CS3244 Tutorial 7

Zhuang Jianning
A0214561M
T13

BL11

Derive incremental version of Bayes rule

$$P(h \mid D_1, D_2) = \frac{P(D_1, D_2 \mid h)\, P(h)}{P(D_1, D_2)}$$

conditional independence of D_1 and D_2 given h

$$= \frac{P(D_1 \mid h) \cdot P(D_2 \mid h) \cdot P(h)}{P(D_1, D_2)}$$

$P(D_1, h) = P(D_1 \mid h) \cdot P(h)$
$= P(h \mid D_1) \cdot P(D_1)$

$$= \frac{P(D_2 \mid h) \cdot P(h \mid D_1) \cdot P(D_1)}{P(D_1, D_2)}$$

marginalization with mutually exclusive $h \in H$

$$= \frac{P(D_2 \mid h) \cdot P(h \mid D_1) \cdot P(D_1)}{\sum_{h \in H} P(D_1, D_2, h)}$$

$P(D_1)$ independent of h

$$= \frac{P(D_2 \mid h) \cdot P(h \mid D_1) \cdot \cancel{P(D_1)}}{\sum_{h \in H} P(D_2, h \mid D_1) \cdot \cancel{P(D_1)}}$$

$$\frac{P(D_2, h, D_1)}{P(D_1)} = \frac{P(D_2, h, D_1)}{P(h, D_1)} \cdot \frac{P(h, D_1)}{P(D_1)}$$

$$= \frac{P(D_2 \mid h) \cdot P(h \mid D_1)}{\sum_{h \in H} P(D_2 \mid h, D_1) \cdot P(h \mid D_1)}$$

conditional independence of D_1 and D_2 given h

$$= \frac{P(D_2 \mid h) \cdot P(h \mid D_1)}{\sum_{h \in H} P(D_2 \mid h) \cdot P(h \mid D_1)}$$

//

TM 6.5

n boolean attributes

m examples.

bits to encode one of the Boolean attributes $= 1 + \log_2 n$

bits to encode misclassification $= 1 + \log_2 m$

a) MDL minimizes $L_{C_1}(h) + L_{C_2}(D|h)$ where $L_C(x)$ is description length of x under encoding C

$= \alpha(1+\log_2 n) + \beta(1+\log_2 m)$ where α is # boolean attributes in h
β is # misclassified examples

b) A maximally general hypothesis with all don't cares will be inconsistent with any negative training example

To minimize the description length of misclassification, let $m = 1 \Rightarrow 1$ training example

description length $= 0(1+\log_2 n) + 1(1+\log_2 1)$
$= 1$

Any hypothesis that is consistent with a negative training example has to have at least 1 boolean attribute. For the description length to be greater than 1 $\Rightarrow 1(1+\log_2 n) + 0(1+\log_2 1) > 1$
$\Rightarrow n \geq 2$

A possible dataset with 1 negative training example with 2 boolean attributes would be

$\{ \langle \text{Sunny}, \text{Warm} \rangle, - \}$

maximally general hypothesis $\langle ?, ? \rangle$ chosen instead of any other consistent hypothesis as MDL $= 1$ lower than any other consistent hypothesis MDL ≥ 2

c) For MDL to output MAP hypothesis

$$\underset{h \in H}{\operatorname{argmin}} \ \alpha(H \log_2 n) + \beta(H \log_2 m) = \underset{h \in H}{\operatorname{argmin}} -\log_2 P(D|h) - \log_2 P(h)$$

$$\underset{h \in H}{\operatorname{argmin}} \ \alpha + \alpha \log_2 n + \beta + \beta \log_2 m$$

$$= \underset{h \in H}{\operatorname{argmin}} \ \log_2 2^\alpha + \log_2 n^\alpha + \log_2 2^\beta + \log_2 m^\beta$$

$$= \underset{h \in H}{\operatorname{argmin}} \ \log_2 (2n)^\alpha + \log_2 (2m)^\beta$$

$$= \underset{h \in H}{\operatorname{argmin}} \ -\log_2 \left(\frac{1}{2n}\right)^\alpha - \log_2 \left(\frac{1}{2m}\right)^\beta$$

which attribute, boolean combination

$$\sum_{h \in H} \left(\frac{1}{2n}\right)^\alpha = \sum_{\alpha=0}^{n} \binom{n}{\alpha} 2^\alpha \left(\frac{1}{2n}\right)^\alpha$$

$$= \sum_{\alpha=0}^{n} \binom{n}{\alpha} \left(\frac{1}{n}\right)^\alpha (1)^{n-\alpha}$$

$$= \left(1 + \frac{1}{n}\right)^n$$

since $\sum_{h \in H} P(h) = 1 \Rightarrow P(h) = \frac{1}{\left(1+\frac{1}{n}\right)^n} \left(\frac{1}{2n}\right)^\alpha$

which examples

$$\sum_{D} \left(\frac{1}{2m}\right)^\beta = \sum_{\beta=0}^{m} \binom{m}{\beta} \left(\frac{1}{2m}\right)^\beta$$

$$= \sum_{\beta=0}^{m} \binom{m}{\beta} \left(\frac{1}{2m}\right)^\beta (1)^{m-\beta}$$

$$= \left(1 + \frac{1}{2m}\right)^m$$

since $\sum_{D} P(D|h) = 1 \Rightarrow P(D|h) = \frac{1}{\left(1+\frac{1}{2m}\right)^m} \left(\frac{1}{2m}\right)^\beta$

B2 12a

prior beliefs of hypotheses w_A, w_B, w_C, w_F are equal and sum to 1

$$p(h) = \frac{1}{4}$$

$h \in \{w_A, w_B, w_C, w_F\}$

posterior belief $\quad p(h|D) = \dfrac{p(D|h) \cdot p(h)}{p(D)} \quad$ (Bayes theorem)

	d_1	d_2	d_3	d_4
A	1/1	1/1	1/−1	−1
B	1/1	1/1	−1/−1	1
$(x_1 \text{ OR } x_2)$ XOR $(x_3 \text{ OR } x_4)$ C	1/1	1/1	−1/−1	−1
D	1/1	1/1	−1/−1	−1

$\boxed{\text{output / target}}$

$$p(D|w_A) = p(t_{d_1}|w_A) \cdot p(t_{d_2}|w_A) \cdot p(t_{d_3}|w_A) = 1 \cdot 1 \cdot 0 = 0$$

$$p(D|w_B) = 1 \cdot 1 \cdot 1 = 1$$

$$p(D|w_C) = 1 \cdot 1 \cdot 1 = 1$$

$$p(D|w_F) = 1 \cdot 1 \cdot 1 = 1$$

$$p(w_A|D) = \frac{0 \cdot \frac{1}{4}}{3 \cdot \frac{1}{4}} = 0$$

$$p(w_B|D) = \frac{1 \cdot \frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

$$p(w_C|D) = \frac{1}{3}$$

$$p(w_F|D) = \frac{1}{3}$$

$$\underset{t \in T}{\operatorname{argmax}}\, p(t|D) = \underset{t \in T}{\operatorname{argmax}} \sum_{h \in H} p(t|h)\, p(h|D)$$

$$\sum_{h \in H} p(+|h)\, p(h|D) = \frac{1}{3} \quad (B)$$

$$\sum_{h \in H} p(-|h)\, p(h|D) = \frac{2}{3} \quad (A, C, D)$$

Bayes Optimal classification is −