Toets deel 2 Data-analyse en retrieval
Vrijdag 29 Juni 2018: 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 informatica vakevaluaties met bijbehorende beoordeling:

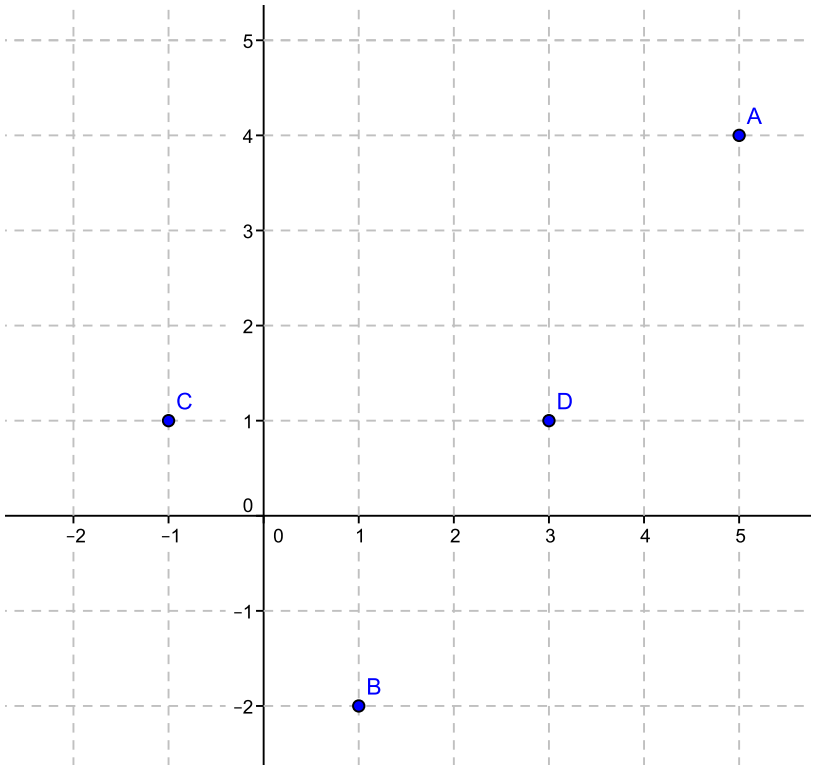
evaluatieID	woorden in evaluatie	Oordeel
e1	`goede docent interessante colleges`	Positief
e2	`goede colleges goede begeleiding`	Positief
e3	`slechte docent onmiddelijk afschaffen`	Negatief
e4	`saaie colleges docent kleunbaviaan`	Negatief

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

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

Opgave 2: Clustering (28 punten)

(a) (16 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 $C_1 = \{A, B\}$, en $C_2 = \{C, D\}$. 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) (12 pnt) We hebben observaties van drie klassen in de aantallen 4, 8, en 2.
Het klasse-label wordt gebruikt als "gouden standaard" om de kwaliteit van een clustering te beoordelen.

Wat is de Rand-Index als we 14 clusters maken met ieder slechts één observatie?

Opgave 3: Gemengde Vragen (18 punten)

(a) (6 pnt) We willen het energieverbruik (electriciteit, gas) voorspellen op basis van temperatuur. We denken dat het verbruik relatief hoog is bij lage temperaturen (omdat we de verwarming moeten aanzetten) en bij hoge temperaturen (omdat we de airco aanzetten), en dat het energieverbruik bij gemiddelde temperaturen relatief laag is. Welke predictorvariabelen moeten we in ons regressiemodel opnemen om dit veronderstelde gedrag te kunnen modelleren?

(b) (8 pnt) In een verzameling song lyrics behoort 90% tot het Metal genre en 10% tot het Rap genre. In 50% van de Rap lyrics komt het woord "bitch" ten minste één keer voor. Voor de Metal lyrics is dit slechts 5%. We trekken willekeurig een liedje uit de verzameling en stellen vast dat dit het woord "bitch" bevat. Wat is de kans dat we een Rap liedje getrokken hebben?

(c) (4 pnt) In een lineair regressiemodel gebruiken we inkomen als enige predictorvariabele om te voorspellen welk bedrag iemand per jaar aan pizza's uitgeeft. Bij het schatten van model A gebruiken we data waarbij inkomen is uitgedrukt in euro's, en bij het schatten van model B gebruiken we dezelfde data, maar met inkomen uitgedrukt in duizenden euro's (er wordt niet afgerond). We beschouwen de geschatte coëfficiënten van inkomen in model A en model B.

Welke uitspraak is juist?

1. De coëfficiënten zijn even groot.
2. De coëfficiënt van model A is duizend keer zo groot als de coëfficiënt van model B.
3. De coëfficiënt van model B is duizend keer zo groot als de coëfficiënt van model A.
4. De verhouding tussen de coëfficiënten hangt af van de variantie van inkomen.

Opgave 4: Logistische Regressie (24 punten)

We analyseren data van de politie van New York, 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 `Stadsdeel` en `Leeftijd` (in jaren). Stadsdeel kan één van de volgende vijf waarden aannemen: Manhattan, Brooklyn, Bronx, Queens, Staten Island. In het model wordt de waarde Manhattan als "baseline" gebruikt. Analyse met R geeft de volgende resultaten:

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.1213601  0.0367730  30.494  < 2e-16
StadsdeelBrooklyn  0.1654819  0.0308809   5.359 8.38e-08
StadsdeelBronx     0.3914823  0.0369222  10.603  < 2e-16
StadsdeelQueens    0.0929490  0.0308089   3.017  0.00255
StadsdeelStatIsl  -0.1822510  0.0382430  -4.766 1.88e-06
Leeftijd          -0.0194676  0.0008508 -22.883  < 2e-16
```

Beantwoord de volgende vragen:

(a) (6 pnt) Wat is de geschatte kans dat een persoon van 25 jaar die in de Bronx staande is gehouden wordt gefouilleerd? (rond voor je berekeningen de coëfficiënten af op 2 decimalen)

(b) (6 pnt) Bij welke leeftijd is de kans precies 50% dat je wordt gefouilleerd als je in stadsdeel Staten Island staande wordt gehouden? (rond voor je berekeningen de coëfficiënten af op 2 decimalen)

(c) (4 pnt) Is de coëfficiënt van `StadsdeelQueens` significant bij $\alpha = 0.01$? Leg uit.

We gebruiken het model om de training set te classificeren, en krijgen dan de onderstaande confusion matrix (rijen: voorspelde klasse, kolommen: werkelijke klasse):

	0	1
0	342	315
1	15,058	29,965

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

Opgave 5: Leren van Word Embeddings met Gradient Descent (14 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:

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

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 $X_{ij} = 64$, en dat

$$u_i^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad v_j^{(0)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

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) (8 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.