Midterm Tentamen
"'Data mining"

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

1. Gegeven een grote medische dataset met zo'n 1000 objecten ("records"). Jij wordt als data mining expert ingehuurd om met deze dataset aan de slag te gaan. Het is van belang eerst de uitbijters ("outliers"), die bijvoorbeeld ontstaan zijn door onjuiste invoer, te identificiëren. Geef aan welke technieken voor visualisatie je hiervoor zou gebruiken en hoe je deze in zou zetten,

   (a) **[6]** als deze dataset bestaat uit 1 continue variabele;

   (b) **[10]** als deze dataset bestaat uit 20 continue variabelen en 2 binaire variabelen.

   Stel je voor dat we naar 1 continue variabele uit bovengenoemde medische dataset kijken en het gemiddelde ("mean") en de mediaan ("median") van deze variabele berekenen.

   (c) **[8]** Hoe gevoelig zijn het gemiddelde en de mediaan voor uitbijters? Maak eventueel gebruik van een schets/figuur en/of enkele formules om je antwoord aan de hand van een voorbeeld duidelijk te maken.

   (d) **[6]** We willen een robuuste maat voor de spreiding in deze variabele. Welke maat komt hiervoor in aanmerking? Geef niet alleen de naam, maar ook de formule.

2. De tabel hieronder vat een data set samen met 3 attributen x, y, z en 2 klasselabels, T ("true") en F ("false").

x	y	z	aantal objecten met klasse T	aantal objecten met klasse F
1	1	1	5	0
1	1	0	0	0
1	0	1	0	20
1	0	0	25	0
0	1	1	20	0
0	1	0	0	0
0	0	1	0	5
0	0	0	0	25

We gaan een beslisboom bouwen met 2 lagen. We gebruiken de klassificatiefout ("classification error rate") als criterium om te bepalen op grond van welk attribuut te splitsen.

(a) **[2]** Wat is de klassificatiefout voordat we gaan splitsen? Hint: de klassificatiefout is de fractie van het aantal objecten dat verkeerd wordt geclassificeerd, wanneer we aan alle objecten de klasse toekennen die het meest voorkomt.

(b) **[12]** Wat wordt de klassificatiefout wanneer we splitsen op grond van elk van de drie attributen x, y en z?

(c) **[2]** Welk attribuut moeten we dus kiezen we om als eerste te gaan splitsen, nog steeds uitgaande van klassificatiefout als criterium?

(d) **[6]** Bouw de tweede laag van de beslisboom, d.w.z. herhaal onderdeel (b) en (c) voor elk van de kinderen en de twee overgebleven attributen.

(e) **[2]** Hoeveel objecten in de gegeven data set worden door deze beslisboom verkeerd geclassificeerd?

(f) **[6]** Schets een methode om de voorspellende waarde van de beslisboom op nieuwe, nog niet geziene data, in te schatten.

Bij de zogenaamde gulzige ("greedy"), die we ook hierboven hebben toegepast, wordt achtereenvolgens gesplitst op grond van steeds 1 variabele. Dit maakt deze methode relatief goedkoop: de rekentijd benodigd om een beslisboom te leren is evenredig met het aantal attributen.

(g) **[2]** Wat is het nadeel van deze "greedy" methode?

(h) **[4]** Wat zijn mogelijke alternatieven voor deze "greedy" aanpak?

(i) **[4]** Geef van de alternatieven die je in de vorige vraag hebt gegeven aan hoe duur ze zijn, d.w.z., hoe schaalt voor deze methoden de tijd benodigd om een boom te leren ruwweg als functie van het aantal attributen?

3. Bij veel algoritmen voor data mining wordt gebruik gemaakt van de afstanden ("distances") of meer algemeen de mate van gelijkenis ("proximities") tussen objecten. Afhankelijk van het type objecten kunnen deze op verschillende manieren worden gedefinieerd.

   (a) **[2]** De "simple matching coefficient" (SMC) is gedefinieerd als het aantal overeenkomstige attribuutwaarden gedeeld door het aantal attributen. Bereken de SMC tussen de vectoren $\mathbf{x} = (1, 0, 0, 0, 0, 1, 0, 0, 0, 1)$ en $\mathbf{y} = (0, 1, 0, 0, 0, 1, 0, 0, 1, 0)$.

   (b) **[3]** De "Jaccard coefficient" lijkt op de SMC, maar negeert (en is onafhankelijk van) het aantal overeenkomstige nullen. Voor welk type attributen is de Jaccard coefficient meer geschikt dan de SMC?

   (c) **[3]** Geef een voorbeeld van een data mining probleem waarbij dit type attributen voorkomt.

   Afstandsmaten worden een metriek genoemd wanneer ze voldoen aan de eigenschappen positiviteit, symmetrie en de driehoeksongelijkheid ("triangular inequality").

   (d) **[6]** Wat zijn voordelen van een metriek boven een willekeurige niet-metrische afstandsmaat?

   We gaan laten zien dat de zogenaamde supremum of L_∞ norm, gedefinieerd aan de hand van de afstand

   $$d(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i| \,,$$

   voldoet aan de driehoeksongelijkheid

   $$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \,,$$

   voor alle vectoren $\mathbf{x}$, $\mathbf{y}$ en $\mathbf{z}$.

   (e) **[8]** Laat eerst zien dat de supremum norm voldoet aan de driehoeksongelijkheid in het special geval wanneer x, y en z geen meer-dimensionale vectoren zijn maar "gewoon" getallen (scalars of, zo je wilt, één-dimensionale vectoren). Hint: beschouw zo nodig alle verschillende ordeningen van x, y en z: $x \leq y \leq z$, $x \leq z \leq y$, $y \leq x \leq z$, ...

   (f) **[8]** Laat nu zien dat de supremum norm ook in het algemene geval voldoet aan de driehoeksongelijkheid. Hierbij mag je gebruik maken van wat je bij de vorige vraag hebt aangetoond. Hint: laat als tussenstap zien dat voor willekeurige vectoren $\mathbf{a}$ en $\mathbf{b}$ geldt dat

   $$\max_i(|a_i| + |b_i|) \leq \max_i |a_i| + \max_j |b_j| \,.$$