Toets Parel 100 — Intelligente Interactie

3 oktober 2014

Dit tentamen bestaat uit 5 vragen. Het cijfer voor deze toets is het aantal punten gedeeld door 100.

☞ **1** *(15 punten)* Voor het vroegtijdig detecteren van de ziekte Z wordt een test T gedaan bij een populatie van vogels. De uitkomst van de test kan positief zijn: $T = pos$, een indicatie dat de vogel de ziekte Z heeft, of negatief: $T = neg$, er is geen indicatie dat de vogel de ziekte Z heeft. Het volgende is bekend over het voorkomen van de ziekte onder vogels en de betrouwbaarheid van de test T.

- $P(Z = true) = 0.001$, de kans dat een vogel de ziekte Z heeft.
- $P(T = pos|Z = true) = 0.9$, de kans dat een vogel die de ziekte Z heeft positief op de test reageert.
- $P(T = pos|Z = false) = 0.01$, de kans dat een vogel die ziekte Z *niet* heeft toch positief op de test reageert.

Bereken de kans $P(Z = true|T = pos)$, de kans dat een vogel de ziekte Z heeft als de uitkomst van de test positief is. (Hint: gebruik Bayes' rule. Gebruik: de wet van de totale kans om $P(T = pos)$ te berekenen.)

☞ **2** *(15 punten)* In urn H_1 zitten **20** knikkers: 4 rode, 8 witte en 8 blauwe. In urn H_2 zitten ook **20** knikkers, maar nu: 10 rode, 5 witte en 5 blauwe. Iemand kiest *willekeurig* een urn en trekt 5 keer (*met teruglegging na iedere trekking*) een knikker uit de gekozen urn. De uitkomst D van de trekking is: 2 rode, 1 witte en 2 blauwe. Welke is de meest waarschijnlijke urn, ofwel: welke kans is groter $P(H_1|D =< 2,1,2 >)$ of $P(H_2|D =< 2,1,2 >)$? Motiveer je antwoord en geef de berekeningen die je hebt gedaan ter ondersteuning van je antwoord. (Let op: er wordt alleen gevraagd welke *de meest waarschijnlijke* urn is waaruit getrokken is gegeven de uitkomst van de trekking.)

☞ **3** *(25 punten)* Deze vraag gaat over een aangepast versie van het brouwerijvoorbeeld zoals beschreven in de Machine Learning reader. Stel de brouwerij heeft nu de volgende dataset waarbij de laatste kolom de assessment voorstelt, dus de classificatie van een locatie.

Ex.	U	H	I	T	S	A
1	Y	M	N	P	M	N
2	N	S	N	P	L	N
3	Y	M	N	A	M	Y
4	N	M	N	P	S	N
5	N	M	Y	P	M	Y
6	Y	N	N	A	S	N
7	N	N	N	A	S	Y
8	N	S	N	A	M	Y
9	N	L	Y	P	L	Y
10	N	M	N	P	S	N

(a) Wat is de information gain van feature U?
(b) Stel dat U de feature is die de hoogste information gain heeft. Dus dan komt U in top van de beslisboom. Wat is de dataset die gebruikt wordt voor het bepalen van het vervolg van de beslisboom onder de tak $U = Y$?
(c) Wat is de information gain voor de feature T voor de data behorende bij de tak $U = Y$?

☞ **4** *(25 punten)* Beschouw de volgende confusion matrix behorende bij een classifier:

		Predicted class		
		C_1	C_2	C_3
Actual Class	C_1	120	35	12
	C_2	10	140	20
	C_3	13	5	130

(a) Wat is de accuracy van deze classifier?
(b) Wat is recall voor klasse C_2?
(c) Wat is de precision voor klasse C_3?

☞ **5** *(20 punten)* Gegeven de classificatielijn $4 + 2x_1 - 3x_2 = 0$ en het featurepunt $x = (-2, 2)$.

(a) Hoe wordt aan de hand van deze classificatielijn het punt x geclassificeerd (1 of 0)?
(b) Stel dat het punt x fout geclassificeerd wordt. Hoe wordt de gewichtenvector $w = (4, 2, -3)$ aangepast vanwege deze misclassificatie? Neem een learning rate van 0.5.