School of Computing and Information Systems
The University of Melbourne
COMP30027 MACHINE LEARNING (Semester 1, 2019)

Tutorial sample solutions: Week 3

Given the following dataset:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
TRAINING INSTANCES					
A	s	h	n	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
TEST INSTANCES					
G	o	m	n	T	?
H	?	h	?	F	?

1. Build a probabilistic **model** based around the given training instances:
   - Okay, we're thinking about training **instances** — there are 6 of them.

   (a) Calculate the **prior** probability $P(\mathit{Outl} = \texttt{s})$. Calculate the prior probabilities of the other attribute values in this data.
   - Here, we're asking "in what proportion of the instances is this true?" (This is known as **maximum likelihood estimation** of the probability(ies).)
   - In the case of $P(\mathit{Outl} = \texttt{s})$, we observe that, of the 6 instances, 2 of them have the attribute value s for *Outl*. Consequently, the prior probability of $P(\mathit{Outl} = \texttt{s})$ is $\frac{2}{6}$.
   - For the value F of *Wind*, it occurs 4 times in the 6 instances, so $P(\mathit{Wind} = \texttt{F})$ is $\frac{4}{6}$.
   - And so on.

   (b) Find the **entropy** of (the distribution of the attribute values) for each of the six attributes, given this probabilistic model.
   - Entropy (in bits) is calculated as follows:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

   - Here, we are going to sum over each attribute value. For *Outl*, this will be s, o, and r:

$$\begin{aligned} H(\mathit{Outl}) &= -[p(\texttt{s})\log_2 p(\texttt{s}) + p(\texttt{o})\log_2 p(\texttt{o}) + p(\texttt{r})\log_2 p(\texttt{r})] \\ &= -[\frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{3}{6}\log_2\frac{3}{6}] \\ &\approx -[0.5(-1) + (0.1667)(-2.585) + (0.3333)(-1.585)] \approx 1.46 \text{ bits} \end{aligned}$$

   (c) Calculate the **joint** probability $P(\mathit{Outl} = \texttt{s} \cap \mathit{Temp} = \texttt{h})$. Calculate some other joint probabilities, for pairs of attribute values from different attributes.
   - For a joint probability, we require multiple things to happen at the same time. In this case, we are asking: "what proportion of instances has both of these as true?"
   - This can be determined by simply counting the number of instances that have the attribute value s for *Outl*, **and** h for *Temp*: there are 2 (A and B), so the required proability $P(\mathit{Outl} = \texttt{s} \cap \mathit{Temp} = \texttt{h})$ is $\frac{2}{6}$.

(d) Calculate the **conditional** probability $P(Outl = \texttt{s}|Temp = \texttt{h})$. Calculate some other conditional probabilities.

- In the case of conditional probabilities, we are asking: "what proportion of instances is the left-hand event true **given** that the right-hand condition is true?"
- Here, we are given *Temp* is h — there are 3 such instances (A, B, and C). Of just these three instances, how many have *Outl* as s? 2 (A and B), so the required probability is $\frac{2}{3}$.
- This can also be derived through the equivalence relationship: $PA|B = \frac{P(A \cap B)}{P(B)}$ — the numerator is the joint count, and the denominator is the prior count of the condition.

2. Ensure that you can derive the **Naive Bayes** formulation.

3. Using the probabilistic model that you developed above, classify the test instances according to the method of **Naive Bayes**.

(a) Using the "epsilon" smoothing method.

- To build a Naive Bayes classifier, we are going to need to calculate all of the prior probabilities of the classes, and all of the conditional probabilities of an attribute (value) given a class.
- The class priors are straight-forward — as with the examples above — we can see that 3 of the 6 instances are Y, and 3 of the 6 instances are N, so $P(\texttt{Y}) = \frac{3}{6}$, and the same for $P(\texttt{N})$.
- For the conditional probabilities, we simply read off the instances of that class, and count the proportion which had that attribute value. For example, $P(Outl{=}\texttt{s}|\texttt{N}) = \frac{2}{3}$; $P(Outl{=}\texttt{o}|\texttt{Y}) = \frac{1}{3}$; and so on.
- Note that we're going to ignore the *ID* attribute below, although it wouldn't change the output of the model if we accounted for it. (You might like to think about why.)
- We classify a test instance T by calculating "probabilities" for each class, as follows: (They aren't truly probabilities.)

$$\text{Score of } c \quad : \quad P(c) \prod_{a \in T} P(a|c)$$

- For instance G, we find the following:

$$\begin{aligned}
\texttt{N} \quad &: \quad P(\texttt{N})P(Outl{=}\texttt{o}|\texttt{N})P(Temp{=}\texttt{m}|\texttt{N})P(Humi{=}\texttt{n}|\texttt{N})P(Wind{=}\texttt{T}|\texttt{N}) \\
&= \quad \frac{3}{6} \times \frac{0}{3} \times \frac{0}{3} \times \frac{2}{3} \times \frac{2}{3} \\
\texttt{Y} \quad &: \quad P(\texttt{Y})P(Outl{=}\texttt{o}|\texttt{Y})P(Temp{=}\texttt{c}|\texttt{Y})P(Humi{=}\texttt{n}|\texttt{Y})P(Wind{=}\texttt{T}|\texttt{Y}) \\
&= \quad \frac{3}{6} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{0}{3}
\end{aligned}$$

- If we were to evaluate, we would find that both values are 0; in fact, as long as there is a single 0, none of the other probabilities in the product matter. This is undesirable behaviour, so we apply a "smoothing" method. In this case, we will simply replace the 0 values with a small positive constant (like 10^{-6}), that we call ϵ:

$$\begin{aligned}
\texttt{N} \quad &: \quad P(\texttt{N})P(Outl{=}\texttt{o}|\texttt{N})P(Temp{=}\texttt{c}|\texttt{N})P(Humi{=}\texttt{n}|\texttt{N})P(Wind{=}\texttt{T}|\texttt{N}) \\
\texttt{N} \quad &: \quad \frac{3}{6} \times \epsilon \times \epsilon \times \frac{2}{3} \times \frac{2}{3} = \frac{2\epsilon^2}{9} \\
\texttt{Y} \quad &: \quad \frac{3}{6} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \epsilon = \frac{\epsilon}{54}
\end{aligned}$$

- By smoothing, we can sensibly the compare the values. Bexcause of the convention of ϵ being very small (i.e. less than $\frac{1}{12}$), Y has the greater score, so G is classified as Y.

- For H, we first observe that the attribute values for *Outl* and *Humi* are missing (?). In Naive Bayes, this just means that we calculate the product without those attributes:

$$\begin{aligned}
\text{N} &: P(\text{N})P(\textit{Outl}{=}\texttt{?}|\text{N})P(\textit{Temp}{=}\texttt{h}|\text{N})P(\textit{Humi}{=}\texttt{?}|\text{N})P(\textit{Wind}{=}\texttt{F}|\text{N}) \\
&\approx P(\text{N})P(\textit{Temp}{=}\texttt{h}|\text{N})P(\textit{Wind}{=}\texttt{F}|\text{N}) \\
&: \frac{3}{6} \times \frac{2}{3} \times \frac{1}{3} = \frac{1}{9} \\
\text{Y} &: \frac{3}{6} \times \frac{1}{3} \times \frac{3}{3} = \frac{1}{6}
\end{aligned}$$

- H is therefore classified as Y.
- A quick note on the epsilons: this isn't a serious smoothing method, but does allow us to sensibly deal with two common cases:
  - Where two classes have the same number of 0s in the product, we essentially ignore the 0s.
  - Where one class has fewer 0s, that class is preferred.

(b) Using "Laplace" smoothing.

- This is similar, but rather than simply changing the probabilities that we have estimated to be equal to 0, we are going to modify the way in which we estimate a conditional probability:

$$\hat{P}_{\mathcal{L}}(a|c) = \frac{1 + \text{freq}(a, c)}{|V| + \text{freq}(c)}$$

- For example, the attribute Outl can take 3 different values. Above, we estimated $P(\textit{Outl}{=}\texttt{o}|\text{Y}) = \frac{1}{3}$; here, we add 1 to the numerator, and 3 to the denominator, to instead have an estimate of $\frac{1+1}{3+3} = \frac{2}{6}$.
- Typically, we would apply this smoothing process when building the model, and then substitute in the Laplace-smoothed values when making the predictions. For brevity, though, I'll make the smoothing corrections in the prediction step. For G, this will look like:

$$\begin{aligned}
\text{N} &: P(\text{N})P(\textit{Outl}{=}\texttt{o}|\text{N})P(\textit{Temp}{=}\texttt{m}|\text{N})P(\textit{Humi}{=}\texttt{n}|\text{N})P(\textit{Wind}{=}\texttt{T}|\text{N}) \\
&= \frac{3}{6} \times \frac{0+1}{3+3} \times \frac{0+1}{3+3} \times \frac{1+2}{2+3} \times \frac{1+2}{2+3} \\
&= \frac{3}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} \times \frac{3}{5} = 0.005 \\
\text{Y} &: P(\text{Y})P(\textit{Outl}{=}\texttt{o}|\text{Y})P(\textit{Temp}{=}\texttt{m}|\text{Y})P(\textit{Humi}{=}\texttt{n}|\text{Y})P(\textit{Wind}{=}\texttt{T}|\text{Y}) \\
&= \frac{3}{6} \times \frac{1+1}{3+3} \times \frac{1+1}{3+3} \times \frac{1+1}{2+3} \times \frac{1+0}{2+3} \\
&= \frac{3}{6} \times \frac{2}{6} \times \frac{2}{6} \times \frac{2}{5} \times \frac{1}{5} \approx 0.0044
\end{aligned}$$

- Unike with the epsilon procedure, N has the greater score — even though there are two attribute values that have never occurred with N.
- For H:

$$\begin{aligned}
\text{N} &: \frac{3}{6} \times \frac{1+2}{3+3} \times \frac{1+1}{2+3} = 0.01 \\
\text{Y} &: \frac{3}{6} \times \frac{1+1}{3+3} \times \frac{1+3}{3+3} \approx 0.013
\end{aligned}$$

- Here, Y has a higher score — which is the same as with the other method, which doesn't do any smoothing here — but this time it is only slightly higher.