a \text{ wint van } b) = \Lambda(\beta_0 + \beta_1(\text{Gem}_a - \text{Gem}_b) + \beta_2(\text{Check}_a - \text{Check}_b))$$

Hierbij is Λ de logistische responsfunctie, Gem_x het gemiddelde van speler x, Check_x het checkoutpercentage van speler x, en a de speler die mag beginnen met werpen. Schatten met de methode van maximum likelihood geeft de volgende resultaten:

Coëfficiënt	Schatting
β_0 (Intercept)	0.120
β_1	0.135
β_2	0.025

Beantwoord de volgende vragen:

(a) (6 pnt) Hoe groot is het voordeel van het recht om te mogen beginnen met werpen volgens dit model?

(b) (6 pnt) Michael van Gerwen heeft een gemiddelde respectievelijk checkout-percentage van 102.7 en 46.2%. Vincent van de Voort heeft een gemiddelde respectievelijk checkout-percentage van 92.6 en 40.4%. Wat is volgens het model de kans dat Michael van Gerwen wint van Vincent van de Voort als van Gerwen mag beginnen?

Op basis van de quoteringen op goksites is het mogelijk te berekenen wat volgens de gokmarkt de winstkansen van de spelers zijn. We zetten de correctheid van de voorspellingen (1 = correct; 0 = incorrect) van het model op de testset uit tegen die van de gokmarkt (rijen: model, kolommen: gokmarkt):

	0	1
0	62	26
1	20	156

(c) (6 pnt) Bereken de accuracy van het model respectievelijk de gokmarkt op de testset. Verslaat het model de gokmarkt?

(d) (6 pnt) We willen de nulhypothese toetsen dat het model en de markt dezelfde accuracy hebben. De alternatieve hypothese is dat ze niet dezelfde accuracy hebben. Volgens de in het college behandelde toets heeft de toetsgrootheid een binomiaalverdeling. Geef de parameterwaarden van deze verdeling onder de nulhypothese voor de gegeven data. Geef tevens een expressie voor de p-waarde van deze toets voor de gegeven data. De p-waarde zelf hoef je niet uit te rekenen.

Opgave 5: Leren van Word Embeddings met Gradient Descent (16 punten)

We beschouwen het probleem van het leren van word embeddings uit een tekstcorpus. Voor centrumwoord i en contextwoord j trachten we de volgende error te minimaliseren:

$$E(u_i, v_j) = \frac{1}{2}(u_i^\top v_j - \log_2 X_{ij})^2.$$

Hierbij is u_i de vector van centrumwoord i, en v_j de vector van contextwoord j.
X_{ij} is het aantal keren dat woord j in de context van woord i voorkomt.

Gegeven is dat $X_{ij} = 128$, en dat

$$u_i^{(0)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad v_j^{(0)} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

de startwaarden zijn voor de woordvectoren. Verder is gegeven dat stapgrootte $\eta = \frac{1}{10}$. We doen een update van de waarden van u_i en v_j.

(a) (10 pnt) Bereken $u_i^{(1)}$ en $v_j^{(1)}$ met behulp van het gradient descent algoritme.

(b) (6 pnt) Is de error inderdaad gedaald na de update? Laat je berekening zien.