National University of Singapore
School of Computing
CS3244 Machine Learning

Tutorial 6: Bayesian Inference

Issue: March 18, 2020 Due: March 23, 2020 (10am)

Important Instructions:

- *Your solutions for this tutorial must be TYPE-WRITTEN.*
- *SUBMIT YOUR SOLUTIONS in PDF format to the 'TUTORIAL 6 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* **TM 6.1** *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.*

TM 6.1 Consider again the medical diagnosis example applying Bayes rule, as described on page 6 of the "Bayesian Inference" lecture slides. Suppose that the doctor decides to order a second lab test for the same patient and the second test returns a positive result as well. What are the posterior probabilities of $cancer$ and $\neg cancer$ following these two tests? Assume that the results of the two tests are conditionally independent given the patient's state of having cancer (or not). Give your answer to 1 decimal place.

Solution. You can certainly use the normal version of Bayes' rule to solve this question. Alternatively, what I will show you below is the "incremental" version of Bayes' rule that utilizes the solution presented during lecture on page 6 and yields the same answer:

$$\begin{aligned}
P(cancer|++) &= \frac{P(+|cancer)P(cancer|+)}{P(+|\neg cancer)P(\neg cancer|+) + P(+|cancer)P(cancer|+)} \\
&= \frac{.98 \times .20851063829}{.03 \times .7914893617 + .98 \times .20851063829} \\
&= 0.89589552238 \\
P(\neg cancer|++) &= 1 - P(cancer|++) \\
&= 1 - 0.89589552238 \\
&= 0.10410447762
\end{aligned}$$

TM 6.4 In the analysis of concept learning on page 10 of the "Bayesian Inference" lecture slides, we assume that the input instances $\mathbf{x}_d$ for $d \in D$ are fixed. Therefore, in deriving an expression for $P(D|h)$, we only need to consider the probability of observing the target outputs $t_d = c(\mathbf{x}_d)$ for $d \in D$ for these fixed input instances $\mathbf{x}_d$ for $d \in D$. Consider the more general setting in which the input instances are not fixed, but are drawn independently from some probability distribution defined over the instance space X. The data D must now be described as the set of ordered pairs $\{\langle \mathbf{x}_d, t_d \rangle\}$ and $P(D|h)$ must now reflect the probability of encountering the specific input instance $\mathbf{x}_d$ as well as the probability of the observed target output t_d. Show that the expression for the posterior belief $P(h|D)$ on page 10 of the "Bayesian Inference" lecture slides holds even under this more general setting.
Hint: Consider the analysis on page 16 of the "Bayesian Inference" lecture slides.

Solution.

$P(h) = \dfrac{1}{|H|}$ for all $h \in H$.

From page 16 of the "Bayesian Inference" lecture slides,

$$\begin{aligned} P(D|h) &= \prod_{d \in D} P(\mathbf{x}_d, t_d|h) \\ &= \prod_{d \in D} P(t_d|h, \mathbf{x}_d) P(\mathbf{x}_d) \\ &= \prod_{d \in D} P(t_d|h, \mathbf{x}_d) \prod_{d \in D} P(\mathbf{x}_d) \\ &= \begin{cases} \displaystyle\prod_{d \in D} P(\mathbf{x}_d) & \text{if } h \text{ is consistent with } D, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

If h is inconsistent with D, then

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{0 \times P(h)}{P(D)} = 0 \, .$$

If h is consistent with D, then

$$\begin{aligned}
P(h|D) &= \frac{P(D|h)P(h)}{P(D)} \\
&= \frac{\prod_{d\in D} P(\mathbf{x}_d) \times \frac{1}{|H|}}{P(D)} \\
\sum_{h\in VS_{H,D}} P(h|D) &= \sum_{h\in VS_{H,D}} \frac{\prod_{d\in D} P(\mathbf{x}_d)}{|H|P(D)} \\
1 &= \frac{\prod_{d\in D} P(\mathbf{x}_d)}{|H|P(D)} \sum_{h\in VS_{H,D}} 1 \\
P(D) &= \prod_{d\in D} P(\mathbf{x}_d) \times \frac{|VS_{H,D}|}{|H|} \\
P(h|D) &= \frac{\prod_{d\in D} P(\mathbf{x}_d) \times \frac{1}{|H|}}{\prod_{d\in D} P(\mathbf{x}_d) \times \frac{|VS_{H,D}|}{|H|}} \\
&= \frac{1}{|VS_{H,D}|}
\end{aligned}$$

Therefore,

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D, \\ 0 & \text{otherwise.} \end{cases}$$

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 $\log_2 n$. 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.

(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.

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

Solution.

(a) $x_h \log_2 n + y_h(1 + \log_2 m)$ where x_h is the number of Boolean attributes present in hypothesis h and y_h is the number of misclassified examples incurred by hypothesis h.

(b) Consider the following $-$ve training example as the training data:

$$\langle Sunny, Warm, High, Strong\rangle, - \ .$$

- Any hypothesis that is consistent with the $-$ve training example has to specify the value of at least one attribute. For example, the hypothesis $\langle Rainy, ?, ?, ?\rangle$ specifies the value of one attribute and misclassifies no example. So, its description length is $1(\log_2 4) + 0(1 + \log_2 1) = 2$.
- The maximally general hypothesis $\langle ?, ?, ?, ?\rangle$, which is inconsistent with the $-$ve training example, specifies no attribute and misclassifies this single example. So, its description length is $0(\log_2 4) + 1(1 + \log_2 1) = 1$ and it is thus selected by MDL.

(c) We assume no '$\emptyset$' symbol in the hypothesis representation.

The MDL algorithm selects the hypothesis that minimizes $x_h \log_2 n + y_h(1 + \log_2 m)$ while the MAP hypothesis minimizes $-\log_2 P(h) - \log_2 P(D|h)$ (page 18 of the "Bayesian Inference" lecture slides).

$$\begin{aligned}
\arg\min_{h\in H} \ x_h \log_2 n + y_h(1 + \log_2 m) &= \arg\min_{h\in H} -\log_2 (1/n)^{x_h}(1/(2m))^{y_h} \\
&= \arg\min_{h\in H} -\log_2 \frac{(1/n)^{x_h}}{(1 + 2/n)^n}\frac{(1/(2m))^{y_h}}{(1 + 1/(2m))^m} \\
&= \arg\min_{h\in H} -\log_2 \frac{(1/n)^{x_h}}{(1 + 2/n)^n} - \log_2 \frac{(1/(2m))^{y_h}}{(1 + 1/(2m))^m} \ .
\end{aligned}$$

To determine the probability distributions for $P(h)$ and $P(D|h)$ such that the MDL algorithm outputs a MAP hypothesis, simply let $P(h) = \dfrac{(1/n)^{x_h}}{(1 + 2/n)^n}$ and $P(D|h) = \dfrac{(1/(2m))^{y_h}}{(1 + 1/(2m))^m}$.

Note that

$$\begin{aligned}
\sum_{h\in H} P(h) &= \frac{1}{(1 + 2/n)^n}\sum_{h\in H}(1/n)^{x_h} \\
&= \frac{1}{(1 + 2/n)^n}\sum_{x_h=0}^{n} C(n, x_h)2^{x_h}(1/n)^{x_h} \\
&= \frac{1}{(1 + 2/n)^n}(1 + 2/n)^n \\
&= 1
\end{aligned}$$

where the third equality is due to Binomial Theorem.

Similarly,

$$\begin{aligned}\sum_D P(D|h) &= \frac{1}{(1+1/(2m))^m}\sum_D (1/(2m))^{y_h} \\ &= \frac{1}{(1+1/(2m))^m}\sum_{y_h=0}^{m} C(m, y_h)(1/(2m))^{y_h} \\ &= \frac{1}{(1+1/(2m))^m}(1+1/(2m))^m \\ &= 1\end{aligned}$$

where the third equality is due to Binomial Theorem.

BL 9 On page 5 of the "Bayesian Inference" lecture slides, it is said that if $P(h) = P(h')$ for any $h, h' \in H$, then $h_{\text{MAP}} = h_{\text{ML}}$. Prove or disprove that if $h_{\text{MAP}} = h_{\text{ML}}$, then $P(h) = P(h')$ for any $h, h' \in H$.
Hints: Consider the concept learning algorithm FIND-S, which outputs a maximally specific consistent hypothesis. Also, suppose that

$$P(D|h) = \begin{cases} 1 & \text{if } h \text{ is consistent with } D, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Solution. Let more specific hypotheses be more probable *a priori*:

$$\forall h, h' \in H \;\; h >_g h' \rightarrow P(h) < P(h') \; . \tag{2}$$

When h is inconsistent with D,

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{0 \times P(h)}{P(D)} = 0 \; .$$

When h is consistent with D,

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{1 \times P(h)}{P(D)} = \frac{P(h)}{P(D)} \; . \tag{3}$$

Then, from (2) and (3),

$$\forall h, h' \in VS_{H,D} \;\; h >_g h' \rightarrow P(h|D) < P(h'|D) \; .$$

Therefore, the maximally specific hypothesis produced by FIND-S is the most probable *a posteriori* and hence a MAP hypothesis. Since such a maximally specific hypothesis is also consistent (see Proposition 2 in "Concept Learning" lecture slides), it is a ML hypothesis, by (1).

One example of such a distribution would be one in which the probability of a hypothesis is proportional to the number of input attributes with specific values, as opposed to '?'.

BL 10 This extra practice question is recycled from my CS3243 Intro. to AI class but won't be discussed during our tutorial session. We will publish the sample solutions for your reference.

Assume that 2% of the population in a country carry a particular virus. A test kit developed by a pharmaceutical firm is able to detect the presence of the virus from a patient's blood sample. The firm claims that the test kit has a high accuracy of detection in terms of the following conditional probabilities obtained from their quality control testing:

P(the kit shows positive | the patient is a carrier) = 0.998
P(the kit shows negative | the patient is not a carrier) = 0.996

(a) Given that a patient is tested to be positive using this kit, what is the posterior belief that he is not a carrier? Give your answer to 3 decimal places.

(b) Suppose that the patient doesn't entirely trust the result offered by the first kit (perhaps because it has expired) and decides to use another test kit. If the patient is again tested to be positive using this second kit, what is the (updated) likelihood that he is not a carrier? You can assume conditional independence between results of different test kits given the patient's state of virus contraction. Give your answer to 4 decimal places.

Solution.

(a) Let X and $\neg X$ represent the test kit shows positive and negative, respectively. Let Y and $\neg Y$ represent the patient is a carrier and not a carrier, respectively. Then,

$$\begin{aligned} P(Y) &= 0.02 \\ \Rightarrow P(\neg Y) &= 1 - P(Y) = 1 - 0.02 = 0.98 \\ P(X|Y) &= 0.998 \\ P(\neg X|\neg Y) &= 0.996 \\ \Rightarrow P(X|\neg Y) &= 1 - P(\neg X|\neg Y) = 1 - 0.996 = 0.004 \end{aligned}$$

Applying Bayes' Rule,

$$\begin{aligned} P(\neg Y|X) &= \frac{P(X|\neg Y)P(\neg Y)}{P(X|\neg Y)P(\neg Y) + P(X|Y)P(Y)} \\ &= \frac{0.004 \times 0.98}{0.004 \times 0.98 + 0.998 \times 0.02} = 0.164. \end{aligned}$$

There is a 16.4% chance that when the test kit shows positive, the patient is not a carrier.

(b) You can certainly use the normal version of Bayes' rule to solve this question. Alternatively, what I will show you below is the "incremental" version of Bayes' rule that utilizes the solution to part (a) and yields the same answer:

$$\begin{aligned} P(\neg Y|X \times 2) &= \frac{P(X|\neg Y)P(\neg Y|X)}{P(X|\neg Y)P(\neg Y|X) + P(X|Y)P(Y|X)} \\ &= \frac{0.004 \times 0.164}{0.004 \times 0.164 + 0.998 \times 0.836} = 0.0008. \end{aligned}$$