Toets deel 2 Data-analyse en retrieval
Dinsdag 2 Juli 2013: 8.30-10.30

Algemene aanwijzingen

1. Het is toegestaan een rekenmachine te gebruiken.
2. Geef bij berekeningen niet alleen het eindresultaat, maar laat ook de belangrijke tussenstappen zien.
3. Er zijn in totaal 5 opgaven.
4. Onderaan pagina 4 is een aantal formules gegeven.

Opgave 1: Naive Bayes voor tekstclassificatie

Gegeven is de volgende collectie recepten met bijbehorende keukens:

receptID	woorden in recept	Keuken
r1	pasta tomaat gehakt	Italiaans
r2	pasta gorgonzola aubergine courgette	Italiaans
r3	rijst olijfolie oesterzwam	Italiaans
r4	feta tomaat olijfolie oregano	Grieks
r5	yoghurt komkommer knoflook	Grieks

(a) Schat de kansen $P(\text{pasta}|\text{Italiaans})$ en $P(\text{yoghurt}|\text{Grieks})$ volgens het *multinomiale* Naive Bayes model. Gebruik hierbij Laplace smoothing.

(b) Schat de kansen $P(\text{pasta}|\text{Italiaans})$ en $P(\text{yoghurt}|\text{Grieks})$ volgens het *Bernoulli* Naive Bayes model. Hierbij is $P(\text{pasta}|\text{Italiaans})$ kort voor $P(\text{pasta} = 1|\text{Italiaans})$. Gebruik wederom Laplace smoothing.

Opgave 2: Clustering

(a) Beschrijf hoe je de recall van een document retrieval systeem zou kunnen verbeteren met behulp van clustering.

(b) Geef een korte beschrijving van een andere toepassing van clustering in information retrieval of web search.

(c) Leg uit waarom je K-means clustering goed zou kunnen omschrijven als 'Rocchio zonder klasselabels'.

Opgave 3: Ranking

We hebben een collectie van 4 documenten, en de ranking van die 4 documenten voor 2 verschillende queries. De gegeven rankings zijn bepaald door een expert. Zoals te doen gebruikelijk staat het meest relevante document bovenaan, etc.

Query 1		
docID	x_1	x_2
101	5	2
102	4	2
103	4	3
104	1	6

Query 2		
docID	x_1	x_2
102	4	3
103	3	3
101	3	5
104	2	3

Voor ieder document-query paar hebben we de waarden van x_1 en x_2 berekend. Hierbij is x_1 de cosine similarity tussen het document en de query ($\times 100$) en x_2 de minimale window-breedte waarin alle query-termen voorkomen in het document. Bijvoorbeeld: de cosine similarity tussen document 101 en query 1 is 0.05, en de minimale window-breedte waarin alle query-termen van query 2 in document 104 voorkomen is 3. We beschouwen de klasse van lineaire ranking functies

$$f(x_1, x_2) = w_1 x_1 + w_2 x_2$$

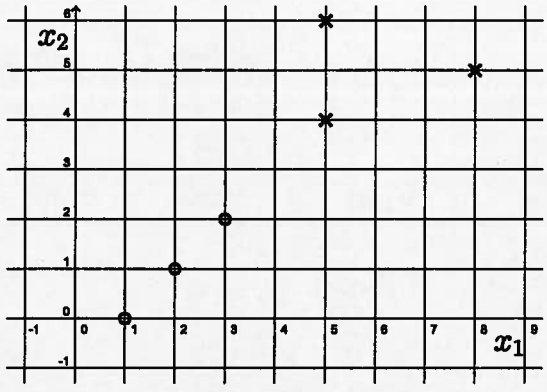
(a) Rank de 4 documenten met behulp van de lineaire ranking functie met $w_1 = 1$ en $w_2 = -1$ voor query 1. Doe hetzelfde voor query 2.

(b) Bereken de waarde van Kendall's τ tussen de onder (a) verkregen ranking en de echte ranking, zowel voor query 1 als voor query 2.

(c) Kun je de onder (a) gegeven ranking functie verbeteren? Zo ja, geef de verbeterde gewichten.

Opgave 4: Support Vector Machines

Beschouw de zes datapunten die in onderstaande grafiek zijn weergegeven:



De punten met klasselabel $y = -1$ zijn weergegeven door cirkels, en de punten met klasselabel $y = +1$ zijn weergegeven door kruisjes.

(a) Geef de vergelijking van de lineaire decision boundary met perfecte separatie en maximale margin.

Let op: je hoeft de coëfficiënten *niet* te schalen zodat $y(w_1x_1 + w_2x_2 + b) = 1$ voor de support vector(s).

(b) Geef de bijbehorende support vector(s).

(c) Geef de vergelijking van de Rocchio decision boundary. Leg uit hoe je aan het antwoord komt.

Opgave 5: Mutual Information

Mutual information wordt gebruikt om de mate van samenhang tussen twee variabelen te kwantificeren. De formele definitie is onderaan de volgende bladzijde gegeven.

(a) Waar wordt mutual information *specifiek* voor gebruikt bij het bouwen van classifiers? Leg uit hoe dit in zijn werk gaat.

Het is altijd zo dat $I(X;Y) \geq 0$ (niet geheel triviaal).

(b) Laat zien dat als X en Y onafhankelijke variabelen zijn dan geldt $I(X;Y) = 0$.

Gegeven is de volgende groep personen (zie achterzijde), en de binaire variabelen:

`geslacht, draagt-hoofddeksel, draagt-bril, heeft-snor, heeft-baard.`

(c) Bereken de mutual information tussen `geslacht` en `draagt-hoofddeksel`.

(d) Is er een variabele (uit het gegeven lijstje variabelen) die een sterkere samenhang heeft met `geslacht`? Zo ja, welke?

Formules

De mutual information $I(X;Y)$ tussen twee discrete variabelen X (met waarden x) en Y (met waarden y) is gedefinieerd als

$$I(X;Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Bij het schatten van de mutual information uit data vervangen we in deze formule de theoretische kansen door uit de data geschatte kansen. We sommeren alleen over die combinaties van waarden x, y die in de data voorkomen, d.w.z. waarvoor geldt dat $\hat{P}(x,y) > 0$. Als je rekenmachine geen log met basis 2 heeft mag je ook de natuurlijke logaritme (ln) gebruiken. Wel even duidelijk aangeven natuurlijk.