# Antwoorden werkbladopgaven

d.r.m.langers@pl.hanze.nl

20 oktober 2021

## 1 Week 1

### 1.1 Informatie entropie

**Opgave 1.1.** Het antwoord "ja" op de vraag "ben je ouder dan 42 jaar?" geeft meer informatie dan het antwoord "ja" op de vraag "ben je ouder dan 100 jaar?" omdat deze eerste vraag een vrijwel gelijke kans heeft op de antwoorden "ja"/"nee" en het antwoord op de tweede vraag meestal al vooraf goed voorspeld kan worden. De antwoorden "ja" en "nee" op de vraag "is de kleur van je ogen bruin?" geven precies evenveel informatie omdat het exact dezelfde onzekerheid over het antwoord oplost. Het antwoord "kop" op de vraag "wat was de uitkomst van de worp van een eerlijke munt?" geeft minder informatie dan het antwoord "6" op de vraag "wat was de uitkomst van de worp van een eerlijke dobbelsteen?" omdat de munt maar twee uitkomsten heeft en de dobbelsteen wel zes.

**Opgave 1.2.** De kans op het antwoord "ja" op de vraag "ben je ouder dan 42 jaar?" bedraagt ongeveer 50% (want 42 jaar is de mediaan) en de kans op het antwoord "ja" op de vraag "ben je ouder dan 100 jaar?" bedraagt naar schatting bijvoorbeeld ongeveer 0.1% (in elk geval zeer klein). De kans op het antwoord "ja" op de vraag "is de kleur van je ogen bruin?" bedraagt 75% en de kans op het antwoord "nee" op de vraag "is de kleur van je ogen bruin?" bedraagt 25% als bruine ogen dominant zijn (in werkelijkheid zit het wat ingewikkelder). De kans op het antwoord "kop" op de vraag "wat was de uitkomst van de worp van een eerlijke munt?" bedraagt 50% en de kans op het antwoord "6" op de vraag "wat was de uitkomst van de worp van een eerlijke dobbelsteen?" bedraagt $\frac{1}{6}$.

**Opgave 1.3.** $P(\text{kop}, \text{zes}) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$.

**Opgave 1.4.** Kansen mogen vermenigvuldigd worden voor onafhankelijke gebeurtenissen. Kleurenblindheid en geslacht zijn echter niet onafhankelijk; mannen zijn verhoudingsgewijs vaker kleurenblind dan vrouwen.

**Opgave 1.5.** $H(\text{kop}) = -\lg\left(\frac{1}{2}\right) = 1$ bit. Dit komt vermoedelijk overeen met je beeld van een bit, want een bit kan net als de worp van een munt twee uitkomsten hebben, waarbij er geen voorkeur voor de ene of de andere uitkomst is.

**Opgave 1.6.** $H\,(\text{zes}) = -\lg\left(\frac{1}{6}\right) = 2.585$ bit.

**Opgave 1.7.** Één "nat" betreft een gebeurtenis met kans $P(G)$ zodanig dat $-\ln(P(G)) = 1$, oftewel $P(G) = e^{-1} = \frac{1}{e}$; de entropie in bits is dan $-\lg(\frac{1}{e}) = \lg(e) \approx 1.443$. Soortgelijk, één "dit" betreft een gebeurtenis met kans $P(G)$ zodanig dat $-\log(P(G)) = 1$, oftewel $P(G) = 10^{-1} = \frac{1}{10}$; de entropie in bits is dan $-\lg(\frac{1}{10}) = \lg(10) \approx 3.322$.

**Opgave 1.8.** $H = -\frac{55}{100}\lg(\frac{55}{100}) - \frac{45}{100}\lg(\frac{45}{100}) \approx 0.993$ bits/worp. Dit is iets minder dan 1 bit omdat de munt weliswaar twee uitkomsten heeft maar niet helemaal eerlijk is.

**Opgave 1.9.** Voor $p = 0$ en voor $p = 1$ gaat het mis omdat de logaritme van nul ongedefinieerd is. In dit geval bedraagt de entropie steeds $H = 0$.

**Opgave 1.10.** $H = -\frac{1}{4}\lg(\frac{1}{4}) - \frac{3}{4}\lg(\frac{3}{4}) \approx 0.811$ bits/waarneming. Dit is een stuk minder dan 1 bit omdat de oogkleur weliswaar twee uitkomsten heeft maar dominante bruine en recessieve blauwe ogen voorkomen in een verhouding 3:1 (op basis van een vereenvoudigde genetica van oogkleur).

**Opgave 1.11.** Vier mogelijkheden die even vaak voorkomen hebben een entropie van twee bits/instance (want $4 = 2^2$). Je kunt dit berekenen als $H = -\frac{1}{4}\lg(\frac{1}{4}) - \frac{1}{4}\lg(\frac{1}{4}) - \frac{1}{4}\lg(\frac{1}{4}) - \frac{1}{4}\lg(\frac{1}{4}) = 2.000$ bits/instance.

**Opgave 1.12.** De entropie wordt kleiner dan twee bits/instance als de kansen niet meer gelijk verdeeld zijn.

**Opgave 1.13.** In de contact-lenses dataset is $P(\text{soft}) = \frac{5}{24}$, $P(\text{hard}) = \frac{4}{24}$ en $P(\text{none}) = \frac{15}{24}$. Sommeren over deze mogelijkheden levert op $H = -\frac{5}{24}\lg(\frac{5}{24}) - \frac{4}{24}\lg(\frac{4}{24}) - \frac{15}{24}\lg(\frac{15}{24}) = 1.326$ bits/instance.

**Opgave 1.14.** Laat $p_i$ lopen over de twintig verschillende waarden in de tabel en bereken daarmee $H = \sum_{i=1}^{20} -p_i \cdot \lg(p_i) = -0.074\lg(0.074) - 0.042\lg(0.042) + \ldots \approx 4.201$ bits/aminozuur. Als er typisch 375 aminozuren in een eiwit zitten, dan levert dit in totaal $375 \cdot 4.201 \approx 1575$ bits aan informatie op.

## 1.2 Information gain & Gain ratio

**Opgave 1.15.** Voor Temperature is $H_{\text{in}} = -\frac{9}{14}\lg\left(\frac{9}{14}\right) - \frac{5}{14}\lg\left(\frac{5}{14}\right) = 0.940$, $H_{\text{hot}} = -\frac{2}{4}\lg\left(\frac{2}{4}\right) - \frac{2}{4}\lg\left(\frac{2}{4}\right) = 1.000$, $H_{\text{mild}} = -\frac{4}{6}\lg\left(\frac{4}{6}\right) - \frac{2}{6}\lg\left(\frac{2}{6}\right) = 0.918$, $H_{\text{cool}} = -\frac{3}{4}\lg\left(\frac{3}{4}\right) - \frac{1}{4}\lg\left(\frac{1}{4}\right) = 0.811$, $H_{\text{uit}} = \frac{4}{14}H_{\text{hot}} + \frac{6}{14}H_{\text{mild}} + \frac{4}{14}H_{\text{cool}} = 0.911$, $\Delta H = H_{\text{in}} - H_{\text{uit}} = 0.029$. Voor Humidity is $H_{\text{in}} = -\frac{9}{14}\lg\left(\frac{9}{14}\right) - \frac{5}{14}\lg\left(\frac{5}{14}\right) = 0.940$, $H_{\text{high}} = -\frac{3}{7}\lg\left(\frac{3}{7}\right) - \frac{4}{7}\lg\left(\frac{4}{7}\right) = 0.985$, $H_{\text{normal}} = -\frac{6}{7}\lg\left(\frac{6}{7}\right) - \frac{1}{7}\lg\left(\frac{1}{7}\right) = 0.592$, $H_{\text{uit}} = \frac{7}{14}H_{\text{high}} + \frac{7}{14}H_{\text{normal}} = 0.788$, $\Delta H = H_{\text{in}} - H_{\text{uit}} = 0.152$. Voor Windy is $H_{\text{in}} = -\frac{9}{14}\lg\left(\frac{9}{14}\right) - \frac{5}{14}\lg\left(\frac{5}{14}\right) = 0.940$, $H_{\text{true}} = -\frac{3}{6}\lg\left(\frac{3}{6}\right) - \frac{3}{6}\lg\left(\frac{3}{6}\right) = 1.000$, $H_{\text{false}} = -\frac{6}{8}\lg\left(\frac{6}{8}\right) - \frac{2}{8}\lg\left(\frac{2}{8}\right) = 0.811$, $H_{\text{uit}} = \frac{6}{14}H_{\text{true}} + \frac{8}{14}H_{\text{false}} = 0.892$, $\Delta H = H_{\text{in}} - H_{\text{uit}} = 0.048$.

**Opgave 1.16.** Er gaan twee "yes" en drie "no" instances de "sunny" tak in. Voeren we daarop de diverse berekeningen uit voor alle attributen, dan vinden we de volgende uitkomsten:

| Attribuut | $H_{\text{in}}$ | $H_{\text{uit}}$ | $\Delta H$ |
|---|---|---|---|
| Outlook | 0.971 | 0.971 | 0.000 |
| Temperature | 0.971 | 0.400 | 0.571 |
| Humidity | 0.971 | 0.000 | 0.971 |
| Windy | 0.971 | 0.951 | 0.020 |

Het beste kan gesplitst worden op Humidity, dan worden de leaf-nodes homogeen. Dit komt overeen met het resultaat van de *Id3-tree* classifier.

Er gaan drie "yes" en twee "no" instances de "rainy" tak in. Voeren we daarop de diverse berekeningen uit voor alle attributen, dan vinden we de volgende uitkomsten:

| Attribuut | $H_{\text{in}}$ | $H_{\text{uit}}$ | $\Delta H$ |
|---|---|---|---|
| Outlook | 0.971 | 0.971 | 0.000 |
| Temperature | 0.971 | 0.951 | 0.020 |
| Humidity | 0.971 | 0.951 | 0.020 |
| Windy | 0.971 | 0.000 | 0.971 |

Het beste kan gesplitst worden op Windy, dan worden de leaf-nodes homogeen. Dit komt overeen met het resultaat van de *Id3-tree* classifier.

**Opgave 1.17.** Voor Temperature is $H_{\text{split}} = -\frac{4}{14}\lg\left(\frac{4}{14}\right) - \frac{6}{14}\lg\left(\frac{6}{14}\right) - \frac{4}{14}\lg\left(\frac{4}{14}\right) = 1.557$, $h = 0.029/1.557 = 0.019$. Voor Humidity is $H_{\text{split}} = -\frac{7}{14}\lg\left(\frac{7}{14}\right) - \frac{7}{14}\lg\left(\frac{7}{14}\right) = 1.000$, $h = 0.152/1.000 = 0.152$. Voor Windy is $H_{\text{split}} = -\frac{6}{14}\lg\left(\frac{6}{14}\right) - \frac{8}{14}\lg\left(\frac{8}{14}\right) = 0.985$, $h = 0.048/0.985 = 0.049$.

**Opgave 1.18.** $H_{\text{in}}$ blijft 0.940, $H_{\text{uit}}$ wordt 0.000, dus $\Delta H = 0.940$. $H_{\text{split}} = \sum_{i=1}^{14} -\frac{1}{14} \cdot \lg(\frac{1}{14}) = 14 \cdot (-\frac{1}{14}\lg(\frac{1}{14})) = 3.807$, waarmee $h = \frac{0.940}{3.807} = 0.246$. Zowel qua information gain als qua gain ratio wint dit datum-attribuut hier toch van de vier andere attributen. De *Id3-* en *J48-trees* zouden beide dus slechts uit één node bestaan die splitst op het datum-attribuut.

**Opgave 1.19.** We vinden door berekening de volgende uitkomsten:

| Attribuut | $H_{\text{in}}$ | $H_{\text{uit}}$ | $\Delta H$ | $H_{\text{split}}$ | $h$ |
|---|---|---|---|---|---|
| Age | 1.326 | 1.287 | 0.039 | 1.585 | 0.025 |
| Spectacle-prescrip | 1.326 | 1.287 | 0.040 | 1.000 | 0.040 |
| Astigmatism | 1.326 | 0.949 | 0.377 | 1.000 | 0.377 |
| Tear-prod-rate | 1.326 | 0.777 | 0.549 | 1.000 | 0.549 |

Op grond hiervan kiezen zowel de *Id3-* als *J48-tree* voor het attribuut Tear-prod-rate.

# 2 Week 2

## 2.1 De formule van Bayes

**Opgave 2.1.** $P(\text{thymine}) = \frac{1}{2} \cdot 60\% = 30\%$.

**Opgave 2.2.** Er zijn drie verschillende oneven worpen (één, drie, vijf) die even waarschijnlijk zijn, dus $P(\text{vijf} \mid \text{oneven}) = \frac{1}{3}$.

**Opgave 2.3.** Bij benadering is $P(\text{'q'}|\text{'u'}) = 0.5\%$. Dit is niet hetzelfde als $P(\text{'u'}|\text{'q'})$.

**Opgave 2.4.** $P(\text{blauwe twee}|\text{oranje drie}) = 1$. De uitkomsten van de dobbelstenen zijn niet onafhankelijk (ze zijn zelfs volledig afhankelijk van elkaar).

**Opgave 2.5.** $P(\text{man}, \text{kleurenblind}) = P(\text{man}) \cdot P(\text{kleurenblind}|\text{man})$, waaruit volgt dat $P(\text{kleurenblind}|\text{man}) = P(\text{man}, \text{kleurenblind})/P(\text{man}) = 0.042/0.500 = 0.084$.

**Opgave 2.6.** Er is één koele, zonnige dag binnen veertien instances, dus $P(\text{cool}, \text{sunny}) = \frac{1}{14} = 0.071$. Of, $P(\text{cool}, \text{sunny}) = P(\text{cool}) \cdot P(\text{sunny} \mid \text{cool}) = \frac{4}{14} \cdot \frac{1}{4} = \frac{1}{14}$, danwel $P(\text{sunny}, \text{cool}) = P(\text{sunny}) \cdot P(\text{cool} \mid \text{sunny}) = \frac{5}{14} \cdot \frac{1}{5} = \frac{1}{14}$.

**Opgave 2.7.** $P(\text{hot}) = \frac{4}{14} = 0.286$; $P(\text{rainy}) = \frac{5}{14} = 0.357$; $P(\text{hot}, \text{rainy}) = \frac{0}{14} = 0.000$; $P(\text{hot}|\text{rainy}) = \frac{0}{5} = 0.000$; $P(\text{rainy}|\text{hot}) = \frac{0}{4} = 0.000$.

**Opgave 2.8.** $P(\text{leverziekte} \mid \text{alcoholist}) = \frac{P(\text{leverziekte}) \cdot P(\text{alcoholist}|\text{leverziekte})}{P(\text{alcoholist})} = \frac{0.10 \cdot 0.07}{0.05} = 0.14$, oftewel 14% kans.

**Opgave 2.9.** $P(\text{staking} \mid \text{laatkomers}) = \frac{P(\text{staking}) \cdot P(\text{laatkomers}|\text{staking})}{P(\text{laatkomers})} = \frac{0.01 \cdot 0.60}{0.04} = 0.15$, oftewel 15% kans.

## 2.2 Het "Naive Bayes" algoritme

**Opgave 2.10.** $P(\text{yes} \mid \text{overcast}) = \frac{\frac{9}{14} \cdot \frac{4}{9}}{\frac{4}{14}} = \frac{4}{4} = 1.00$ en $P(\text{no} \mid \text{overcast}) = \frac{\frac{5}{14} \cdot \frac{0}{5}}{\frac{4}{14}} = \frac{0}{4} = 0.00$, dus het label voor bewolkte dagen wordt "yes"; $P(\text{yes} \mid \text{rainy}) = \frac{\frac{9}{14} \cdot \frac{3}{9}}{\frac{5}{14}} = \frac{3}{5} = 0.60$ en $P(\text{no} \mid \text{rainy}) = \frac{\frac{5}{14} \cdot \frac{2}{5}}{\frac{5}{14}} = \frac{2}{5} = 0.40$, dus het label voor regenachtige dagen wordt ook "yes".

**Opgave 2.11.** $P(\text{yes} \mid \text{rainy}, \text{cool}, \text{high}) = \frac{\frac{9}{14} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}}{P(\text{rainy,cool,high})} = \frac{0.024}{P(\text{rainy,cool,high})}$ en $P(\text{no} \mid \text{rainy}, \text{cool}, \text{high}) = \frac{\frac{5}{14} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{4}{5}}{P(\text{rainy,cool,high})} = \frac{0.023}{P(\text{rainy,cool,high})}$, waaruit volgt $P(\text{yes} \mid \text{rainy}, \text{cool}, \text{high}) = \frac{0.024}{0.024+0.023} = 0.510$ en $P(\text{no} \mid \text{rainy}, \text{cool}, \text{high}) = \frac{0.023}{0.024+0.023} = 0.490$. Het klasselabel wordt dan "yes".

**Opgave 2.12.** $P(\text{soft} \mid \text{presbyopic}, \text{myope}, \text{no}, \text{normal}) = \frac{\frac{5}{24} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{5} \cdot \frac{5}{5}}{P(\text{presbyopic,myope,no,normal})} = \frac{0.017}{P(\text{presbyopic,myope,no,normal})}$ en $P(\text{hard} \mid \text{presbyopic}, \text{myope}, \text{no}, \text{normal}) = \frac{\frac{4}{24} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{0}{4} \cdot \frac{4}{4}}{P(\text{presbyopic,myope,no,normal})} = \frac{0.000}{P(\text{presbyopic,myope,no,normal})}$ en $P(\text{none} \mid \text{presbyopic}, \text{myope}, \text{no}, \text{normal}) = \frac{\frac{15}{24} \cdot \frac{6}{15} \cdot \frac{7}{15} \cdot \frac{7}{15} \cdot \frac{3}{15}}{P(\text{presbyopic,myope,no,normal})} = \frac{0.011}{P(\text{presbyopic,myope,no,normal})}$, waaruit volgt $P(\text{soft} \mid \text{presbyopic}, \text{myope}, \text{no}, \text{normal}) = \frac{0.017}{0.017+0.000+0.011} = 0.605$ en $P(\text{hard} \mid \text{presbyopic}, \text{myope}, \text{no}, \text{normal}) = \frac{0.000}{0.017+0.000+0.011} = 0.000$ en $P(\text{none} \mid \text{presbyopic}, \text{myope}, \text{no}, \text{normal}) = \frac{0.011}{0.017+0.000+0.011} = 0.395$. Het klasselabel wordt dan "soft", terwijl de data "none" vermeldt.

**Opgave 2.13.** $P(\text{soft} \mid \text{presbyopic, myope, no, normal}) = \frac{\frac{6}{27}\cdot\frac{2}{8}\cdot\frac{3}{7}\cdot\frac{6}{7}\cdot\frac{6}{7}}{P(\text{presbyopic,myope,no,normal})} = \frac{0.017}{P(\text{presbyopic,myope,no,normal})}$ en $P(\text{hard} \mid \text{presbyopic, myope, no, normal}) = \frac{\frac{5}{27}\cdot\frac{2}{7}\cdot\frac{4}{6}\cdot\frac{1}{6}\cdot\frac{5}{6}}{P(\text{presbyopic,myope,no,normal})} = \frac{0.005}{P(\text{presbyopic,myope,no,normal})}$ en $P(\text{none} \mid \text{presbyopic, myope, no, normal}) = \frac{\frac{16}{27}\cdot\frac{7}{18}\cdot\frac{8}{17}\cdot\frac{8}{17}\cdot\frac{4}{17}}{P(\text{presbyopic,myope,no,normal})} = \frac{0.012}{P(\text{presbyopic,myope,no,normal})}$, waaruit volgt $P(\text{soft} \mid \text{presbyopic, myope, no, normal}) = \frac{0.017}{0.017+0.005+0.012} = 0.509$ en $P(\text{hard} \mid \text{presbyopic, myope, no, normal}) = \frac{0.005}{0.017+0.005+0.012} = 0.142$ en $P(\text{none} \mid \text{presbyopic, myope, no, normal}) = \frac{0.012}{0.017+0.005+0.012} = 0.349$. Het klasselabel blijft dan "soft".

**Opgave 2.14.** Naive Bayes met één attribuut maximaliseert $P(C \mid A)$, wat wil zeggen dat per waarde van het attribuut $A$ dat klasselabel $C$ wordt gekozen dat de hoogste waarschijnlijkheid heeft. Dat is dus ook het klasselabel met het grootste aantal instances voor die waarde van het attribuut, en dat is exact wat OneR zou doen. Naive Bayes met nul attributen maximaliseert $P(C)$, wat wil zeggen dat over de hele dataset dat klasselabel $C$ wordt gekozen dat de hoogste waarschijnlijkheid heeft. Dat is dus ook het klasselabel met het grootste aantal instances in de dataset, en dat is exact wat ZeroR zou doen.

# 3 Week 3

## 3.1 De confusion matrix

**Opgave 3.1.** ACC $= \frac{9}{14}$; ERR $= \frac{5}{14}$; FPR $= \frac{3}{5}$; TPR $= \frac{7}{9}$; FNR $= \frac{2}{9}$; TNR $= \frac{2}{5}$; PPV $= \frac{7}{10}$; NPV $= \frac{2}{4}$; FDR $= \frac{3}{10}$; FOR $= \frac{2}{4}$.

**Opgave 3.2.** Deze kans is gelijk aan de False Discovery Rate (FDR).

**Opgave 3.3.** De onvoorwaardelijke kans $P$ (voorspelde klasse=ware klasse) komt overeen met de accuracy (ACC). De voorwaardelijke kans $P$ (voorspelde klasse=negatief|ware klasse=positief) komt overeen met de False Negative Rate (FNR). De Precision oftewel Positive Predictive Value staat voor de voorwaardelijke kans $P$ (ware klasse=positief|voorspelde klasse=positief).

**Opgave 3.4.** De confusion matrix bevat de volgende aantallen (of een veelvoud daarvan).

| | | Toegekende klasse: | |
|---|---|---|---|
| | | Positief | Negatief |
| Ware klasse: | Positief | 12 | 8 |
| | Negatief | 3 | 12 |

De accuracy is ACC $= \frac{24}{35} = 0.686$.

**Opgave 3.5.** De confusion matrix bevat de volgende aantallen (of een veelvoud daarvan).

| | | Toegekende klasse: | |
|---|---|---|---|
| | | Positief | Negatief |
| Ware klasse: | Positief | 171 | 9 |
| | Negatief | 19 | 25 |

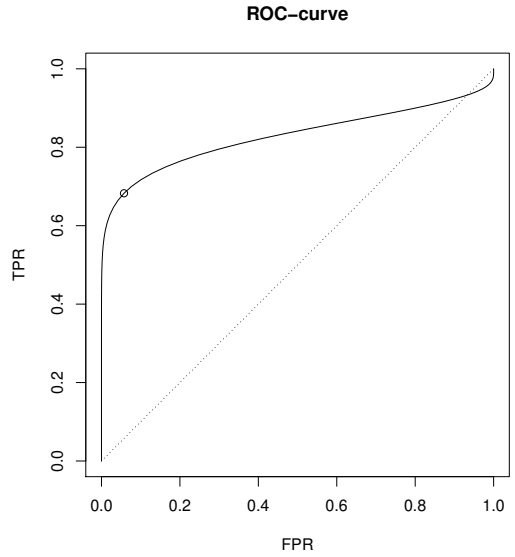
De False Positive Rate is FPR $= \frac{19}{44} = 0.432$.

**Opgave 3.6.** Voor *Iris setosa* zijn de Precision en Recall 100.0%; voor *Iris versicolor* is de Precision $\frac{45}{52} = 86.5\%$ en de Recall $\frac{45}{50} = 90.0\%$; voor *Iris virginica* is de Precision $\frac{43}{48} = 89.6\%$ en de Recall $\frac{43}{50} = 86.0\%$.

## 3.2 De ROC-curve

**Opgave 3.7.** Omdat de rode cruve vrijwel geheel links van de grenswaarde ligt wordt geen enkele negatieve als positief geclassificeerd, zodat bijvoorbeeld FPR=0.1%. Tegelijkertijd ligt de helft van de groene curve aan de verkeerde kant van de grenswaarde, zodat TPR=50%.

**Opgave 3.8.** De hieronder getoonde ROC-curve heeft AUC = 0.829. Een geschikte grenswaarde zou om en nabij $x' = -\frac{1}{2}$ gelegd kunnen worden, hetgeen een behoorlijke TPR = 0.683 combineert met een goede FPR = 0.057 en de accuracy nagenoeg maximaliseert (zie het gemarkeerde punt in de grafiek).




**Opgave 3.9.** Voor $P\,(\text{Yes}) = \frac{1}{4}$ vinden we als confusion matrix:

| | | Toegekende klasse: | |
|---|---|---|---|
| | | Yes | No |
| Ware klasse: | Yes | 9 | 0 |
| | No | 3 | 2 |

Hieruit volgt dat TPR = $\frac{9}{9} = 1.00$ en FPR = $\frac{3}{5} = 0.60$.

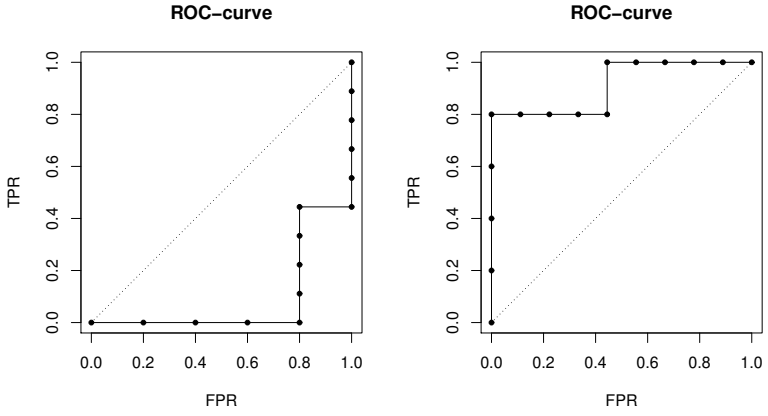
**Opgave 3.10.** Het punt met TPR = $\frac{9}{9} = 1.00$ en FPR = $\frac{3}{5} = 0.60$ vind je midden in het vlak lopende deel van de ROC-curve uiterst bovenin.

**Opgave 3.11.** Bij cost-sensitive classification wordt de drempel gelegd op $P(\text{Yes}) = \frac{C(\text{FP})}{C(\text{FP})+C(\text{FN})} = \frac{3.5}{3.5+1.0} = \frac{7}{9} = 0.778$. Dit leidt tot de volgende confusion matrix:

| | | Toegekende klasse: | |
|---|---|---|---|
| | | Yes | No |
| Ware klasse: | Yes | 5 | 4 |
| | No | 0 | 5 |

Hieruit volgt dat TPR $= \frac{5}{9} = 0.556$ en FPR $= \frac{0}{5} = 0.00$. Je zit dan in de meest linker knie van de ROC-curve.

**Opgave 3.12.** Wanneer alleen de door het algoritme toegekende klasselabels worden omgekeerd verkrijgen we de linker ROC-curve hieronder; de nieuwe $\text{AUC}_{\text{omgekeerd}} = 1 - \text{AUC}_{\text{origineel}}$, met een slecht presterend algoritme als resultaat omdat het meestal precies het verkeerde label toekent. Wanneer de No-instances als positief worden gezien en de Yes-instances als negatief verkrijgen we de rechter ROC-curve; de nieuwe $\text{AUC}_{\text{omgekeerd}} = \text{AUC}_{\text{origineel}}$, met een goed presterend algoritme als resultaat omdat de juistheid van de classificaties niet afhangt van welke klasse als de positieve of als de negatieve wordt beschouwd.




**Opgave 3.13.** Samengevat vinden we:

| | $P(\text{Yes}) = \frac{1}{4} = 0.250$ | $P(\text{Yes}) = \frac{1}{2} = 0.500$ | $P(\text{Yes}) = \frac{7}{9} = 0.778$ |
|---|---|---|---|
| TPR | $\frac{9}{9} = 1.000$ | $\frac{9}{9} = 1.000$ | $\frac{5}{9} = 0.556$ |
| TNR | $\frac{2}{5} = 0.400$ | $\frac{4}{5} = 0.800$ | $\frac{5}{5} = 1.000$ |
| PPV | $\frac{9}{12} = 0.750$ | $\frac{9}{10} = 0.900$ | $\frac{5}{5} = 1.000$ |
| NPV | $\frac{2}{2} = 1.000$ | $\frac{1}{1} = 1.000$ | $\frac{5}{9} = 0.556$ |
| ACC | $\frac{11}{14} = 0.786$ | $\frac{13}{14} = 0.929$ | $\frac{10}{14} = 0.714$ |

De drempels $P(\text{Yes}) = 0.250$ en $P(\text{Yes}) = 0.500$ leveren allebei de beste TPR en NPV omdat deze respectievelijk het hoogst in de grafiek liggen en de hoek $\beta$ maximaliseren; de drempel $P(\text{Yes}) = 0.778$ levert de beste TNR en PPV omdat deze respectievelijk het meest links in de grafiek ligt en de hoek $\alpha$ maximaliseert. De beste ACC wordt behaald met de drempel $P(\text{Yes}) = 0.500$ omdat dit punt van de ROC-curve het verst schuin boven

de diagonaal ligt.

# 4 Week 4

## 4.1 Medische diagnostiek

**Opgave 4.1.** Het Id3-Tree algoritme is niet in staat om om te gaan met numerieke attributen of missing values; al deze komen in deze datasets wel voor.

**Opgave 4.2.** Maak gebruik van de Weka Experimenter om te vinden dat OneR, J48-Tree, Random Forest en Adaboost op de AUTISM dataset de hoogste accuracy van 100.00% behalen; Zero-R behaalt op de SCHIZO dataset de laagste accuracy van 52.06%.

**Opgave 4.3.** Op de breast-cancer dataset behaalt Classificatie via Clustering een significant hogere error rate dan Zero-R, en op de sick dataset geldt dat voor zowel Naive Bayes als Classificatie via Clustering.

**Opgave 4.4.** Het Random Forest behaalt de hoogste gemiddelde AUC van 0.88.

## 4.2 Granulocyten

**Opgave 4.5.** $a_0 = 1641.0662$; $a_1 = -96.0918$; $a_2 = -0.269$.

**Opgave 4.6.** Resubstitutie: 100%; leave-one-out: 90%.

**Opgave 4.7.** Resubstitutie: 1.00; leave-one-out: 0.99.

**Opgave 4.8.** Ja, deze set is lineair separabel, want het logistische model scheidt de trainingsdata perfect middels een rechte lijn aangezien de resubstitution error 0% is.

**Opgave 4.9.** Cost-sensitive classification leidt hier wel tot een verandering, maar cost-sensitive learning niet: geen weging van kosten geeft 3 classificatiefouten (2 FP, 1 FN); cost-sensitive classification geeft 2 classificatiefouten (1 FP, 1FN); cost-sensitive learning geeft 3 classificatiefouten (2 FP, 1FN). Cost-sensitive classification houdt rekening met kosten door tijdens de classificatie de drempel aan te passen om de kosten te minimaliseren, maar het getrainde model blijft hetzelfde; cost-sensitive learning houdt rekening met de kosten door de instances uit de klasse met de kostbaarste fouten zwaarder te wegen tijdens de training, maar de drempeling blijft hetzelfde.

**Opgave 4.10.** Er worden 3 clusters gerapporteerd (met ignore class!).

**Opgave 4.11.** De centroids zijn $(d = -0.2934, V = -0.7023)$ en $(d = 0.8801, V = 1.2578)$. Er wordt 1 instance verkeerd geclusterd (met ignore class!).

## 4.3 Wie Is Het?

**Opgave 4.12.** De naam van de persoon is de klasse; dit is een kwalitatieve categorische discrete nominale variabele.

**Opgave 4.13.** De entropie is gelijk aan $H = 24 \cdot -\frac{1}{24} \lg\left(\frac{1}{24}\right) = \lg(24) = 4.58$ bits; omdat een ja/nee-vraag maximaal 1 bit informatie oplevert vereist dit dus gemiddeld minstens vijf vragen.

**Opgave 4.14.** De resubstitutie error rate is 0% omdat je test op de data waarmee ook al getraind is en de tree zuivere leaf-nodes heeft die de instances volledig scheiden; de kruisgevalideerde error rate is 100% omdat de (unieke) kaartjes in de testset en de trainingsset niet overlappen en de tree dus nooit de kans krijgt om getraind wordt op de namen uit de testset.

**Opgave 4.15.** De root-node test op het attribuut bigmouth, dus de beste vraag luidt "heeft jouw persoon een grote mond?". Dit splitst de data in een verhouding 10:14. In de ene tak is de entropie $H = \lg(10) = 3.32$ bits, in de andere $H = \lg(14) = 3.81$ bits, dus de verwachtingswaarde is $\frac{10}{24} \cdot 3.32 + \frac{14}{24} \cdot 3.81 = 3.61$ bits.

**Opgave 4.16.** Volg de tak bigmouth = no om bij bald = ? terecht te komen. Volg dan beide takken; met weging 9: bald = no naar hairstuff = no naar gingerhair = no naar curlyhair = yes naar blondhair = no, voorspelt "Anne"; en met weging 5: bald = yes naar gingerhair = no naar whitehair = no naar brownhair = ?: met weging 1 naar brownhair = yes, voorspelt "Richard", en met weging 1 naar brownhair = no, voorspelt "Tom". Dit geeft 5/28 kans op Richard, 5/28 kans op Tom, 9/14 kans op Anne, dus de toegewezen klasse wordt Anne.

**Opgave 4.17.** Op basis van $2^{22} = 4194304$ subset evaluaties kunnen hat, longhair, facialhair, beard en earrings worden verwijderd.

**Opgave 4.18.** De Id3-Tree splitst op name en de J48-Tree splitst op hairpartition in de root-node. Dit verschil wordt veroorzaakt doordat Id3 information gain en J48 gain ratio als criterium gebruiken.

**Opgave 4.19.** Middels de CVParameterSelection meta-classifier wordt een optimum $k = 5$ gevonden.