National University of Singapore
School of Computing
CS3244 Machine Learning

Tutorial 7: Bayesian Inference and Computational Learning Theory

Issue: April 1, 2020 Due: April 6, 2020 (10am)

Important Instructions:

- *Your solutions for this tutorial must be TYPE-WRITTEN.*
- *SUBMIT YOUR SOLUTIONS in PDF format to the 'TUTORIAL 7 SUBMISSION' folder under Files in LumiNUS by the DUE DATE specified above. Late submissions will NOT be entertained.*
- *Indicate your NAME, STUDENT NUMBER, and TUTORIAL GROUP in your submitted solution.*
- *YOUR SOLUTION TO QUESTION* **BL 11** *will be GRADED for this tutorial.*
- *You may discuss the content of the questions with your classmates. But everyone should work out and write up ALL the solutions by yourself.*

BL 11 In the solution to question TM 6.1 in Tutorial 6, 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.

Specifically, let $h \in H$. Using the Bayes' Theorem, we know that

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

By assuming the conditional independence of data D_1 and data D_2 given hypothesis h, give a step-by-step derivation of the following "incremental" version of Bayes' rule:

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

and state any result/assumption that you have used in each step.

Solution.

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

The second equality is due to conditional independence assumption in the numerator and product rule in the denominator. The third equality is due to product rule or Bayes' Theorem: $P(D_1, h) = P(D_1|h)P(h) = P(h|D_1)P(D_1)$ in the numerator and marginalization in the denominator. The fourth equality follows from product rule in the denominator. The last equality is due to conditional independence assumption in the denominator.

BL 12 On page 39 of the "Bayesian Inference" lecture slides, it is claimed that setting the m-th Gaussian mean parameter to

$$\mu'_m \leftarrow \frac{\sum_{d\in D} \mathbb{E}[z_{dm}]\, x_d}{\sum_{d\in D} \mathbb{E}[z_{dm}]}$$

minimizes

$$\sum_{d\in D}\sum_{m=1}^{M} \mathbb{E}[z_{dm}](x_d - \mu'_m)^2 \ .$$

Show that this claim is true.

Solution.

$$\begin{aligned} \frac{\partial}{\partial \mu'_m}\sum_{d\in D}\sum_{m=1}^{M} \mathbb{E}[z_{dm}](x_d - \mu'_m)^2 &= \sum_{d\in D} \mathbb{E}[z_{dm}]\frac{\partial}{\partial \mu'_m}(x_d - \mu'_m)^2 \\ &= -2\sum_{d\in D} \mathbb{E}[z_{dm}](x_d - \mu'_m) \\ &= -2\left(\sum_{d\in D} \mathbb{E}[z_{dm}]x_d - \mu'_m \sum_{d\in D} \mathbb{E}[z_{dm}]\right) \\ &= 0\ . \\ \mu'_m &= \frac{\sum_{d\in D} \mathbb{E}[z_{dm}]\, x_d}{\sum_{d\in D} \mathbb{E}[z_{dm}]} \end{aligned}$$

BL 13 On page 9 of the "Computational Learning Theory" lecture slides, I have shown you during lecture how the resulting expression of the true error $error_Q(h)$ of hypothesis h can be derived and interpreted if Q is a uniform probability distribution over X.

In this question, show that if $Q(x) = 0$ for all $x \in X$ such that $h(x) = c(x)$, then $error_Q(h) = 1$.

Solution. Let $X' = \{x \in X | c(x) = h(x)\}$. Then, $X \setminus X' = \{x \in X | c(x) \neq h(x)\}$. Then,

$$\begin{aligned}
error_Q(h) &= \sum_{x \in X} (1 - \delta_{h(x),c(x)}) Q(x) \\
&= \sum_{x \in X'} (1 - \delta_{h(x),c(x)}) Q(x) + \sum_{x \in X \setminus X'} (1 - \delta_{h(x),c(x)}) Q(x) \\
&= \sum_{x \in X'} 0 \times Q(x) + \sum_{x \in X \setminus X'} 1 \times Q(x) \\
&= \sum_{x \in X \setminus X'} Q(x) \\
&= \sum_{x \in X'} \underbrace{Q(x)}_{=0} + \sum_{x \in X \setminus X'} Q(x) \\
&= \sum_{x \in X} Q(x) \\
&= 1
\end{aligned}$$

BL 14 This question pertains to analyzing the sample complexity in setting 2 on page 7 of the "Computational Learning Theory" lecture slides where the teacher (who knows the target concept c) selects training examples of the form $\langle x, c(x) \rangle$ for the learner (assume c is in learner's hypothesis space H).

(a) Consider H = conjunctions of up to 3 Boolean literals and their negations. How many training examples suffice to learn the target concept $c = \langle 0, ?, ? \rangle$? What are they? Trace the version space.

(b) Consider H = conjunctions of up to 3 Boolean literals and their negations. How many training examples suffice to learn the target concept $c = \langle 0, ?, 1 \rangle$? What are they? Trace the version space.

Solution.

(a)
- $x_1 = \langle 0, 1, 1 \rangle, c(x_1) = 1$
  $VS_{H,D} = \{\langle 0, 1, 1 \rangle, \langle ?, 1, 1 \rangle, \langle 0, ?, 1 \rangle, \langle 0, 1, ? \rangle, \langle ?, ?, 1 \rangle, \langle ?, 1, ? \rangle, \langle 0, ?, ? \rangle, \langle ?, ?, ? \rangle\}$

- $x_2 = \langle 0, 0, 0 \rangle$, $c(x_2) = 1$
  $VS_{H,D} = \{\langle 0, ?, ? \rangle, \langle ?, ?, ? \rangle\}$
- $x_3 = \langle 1, 1, 1 \rangle$, $c(x_3) = 0$
  $VS_{H,D} = \{\langle 0, ?, ? \rangle\}$

(b)
- $x_1 = \langle 0, 1, 1 \rangle$, $c(x_1) = 1$
  $VS_{H,D} = \{\langle 0, 1, 1 \rangle, \langle ?, 1, 1 \rangle, \langle 0, ?, 1 \rangle, \langle 0, 1, ? \rangle, \langle ?, ?, 1 \rangle, \langle ?, 1, ? \rangle, \langle 0, ?, ? \rangle, \langle ?, ?, ? \rangle\}$
- $x_2 = \langle 0, 0, 1 \rangle$, $c(x_2) = 1$
  $VS_{H,D} = \{\langle 0, ?, 1 \rangle, \langle ?, ?, 1 \rangle, \langle 0, ?, ? \rangle, \langle ?, ?, ? \rangle\}$
- $x_3 = \langle 1, 1, 1 \rangle$, $c(x_3) = 0$
  $VS_{H,D} = \{\langle 0, ?, 1 \rangle, \langle 0, ?, ? \rangle\}$
- $x_4 = \langle 0, 1, 0 \rangle$, $c(x_4) = 0$
  $VS_{H,D} = \{\langle 0, ?, 1 \rangle\}$