Toets deel 2 Data-analyse en retrieval
Vrijdag 1 Juli 2016: 11.00-13.00
Uitwerking

Opgave 1: Naive Bayes voor tekstclassificatie (16 punten)

Het Vocabulaire bestaat uit:

1. goed
2. onderhouden
3. mooie
4. ligging
5. speeltuin
6. geweldig
7. sanitair
8. verwaarloosd
9. slecht
10. schoongemaakt
11. oud
12. vies

$|V|$ is dus 12.

(a)

$$\hat{P}(\texttt{slecht}|\text{Negatief}) = \frac{2+1}{9+12} = \frac{1}{7}$$

$$P(\texttt{slecht}|\text{Positief}) = \frac{0+1}{8+12} = \frac{1}{20}$$

(b)

$$\hat{P}(\texttt{mooie} = 1|\text{Positief}) = \frac{2+1}{2+2} = \frac{3}{4}$$
$$\hat{P}(\texttt{mooie} = 1|\text{Negatief}) = \frac{0+1}{2+2} = \frac{1}{4}$$

Opgave 2: Clustering (24 punten)

(a) Zie het boek en/of de slides.

(b) Nee, het algoritme vindt slechts een lokaal optimum. Je kunt voor $K = 2$ een goede vinden en voor $K = 3$ een slechte. Verzin zelf een voorbeeld. Overigens had ik de vraag niet goed genoeg dicht getimmerd zodat er natuurlijk weer slimmerikken waren die een dataset met slechts 2 punten namen. Dan werken ze inderdaad allebei even goed. Dit heb ik ook maar goed gerekend.

(c)

$$\text{RI} = \frac{17+71}{136} = 0.647$$

Opgave 3: Gemengde Vragen (18 punten)

(a) Definieer predictorvariabelen:

- X_1: perceeloppervlakte (in m^2)
- X_2: aantrekkelijke locatie (ja=1, nee=0)
- X_3: $X_1 \times X_2$

Je mag X_2 ook weglaten, in dat geval is er geen "vaste premie" voor een aantrekkelijke locatie. Dan zit $X_1 \times X_2$ wel als predictor in het model, maar X_2 niet. Dat is in strijd met het hiërarchieprincipe, maar er stond niet in de opgave dat je je daar per se aan moest houden.

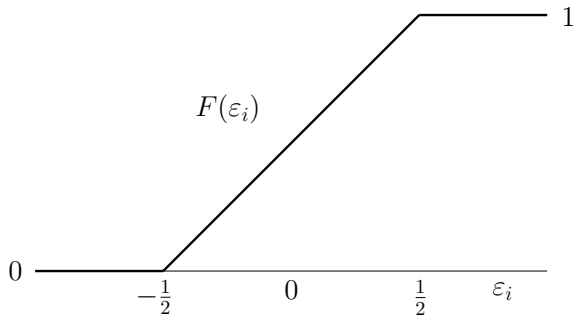
Het weglaten van X_1 is wel zeer dubieus, omdat in dat geval voor een huis dat niet op een aantrekkelijke locatie staat, de prijs voor een extra vierkante meter nul is. Dat zou tamelijk absurd zijn.

(b) D
Uitleg: er is geen informatie gegeven over de variantie.

(c)

$$P(y_i = 1 \mid x_i) = P(\varepsilon_i < \beta_0 + \beta_1 x_i) = \begin{cases} 0 & \text{als } \beta_0 + \beta_1 x_i < -\frac{1}{2} \\ 1 & \text{als } \beta_0 + \beta_1 x_i > \frac{1}{2} \\ \frac{1}{2} + \beta_0 + \beta_1 x_i & \text{anders.} \end{cases}$$

Noteer $P(\varepsilon_i < \beta_0 + \beta_1 x_i) = F(\beta_0 + \beta_1 x_i)$, waarbij $F(\varepsilon_i)$ de cumulatieve kansdichtheidsfunctie hieronder is (de vette lijn):



Opgave 4: Logistische Regressie (24 punten)

(a) Vul `TLOC` = 10 in in de geschatte logistische responsfunctie:

$$\hat{P}(\texttt{defect}{=}1 \mid \texttt{TLOC}{=}10) = \frac{e^{-1.05+10\times0.064}}{1+e^{-1.05+10\times0.064}} = 0.399$$

Dus ongeveer 40%.

(b) B

(c) Ja,

$$z = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{0.064}{0.007} \approx 9.14$$

De kritieke waarde ligt ongeveer bij $z = 2$, dus daar gaan we ruimschoots overheen.

(d) De decision boundary is:

$$\begin{aligned} -1.05 + 0.064 \times \texttt{TLOC} &= 0 \\ 0.064 \times \texttt{TLOC} &= 1.05 \\ \texttt{TLOC} &= \frac{1.05}{0.064} = 16.41 \end{aligned}$$

De classificatieregel is dus: als het aantal regels code meer dan 1641 is, voorspel defect, anders geen defect.

(e) We gaan er van uit dat, zoals gebruikelijk, klasse 1 met de "positives" correspondeert, bijvoorbeeld $TP = 159$. Als je klasse 0 als positives hebt aangemerkt, dan kon je 4 punten verdienen.

$$\begin{aligned} \text{Accuracy} &= \frac{290+159}{661} = 0.697 \\ \text{Precision} &= \frac{159}{159+58} = 0.733 \\ \text{Recall} &= \frac{159}{159+154} = 0.508 \\ F_1 &= \frac{2\times0.733\times0.508}{0.733+0.508} = 0.6 \end{aligned}$$

Opgave 5: Ordinale Classificatie (18 punten)

(a) De handigste manier om dit uit te rekenen is:

$$P(y \leq 2) = \frac{e^{2.7270 - 0.7643 \times 0.5 + 0.4590 \times 0.5}}{1 + e^{2.7270 - 0.7643 \times 0.5 + 0.4590 \times 0.5}} = 0.929$$
$$P(y = 3) = 1 - P(y \leq 2) = 1 - 0.929 = 0.071.$$

De kans is dus ongeveer 7.1%.

(b) De correct ingevulde tabel is:

	daalt	stijgt	kan beide
$\hat{P}(y = 1)$	X		
$\hat{P}(y = 2)$			X
$\hat{P}(y = 3)$		X	

(c) C

Uitleg: zelfs in het "meest relevante" geval (volgens het model), namelijk `TF` =1 en `DL` = 0 (een absurde combinatie natuurlijk), hebben we

$$P(y = 1) = \frac{e^{1.3776 - 0.7643 \times 1 + 0.4590 \times 0}}{1 + e^{1.3776 - 0.7643 \times 1 + 0.4590 \times 0}} = 0.65.$$

Dus het model wijst alle gevallen aan klasse 1 toe en maakt dan 427+190=617 fouten.