CS3244 Tutorial 7

Wu Zhaoxuan
wu.zhaoxuan@u.nus.edu

2022.4.13

BL 11

In the solution to question TM 6.1 in Tutorial 5, I have shown you the use of the "incremental" version of Bayes' rule. One may wonder how this "incremental" version can be derived from the original Bayes' Theorem. In this question, you are asked to derive the "incremental" version of Bayes' rule.

We know

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

By assuming the conditional independence of data D_1 and data D_2 given h, derive

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

BL 11

Solution

$$
\begin{aligned}
P(h \mid D_1, D_2) &= \frac{P(D_1, D_2 \mid h) P(h)}{P(D_1, D_2)} \\
&= \frac{P(D_2 \mid h) P(D_1 \mid h) P(h)}{P(D_2 \mid D_1) P(D_1)} \\
&= \frac{P(D_2 \mid h) P(h \mid D_1) P(D_1)}{\sum_{h \in H} P(D_2, h \mid D_1) P(D_1)} \\
&= \frac{P(D_2 \mid h) P(h \mid D_1)}{\sum_{h \in H} P(D_2 \mid h, D_1) P(h \mid D_1)} \\
&= \frac{P(D_2 \mid h) P(h \mid D_1)}{\sum_{h \in H} P(D_2 \mid h) P(h \mid D_1)}
\end{aligned}
$$

Minimum Description Length (MDL)

- Remember the MAP hypothesis

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h)P(h)$$
$$= \arg\min_{h \in H} -\log_2 P(h) - \log_2 P(D \mid h)$$

- From information theory point of view,
  - The first term: description length of h under the optimal code
  - The second term: description length of D given h under the optimal code
  - Also, optimal code for a message with probability p is $-\log_2 p$ bits
- Write it in another way, using $L_C(x)$ to denote description length of x under encoding C

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D \mid h)$$

TM 6.5

Consider the Minimum Description Length principle applied to the hypothesis space H consisting of conjunctions of up to n Boolean attributes (e.g., $Sunny \wedge Warm$). Assume each hypothesis is encoded simply by listing the attributes present in the hypothesis where the number of bits needed to encode any one of the n Boolean attributes is $1 + \log_2 n$ (i.e., $\log_2 n$ bits to indicate which of the n Boolean attributes is present in the hypothesis and 1 bit to indicate its corresponding Boolean value). Suppose that the encoding of an example given the hypothesis uses zero bits if the example is consistent with the hypothesis and uses $1 + \log_2 m$ bits otherwise (i.e., 1 bit to indicate the correct classification and $\log_2 m$ bits to indicate which of the m examples was misclassified).

(a) Write down the expression for the quantity to be minimized according to the Minimum Description Length principle.

TM 6.5

(a) Write down the expression for the quantity to be minimized according to the Minimum Description Length principle.

Solution

Minimize $x_h(1 + \log_2 n) + y_h(1 + \log_2 m)$.

Where x_h is the number of Boolean attributes present in hypothesis h

And y_h is the number of misclassified examples by hypothesis h

TM 6.5

(b) Is it possible to construct a set of training data such that a consistent hypothesis exists, but MDL chooses an inconsistent hypothesis? If so, give such a set of training data. If not, explain why not.

Solution

- Yes, it is possible.
- Recall that it minimises $x_h(1 + \log_2 n) + y_h(1 + \log_2 m)$. Make the number m of training examples much small than the number n of boolean attributes.
- For example, one training example with lots of attributes. $\langle Sunny, Warm, High, Strong \rangle, -ve$
- Any consistent hypothesis needs to specific at least one attribute, e.g. $\langle Rainy, ?, ?, ? \rangle$. Then the description length is $1(1 + \log_2 4) + 0(1 + \log_2 1) = 3$.
- A hypothesis $\langle ?, ?, ?, ? \rangle$ is inconsistent. The description length is $0(1 + \log_2 4) + 1(1 + \log_2 1) = 1$. Thus, this inconsistent hypothesis is selected by MDL.

TM 6.5

(c) Give probability distributions for $P(h)$ and $P(D \mid h)$ such that the above MDL algorithm outputs MAP hypotheses.

Solution

Recall

- $h_{MAP} = \arg\min_{h \in H} -\log_2 P(D \mid h) - \log_2 P(h)$

$$h_{MDL} = \arg\min_{h \in H}(x_h(1 + \log_2 n) + y_h(1 + \log_2 m))$$

$$= \arg\min_{h \in H}(\log_2 2^{x_h} + \log_2 n^{x_h} + \log_2 2^{y_h} + \log_2 m^{y_h})$$

- $$= \arg\min_{h \in H} -\log_2(1/(2n))^{x_h} - \log_2(1/(2m))^{y_h}$$

- Does this mean $P(h) = (1/(2n))^{x_h}$ and $P(D \mid h) = (1/(2m))^{y_h}$?

TM 6.5

- $P(h) = (1/(2n))^{x_h}$ and $P(D \mid h) = (1/(2m))^{y_h}$ are not yet probability distributions. Probability distribution has to fulfil an important property: it must sums to 1 => i.e. we need to find the normalizing constant.

$$\sum_{h \in H} P(h) = \frac{1}{(1+1/n)^n} \sum_{h \in H} (1/(2n))^{x_h}$$

$$= \frac{1}{(1+1/n)^n} \sum_{x_h=0}^{n} C(n, x_h) 2^{x_h} (1/(2n))^{x_h}$$

$$= \frac{1}{(1+1/n)^n} \sum_{x_h=0}^{n} C(n, x_h) (1/n)^{x_h} 1^{x_h}$$

By Binomial Theorem, $(x+y)^n = \sum_{k=0}^{n} C(n,k) x^k y^{n-k}$

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

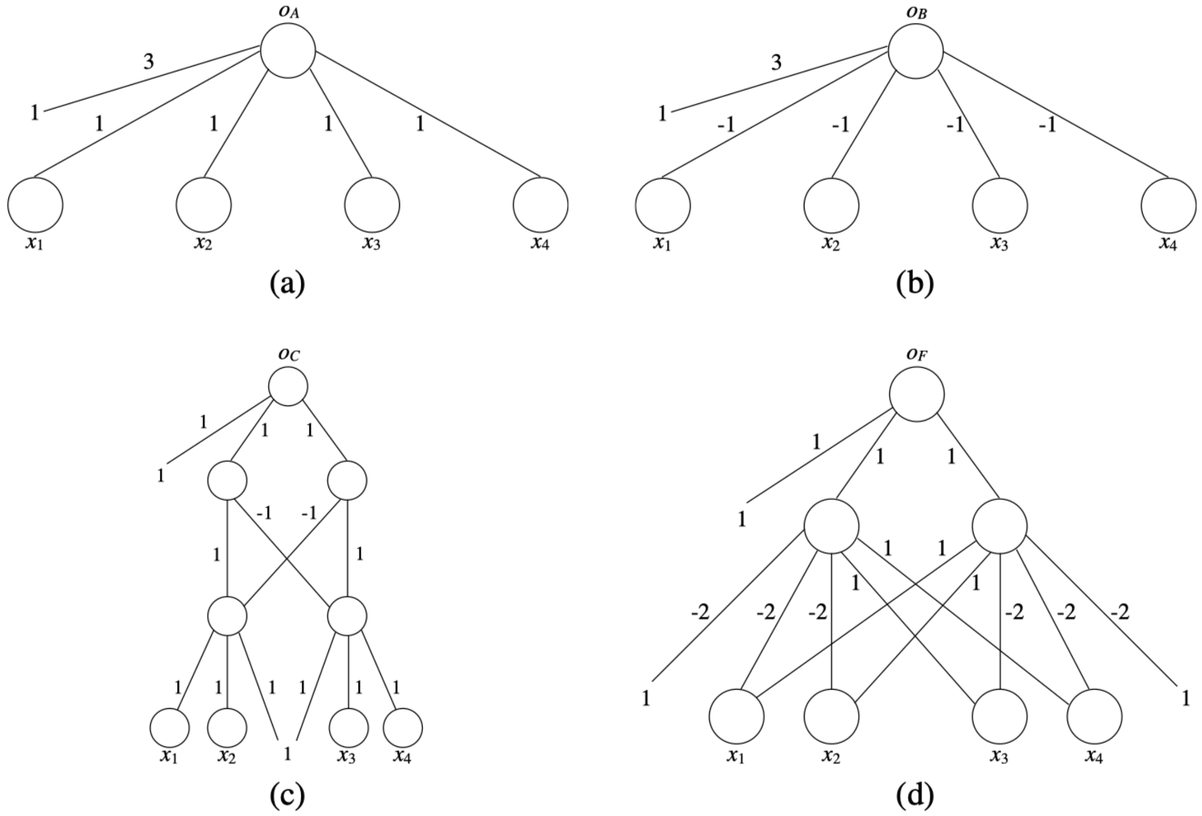
$$= 1$$

- Therefore, we have $P(h) = \frac{(1/(2n))^{x_h}}{(1+1/n)^n}$. Similarly, $P(D \mid h) = \frac{(1/(2m))^{y_h}}{(1+1/(2m))^m}$.

BL 12a

$D = \{d_1 = \langle(-1,-1,-1,1),1\rangle, d_2 = \langle(1,1,-1,-1),1\rangle, d_3 = \langle(1,1,1,1),-1\rangle\}.$

Suppose that the prior beliefs of weights $\mathbf{w}_A, \mathbf{w}_B, \mathbf{w}_C, \mathbf{w}_F$ *are equal and they sum to 1. Using Bayes' Theorem, calculate posterior beliefs* $P(\mathbf{w}_A \mid D), P(\mathbf{w}_B \mid D), P(\mathbf{w}_C \mid D), P(\mathbf{w}_F \mid D)$.

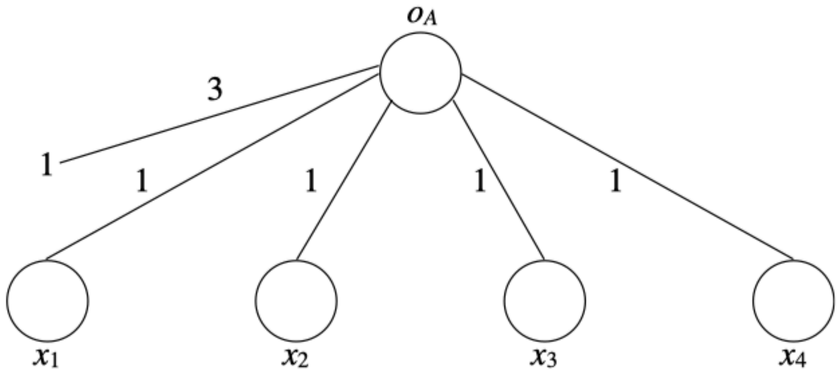


BL 12a

$$P(\mathbf{w}_A \mid D) = \frac{P(D \mid \mathbf{w}_A)P(\mathbf{w}_A)}{P(D)}$$

$$= \frac{\left[\prod_{i=1}^{3} P(t_{d_i} \mid \mathbf{w}_A, \mathbf{x}_{d_i})\right] P(\mathbf{w}_A)}{P(D)}$$

$$= 0 \quad \text{since } P(t_{d_3} \mid \mathbf{w}_A, \mathbf{x}_{d_3}) = 0$$



(a)

BL 12a

Follow the same procedure for all three other networks.

For fast treatment, you might see that A implements OR gate, B implements NAND gate, and both C and F implements $(x_1 OR x_2) XOR (x_3 OR x_4)$.

$$D = \{d_1 = \langle(-1,-1,-1,1),1\rangle, d_2 = \langle(1,1,-1,-1),1\rangle, d_3 = \langle(1,1,1,1),-1\rangle\}$$

$P(t_{d_i} | \mathbf{w}, \mathbf{x}_{d_i}) = 1$ for all $\mathbf{w} \in \{\mathbf{w}_B, \mathbf{w}_C, \mathbf{x}_F\}$ and $d_i \in D$. Thus,

$$P(\mathbf{w}_F | D) = P(\mathbf{w}_C | D) = P(\mathbf{w}_B | D) = \frac{\left[\prod_{i=1}^{3} P(t_{d_i} | \mathbf{w}_B, \mathbf{x}_{d_i})\right] P(\mathbf{w}_B)}{P(D)} = \frac{0.25}{P(D)}$$

$$P(\mathbf{w}_F | D) + P(\mathbf{w}_C | D) + P(\mathbf{w}_B | D) + P(\mathbf{w}_A | D) = \frac{0.75}{P(D)} \Rightarrow 1 = \frac{0.75}{P(D)} \Rightarrow P(D) = 0.75$$

Therefore, $P(\mathbf{w}_F | D) = P(\mathbf{w}_C | D) = P(\mathbf{w}_B | D) = \frac{0.25}{0.75} = \frac{1}{3}$.

Bayes-Optimal Classifier

- Motivation
  - Consider H with 3 possible hypotheses:
    - $P(h_1 \mid D) = 0.4, P(h_2 \mid D) = 0.3, P(h_3 \mid D) = 0.3.$
  - The h_{MAP} (h_1 in this case) is the most probable hypothesis, but it does not give the most probable prediction. Because…
  - Suppose for a new instance $\mathbf{x}$,
    - $h_1(\mathbf{x}) = +, h_2(\mathbf{x}) = -, h_3(\mathbf{x}) = -.$
  - The most probable classification should be $-$.
- Therefore, we define ***Bayes-optimal Classification*** as

$$\arg\max_{t \in T} P(t \mid D) = \arg\max_{t \in T} \sum_{h \in H} P(t \mid h) P(h \mid D)$$

BL 12a (Last Part)

Using the posterior beliefs calculated above, compute the Bayes-optimal classification for the new input instance $\mathbf{x}_{d_4} = (-1, -1, -1, -1)$*. Show the steps of your derivation.*

Solution

Bayes-optimal classification: $\arg\max_{t \in T} P(t \mid D) = \arg\max_{t \in T} \sum_{h \in H} P(t \mid h) P(h \mid D)$

$$
\begin{aligned}
P(t_{d_4} = -1 \mid D, \mathbf{x}_{d_4}) &= P(t_{d_4} = -1 \mid \mathbf{w}_B, \mathbf{x}_{d_4}) P(\mathbf{w}_B \mid D) \\
&\quad + P(t_{d_4} = -1 \mid \mathbf{w}_C, \mathbf{x}_{d_4}) P(\mathbf{w}_C \mid D) \\
&\quad + P(t_{d_4} = -1 \mid \mathbf{w}_F, \mathbf{x}_{d_4}) P(\mathbf{w}_F \mid D) \\
&= (0 + 1 + 1)(1/3) \\
&= 2/3
\end{aligned}
$$

$P(t_{d_4} = 1 \mid D, \mathbf{x}_{d_4}) = 1/3.$

Therefore, the Bayes-optimal classification for the new instance $\mathbf{x}_{d_4} = (-1, -1, -1, -1)$ is $t_{d_4} = -1$.

Thank you!

- Any questions?