

Tehtävä 1. Naivi-Bayes ja SPAM (2 pistettä).

Ehdolliset todennäköisyydet $P(\text{SANA}_i = s \mid \text{spam})$ ja $P(\text{SANA}_i = s \mid \text{ham})$, missä $s \in \{\text{million, dollars, adclick, conferences}\}$:*)

sana	spam	ham
million	$156/95\,791 \approx 0.001629$	$98/306\,438 \approx 0.000320$
dollars	$29/95\,791 \approx 0.000303$	$119/306\,438 \approx 0.000388$
adclick	$51/95\,791 \approx 0.000532$	$0/306\,438 = 0.0$
conferences	$0/95\,791 = 0.0$	$12/306\,438 \approx 0.000039$

*) Huom: taulukon frekvenssit ovat (vahingossa) eri lukuja kuin tehtävän 2 tiedostoissa olevat.

i.

$$\begin{aligned}
 P(\text{SANA}_i \neq \text{million}) &= 1 - P(\text{SANA}_i = \text{million}) \\
 &= 1 - (P(\text{SANA}_i = \text{million} \mid \text{spam}) P(\text{spam}) + \\
 &\quad P(\text{SANA}_i = \text{million} \mid \text{ham}) P(\text{ham})) \\
 &= 1 - (0.001629 \cdot 0.5 + 0.000320 \cdot 0.5) \approx 1 - 0.000974 = 0.999026
 \end{aligned}$$

Laskussa tarvittiin ns. ärsyttävän nimittäjän laskemista, mikä onnistuu marginalisointi-kaavalla $P(A) = P(A, B) + P(A, \neg B)$. Kumpikin tapaus on vielä jaettu ketjusäännöllä muotoon $P(A, B) = P(A \mid B) P(B)$, jolloin kaikki tarvittavat todennäköisyysarvot (tai niiden arviot) tunnetaan.

ii.

$$P(\text{spam} \mid \text{million}) = \frac{P(\text{million} \mid \text{spam}) P(\text{spam})}{P(\text{million})} \approx \frac{0.001629 \cdot 0.5}{0.000974} \approx 0.836.$$

Kyseessä oli suoraviivainen Bayesin kaavan sovellus. Ärsyttävä nimittäjä oli laskettu edellisessä kohdassa.

iii. Nyt on kätevintä laskea osamäärä

$$\begin{aligned}
 \text{Odds} &= \frac{P(\text{spam} \mid \text{million, dollars, adclick, conferences})}{P(\text{ham} \mid \text{million, dollars, adclick, conferences})} \\
 &= \frac{P(\text{spam})}{P(\text{ham})} \times \frac{P(\text{million} \mid \text{spam})}{P(\text{million} \mid \text{ham})} \times \frac{P(\text{dollars} \mid \text{spam})}{P(\text{dollars} \mid \text{ham})} \\
 &\quad \times \frac{P(\text{adclick} \mid \text{spam})}{P(\text{adclick} \mid \text{ham})} \times \frac{P(\text{conferences} \mid \text{spam})}{P(\text{conferences} \mid \text{ham})} \\
 &\approx \frac{0.5}{0.5} \times \frac{0.001629}{0.000320} \times \frac{0.000303}{0.000388} \times \frac{0.000532}{0.0} \times \frac{0.0}{0.000039} = \frac{0.0}{0.0}
 \end{aligned}$$

Ongelmaksi muodostuivat siis nollatodennäköisyydet. Korvataan ne lukuarvolla 0.000010, jolloin edellisestä kaavasta saadaan tulokseksi $\text{Odds} \approx 53.97$. Tämä tuottaa todennäköisyyden

$$P(\text{spam} \mid \text{million, dollars, adclick, conferences}) \approx \frac{53.97}{1 + 53.97} \approx 0.982.$$

Tehtävä 2. Roskapostisuodattimen toteutus (2 pistettä).

Java-ratkaisu tulee kurssin sivulle.

Esimerkkiviestien kohdalla osamääräksi tuli

$$\text{spamesim.txt} : \text{Odds} = 1.220 \times 10^{26} \Leftrightarrow P(\text{spam} \mid \text{evidenssi}) \approx 1.000$$

$$\text{hamesim.txt} : \text{Odds} = 1.118 \times 10^{-45} \Leftrightarrow P(\text{spam} \mid \text{evidenssi}) \approx 0.000,$$

kun käytettiin nollatodennäköisyyksien paikalla vakiota 0.00001.

Tehtävä 3. Shakki. (2 pistettä)

Ei valmista esimerkkiratkaisua. Heuristisessa arviointifunktiossa oli hyvä oottaa huomioon mm. kummankin pelaajan nappuloiden määrä, laatu ja sijainti ruudulla. Kivin monimutkaisia funktioita ei tarvinnut. Tärkeämpää on säätää painotuksia ja kokeilla, miten hyvin botti pärjää. Yksi suoritusta parantava keino näytti olevan myös se, että kannustaa bottia etenemään melko aggressiivisesti, jottei peli pääty liian usein tasapeliin.