

1 Bayesin teoreeman käyttö luokittelijana

Bayesin kaavan mukaan merkityksen s_k kontekstille c ehdollistettu todennäköisyys voidaan määrittää alla olevan yhtälön perusteella:

$$P(s_k|c) = \frac{P(c|s_k)P(s_k)}{P(c)} \quad (1)$$

Tämä yhtälö voidaan tulkita sanoin seuraavasti. Todennäköisyys, että s_k on se luokka, joka kuvaa kontekstia c on sama kuin todennäköisyys sille, että konteksti c löytyy luokasta s_k kertaa luokan s_k esiintymisen todennäköisyys jaettuna kontekstin c ”esiintymisen” (priori) todennäköisyydellä.

1.1 Naiivi bayesluokitin

Yhtälö 1 voidaan yksinkertaistaa ns. ”naiiviksi” olettamalla, että konteksti c koostuu piirteistä w_i , jotka ovat toisistaan riippumattomia. Tämän oletuksen perusteella saadaan yhtälö

$$s' = \operatorname{argmax}_{s_k} P(c|s_k)P(s_k) \quad (2)$$

$$s' = \operatorname{argmax}_{s_k} (P(s_k) \prod_{i=0}^{N-1} P(w_i|s_k)) \quad (3)$$

Nyt yhtälössä (3) esiintyvä w_i kuvaa jonkin piirteen esiintymistodennäköisyyttä kontekstissa c ehdolla (luokassa) s_k . Yhtälöissä esiintyvä s' kuvaa luokituksen tulosta eli luokkaa mihin tutkittava näyte bayesiläisen todennäköisyyden mukaan kuuluu. Edellä esitetty yhtälö voidaan kirjoittaa myös logaritimuotoon seuraavasti:

$$s' = \operatorname{argmax}_{s_k} (\log P(s_k) + \sum_{i=0}^{N-1} \log P(w_i|s_k)) \quad (4)$$

1.2 Naivin bayesluokittimen käyttäminen

Olkoon, että bayesluokittimemme tuntee 32 sanaa ($N_l = 32$). Nämä sanat on saatu kahdeksasta lauseesta, joista kolme on lauseita jotka ovat mielestämme ”kiinnostavia” (merk. s_{kiin}) ja loput viisi ovat ”roskaa” (merk. s_{rosk}). Näin ollen bayesluokittimemme voisi pitää sisällään seuraavaa tietoa luokasta s_{roska} :

sana	1	2	3	4	5	esiintymiskerrat yht.
<i>hello</i>	1	0	0	1	1	3
<i>china</i>	0	0	0	0	0	0
<i>girl</i>	1	0	0	1	1	3
<i>viagra</i>	1	1	1	1	0	4
<i>thomas</i>	0	0	0	0	0	0
...	...	...				

ja vastavasti tietoa luokasta s_{kiin} :

sana	1	2	3	esiintymiskerrat yht.
<i>hello</i>	1	1	0	2
<i>china</i>	0	1	0	1
<i>girl</i>	0	0	1	1
<i>viagra</i>	0	0	0	0
<i>thomas</i>	1	1	0	2
...	...	...		

Nyt haluamme tutkia, ovatko seuraavat tuntemattomat lauseet yllä olevan luokittimemme perusteella ”roskaa” vai ”kiinnostavia”.

- a) $w = \{\textit{hello}, \textit{thomas}, \textit{can}, \textit{you}, \textit{buy}, \textit{me}, \textit{a}, \textit{drink}\}$
- b) $w = \{\textit{hello}, \textit{do}, \textit{you}, \textit{need}, \textit{a}, \textit{viagra}\}$

Ennen kuin voimme tutkia lauseiden ”kiinnostavuutta”, tulee meidän määrittää mitä on todennäköisyys $P(w_i|s_k)$ eli piirteen w_i todennäköisyys luokassa s_k . Esimerkkimme tapauksessa todennäköisyys voidaan tulkita seuraavasti: mikä on kontekstissa esiintyvän sanan w_i esiintymisen todennäköisyys luokassa s_k ? Nyt määritetään, että

$$P(w_i|s_k) = \frac{C(w_i|s_k) + \lambda}{C(s_k) + N_l \lambda} \quad (5)$$

missä $C(w_i|s_k)$ on sanan w_i esiintymisen lukumäärä luokassa s_k ja termi λ on menetelmän käyttäjän määrittämä arvo, joka kuvaa prioriuskomuksen vahvuutta. Termin tarkoituksena on estää tilanne, jossa jonkin sanan esiintymistodennäköisyys olisi nolla. Termi $C(s_k)$ on arvo, joka kertoo kuinka monta opetusnäytettä kuuluu luokkaan s_k . Edellisessä yhtälössä esiintyvä N_l on luku, joka kuvaa sitä kuinka monta erillistä sanaa luokittimemme tunnistaa.

- a) Nyt lauseemme on muotoa $w = \{\textit{hello}, \textit{thomas}, \textit{can}, \textit{you}, \textit{buy}, \textit{me}, \textit{a}, \textit{drink}\}$. Olkoon prioriuskomuksemme vahvuus $\lambda = 0.5$. Tutkitaan ensin, kuuluuko lause luokkaan s_{rosk} . Sovelletaan yhtälöitä 2 ja 5 sekä edellä esi-

tettyä taulukkoa. Näiden perusteella saadaan sanojen esiintymistodennäköisyyksiksi

$$\begin{aligned} P(\text{"hello"}|s_{\text{rosk}}) &= \frac{3 + 0.5}{5 + 0.5 \times 32} \\ P(\text{"thomas"}|s_{\text{rosk}}) &= \frac{0 + 0.5}{5 + 0.5 \times 32} \\ P(\text{"can"}|s_{\text{rosk}}) &= P(\text{"you"}|s_{\text{rosk}}) = \dots = P(\text{"drink"}|s_{\text{rosk}}) = \frac{0 + 0.5}{5 + 0.5 \times 32} \end{aligned}$$

Yhdistelemällä yllä lasketut arvot saamme todennäköisyydeksi, että lause kuuluu luokkaan s_{rosk} :

$$\begin{aligned} s_{\text{rosk}} &= \frac{5}{8} \times P(\text{"hello"}|s_{\text{rosk}}) \times \dots \times P(\text{"drink"}|s_{\text{rosk}}) \\ s_{\text{rosk}} &= \frac{5}{8} \times \frac{3.5}{21} \times \frac{0.5^7}{21^7} \\ s_{\text{rosk}} &= 4.51839 \times 10^{-13} \end{aligned}$$

Seuraavaksi laskemme todennäköisyyden, että lause kuuluu luokkaa s_{kiin} . Nyt sanojen esiintymistodennäköisyydet ovat

$$\begin{aligned} P(\text{"hello"}|s_{\text{kiin}}) &= \frac{2 + 0.5}{3 + 0.5 \times 32} \\ P(\text{"thomas"}|s_{\text{kiin}}) &= \frac{2 + 0.5}{3 + 0.5 \times 32} \\ P(\text{"can"}|s_{\text{kiin}}) &= P(\text{"you"}|s_{\text{kiin}}) = \dots = P(\text{"drink"}|s_{\text{kiin}}) = \frac{0 + 0.5}{5 + 0.5 \times 32} \end{aligned}$$

Yhdistelemällä yllä lasketut arvot saamme todennäköisyydeksi, että lause kuuluu luokkaan s_{kiin} :

$$\begin{aligned} s_{\text{kiin}} &= \frac{3}{8} \times P(\text{"hello"}|s_{\text{kiin}}) \times \dots \times P(\text{"drink"}|s_{\text{kiin}}) \\ s_{\text{kiin}} &= \frac{3}{8} \times \frac{2.5}{19} \times \frac{2.5}{19} \times \frac{0.5^6}{19^6} \\ s_{\text{kiin}} &= 2.15626 \times 10^{-12} \end{aligned}$$

Koska $s_{\text{kiin}} > s_{\text{rosk}}$ päättelemme, että lauseen w luokitus s' on "kiinnostava".

- b) Nyt lauseemme on muotoa $w = \{hello, do, you, need, a, viagra\}$ Tutkitaan ensin, että kuuluuko lause luokkaan s_{rosk} . Sovelletaan yhtälöitä 2 ja 5 sekä edellä esitettyä taulukkoa. Näiden perusteella saadaan sanojen esiintymistodennäköisyyksiksi

$$P("hello" | s_{rosk}) = \frac{3 + 0.5}{5 + 0.5 \times 32}$$

$$P("do" | s_{rosk}) = \frac{0 + 0.5}{5 + 0.5 \times 32}$$

$$P("you" | s_{rosk}) = P("need" | s_{rosk}) = P("a" | s_{rosk}) = \frac{0 + 0.5}{5 + 0.5 \times 32}$$

$$P("viagra" | s_{rosk}) = \frac{4 + 0.5}{5 + 0.5 \times 32}$$

Yhdistelemällä yllä lasketut arvot saamme todennäköisyydeksi, että lause kuuluu luokkaan s_{rosk} :

$$s_{rosk} = \frac{5}{8} \times P("hello" | s_{rosk}) \times \dots \times P("viagra" | s_{rosk})$$

$$s_{rosk} = \frac{5}{8} \times \frac{3.5}{21} \times \frac{0.5^4}{21^4} \times \frac{4.5}{21}$$

$$s_{rosk} = 0.0430 \times 10^{-6}$$

Seuraavaksi laskemme todennäköisyyden, että lause kuuluu luokkaa s_{kiin} . Nyt sanojen esiintymistodennäköisyydet ovat

$$P("hello" | s_{kiin}) = \frac{2 + 0.5}{3 + 0.5 \times 32}$$

$$P("do" | s_{kiin}) = \frac{0 + 0.5}{3 + 0.5 \times 32}$$

$$P("you" | s_{kiin}) = P("need" | s_{kiin}) = P("a" | s_{kiin}) = \frac{0 + 0.5}{5 + 0.5 \times 32}$$

$$P("viagra" | s_{kiin}) = \frac{0 + 0.5}{5 + 0.5 \times 32}$$

Yhdistelemällä yllä lasketut arvot saamme todennäköisyydeksi, että lause kuuluu luokkaan s_{kiin} :

$$s_{kiin} = \frac{3}{8} \times P("hello" | s_{kiin}) \times \dots \times P("viagra" | s_{kiin})$$

$$s_{kiin} = \frac{3}{8} \times \frac{2.5}{19} \times \frac{0.5^5}{19^5}$$

$$s_{kiin} = 0.00166 \times 10^{-6}$$

Koska $s_{rosk} > s_{kiin}$ päättelemme, että lauseen w luokitus s' on ”roskaa”.