

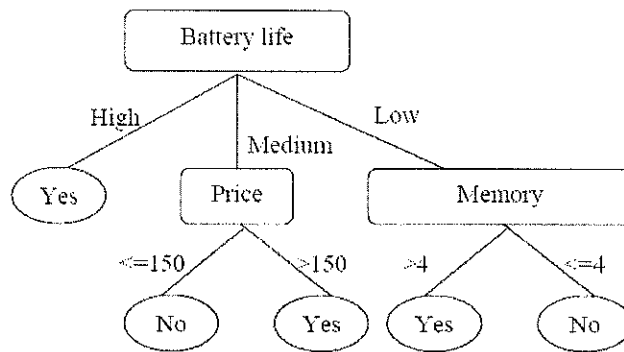
Tentamen

“Lerende en Redenerende Systemen” / “Data mining”

22 april 2010

- Schrijf op elk blad je naam en studentnummer.
- Bij elke opgave staat hoeveel punten er te verdienen zijn (100 totaal).
- Succes!

1. Gegeven onderstaande beslisboom die bedoeld is om de klanttevredenheid van batterijen te modelleren.



(a) [3] De beslisboom impliceert een set van regels. Welk van deze regels komt niet overeen met wat je zou verwachten?

De beslisboom is getraind op een trainingset (hier niet getoond). Daarnaast hebben we een validatieset als in de volgende tabel (eerste vier kolommen).

battery life	price	memory	satisfaction	P(satisfaction = yes)
low	80	3	no ✓	0.42
high	300	6	no -	0.18
medium	200	4	yes ✓	0.82
medium	350	3	yes ✓	0.37
low	70	4	no ✓	0.51
high	400	7	yes ✓	0.92
high	200	6	no -	0.61
medium	100	8	yes -	0.49
medium	280	3	yes ✓	0.78
low	70	5	yes ✓	0.81

Handwritten notes next to the table:

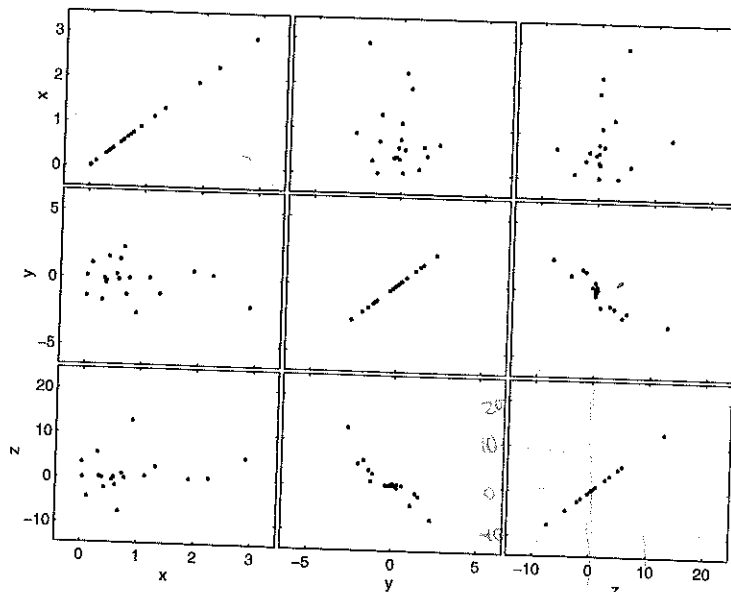
- 0.92 yes
- 0.82 yes
- 0.81 yes
- 0.78 yes
- 0.61 no
- 0.51 no
- 0.49 yes
- 0.42 no
- 0.37 yes
- 0.18 no

- (b) [3] Wat is de (geschatte) generalisatiefout op grond van deze validatieset?
- (c) [6] Noem minimaal twee nadelen voor het gebruik van een enkele validatieset. Hoe kun je deze nadelen ondervangen?
- (d) [10] Daarnaast hebben we een andere, probabilistische classifier getraind. Deze voorspelt niet slechts de klasse, maar geeft de kans op een klasse aan. Deze staat in de vijfde kolom van bovenstaande tabel. Geef de ROC (receiver operating characteristic) curve weer voor deze classifier, waarbij je de "false positive rate" (aantal false positives gedeeld door het aantal false positives + true negatives) langs de x-as uitzet tegen de "true positive rate" (aantal true positives gedeeld door het aantal true positives + false negatives) langs de y-as. Wat valt het meest op aan deze ROC curve?

We gaan nu de twee classifiers met elkaar vergelijken. Hiertoe "dwingen" we de probabilistische classifier een uitspraak te doen door deze steeds te laten kiezen voor de meest waarschijnlijke klasse.

- (e) [2] Welke classifier doet het beter op de validatieset?
- (f) [8] Beschrijf een algemene procedure om te bepalen of het verschil tussen twee classifiers die uitspraken doen over *dezelfde* validatieset significant genoemd kan worden en pas deze toe op dit specifieke geval.
- (g) [3] Is het verschil significant, uitgaande van een confidence level van 0.05? Je hoeft dit niet persé uit te rekenen, mag het ook beargumenteren.

2. Gegeven een dataset van 20 objecten en 3 attributen (variabelen), x , y en z . De scatterplot staat weergegeven in onderstaande figuur.



$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

- (a) [6] Geef van elk paar variabelen aan of ze positief, negatief of (zo goed als) ongecorrleerd zijn.
- (b) [8] Schets het histogram van variabele z . Wat valt vooral op?
- (c) [8] Schets een boxplot voor variabele x . Wat kun je hier goed zien?
- (d) [6] Het vervelende van een scatterplot is dat je niet direct verbanden kunt zien tussen meer dan 2 variabelen. Met welke methode kan dit wel? Noem niet alleen de methode, maar leg ook uit hoe je dan deze verbanden tussen de variabelen kunt zien, eventueel aan de hand van een voorbeeld.
- (e) [6] In de boxplot worden o.a. kwantielen (waaronder de mediaan) weergegeven. Het voordeel van de mediaan t.o.v. het gemiddelde als samenvatting van de data is dat deze "robuust" is, dat wil zeggen, minder gevoelig voor uitbijters. Wat is een robuuste maat voor de spreiding van de data? Geef ook de bijbehorende formule.

3. Veel data mining algoritmen maken gebruik van de afstanden tussen objecten i.p.v. de attribuutwaarden van de objecten zelf.

- (a) [3] De "som van de kwadraten",

$$d_{\text{kwadr}}(\mathbf{x}, \mathbf{y}) = \sum_i (x_i - y_i)^2,$$

wordt soms gebruikt als afstandsmaat. Volgens de driehoeksongelijkheid is de directe weg van een punt \mathbf{x} naar een punt \mathbf{y} nooit langer dan een indirecte weg via een willekeurig punt \mathbf{z} . De som van de kwadraten voldoet *niet* aan de driehoeksongelijkheid. Laat dit zien aan de hand van een tegenvoorbeeld.

- (b) [3] De Euclidische afstand,

$$d_{\text{Euclid}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2},$$

voldoet wel aan de driehoeksongelijkheid. Laat zien dat je onder a) bedachte tegenvoorbeeld voor d_{kwadr} nu voor d_{Euclid} geen tegenvoorbeeld meer is.

- (c) [6] Met wat rekenwerk kun je laten zien dat $d_{\text{Euclid}}(\mathbf{x}, \mathbf{z}) + d_{\text{Euclid}}(\mathbf{z}, \mathbf{y})$ minimaal is voor een punt \mathbf{z} op de lijn tussen \mathbf{x} en \mathbf{y} , d.w.z., voor een punt \mathbf{z} dat voldoet aan $\mathbf{z} = \alpha \mathbf{x} + (1 - \alpha) \mathbf{y}$ met $0 \leq \alpha \leq 1$. [Je kunt 8 bonuspunten verdienen door zelf te bewijzen dat dit zo is. Let wel: dit is waarschijnlijk niet eenvoudig, dus doe dit alleen als je echt tijd over hebt!] Gebruik dit om aan te tonen dat de Euclidische afstand inderdaad voldoet aan de driehoeksongelijkheid.

Een afstandsmaat die voldoet aan (onder andere) de driehoeksongelijkheid heeft als voordeel dat deze overeenkomt met onze intuïtie. Ook kan het helpen data mining algoritmen efficiënter te maken, zoals blijkt uit het volgende voorbeeld.

- (d) [10] Stel we weten de afstanden van alle objecten in onze dataset tot een punt \mathbf{x} , kennen de afstand tussen \mathbf{x} en een ander punt \mathbf{y} in de buurt van \mathbf{x} , maar (nog) niet de afstanden van alle objecten tot dit punt \mathbf{y} . Nu willen we alle objecten vinden die een afstand kleiner dan ϵ hebben tot dit punt \mathbf{y} . Beschrijf een efficiënt algoritme dat deze objecten vindt en daarvoor zo min mogelijk nieuwe afstanden uitrekent. Hint: gebruikmakend van de driehoeksongelijkheid zijn er 2 “trucs” te bedenken; één voor kleine afstanden en één voor grote. Het kan helpen een schets voor jezelf te maken.

De cosinus gelijkenis (“cosine similarity”)

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|},$$

wordt wel gebruikt als een maat voor de gelijkenis tussen 2 vectoren.

- (e) [3] Schrijf de cosinus gelijkenis om naar een afstandsmaat $d_{\cos}(\mathbf{x}, \mathbf{y})$ die voldoet aan $0 \leq d_{\cos}(\mathbf{x}, \mathbf{y}) \leq 1$ en verder zodanig is dat grotere gelijkenis overeenkomt met kleinere afstand.

Een afstandsmaat is een metriek als deze voldoet aan de driehoeksongelijkheid, symmetrie en als $d(\mathbf{x}, \mathbf{y}) \geq 0$ met $d(\mathbf{x}, \mathbf{y}) = 0$ d.e.s.d.a. $\mathbf{x} = \mathbf{y}$.

- (f) [6] Is d_{\cos} een metriek? Zo ja, bewijs dit. Zo nee, laat zien waarom niet.

KLAAR!