School of Computing and Information Systems
The University of Melbourne
COMP30027 Machine Learning (Semester 1, 2021)

Week 3: Sample Solution

1. Approximately 1% of women aged between 40 and 50 have breast cancer. 80% of mammogram screening tests detect breast cancer when it is there. 90% of mammograms DO NOT show breast cancer when it's **NOT** there[1]. Based on these information complete the following table.

Cancer	Probability
No	99%
Yes	1%

Cancer	Test	Probability
Yes	Positive	80%
Yes	Negative	?
No	Positive	?
No	Negative	90%

Based on the probability rule of sum for mutually exclusive events (events that cannot both happen at the same time), we know that the sum of positive and negative test results should sum up to 1 (or 100%).

Therefore, when we have a patient with cancer (`Cancer` = '`Yes`'), and we know that there is 80% probability that the test detects it (`Test` returns '`Positive`'), it means that there is 20% chance (1 – 0.80 = 0.20) that the test does not detect the cancer (`Test` returns '`Negative`' results). We call this a **False Negative** (wrong negative), you will learn more about it later in lectures.

Similarly, when a patient does not have cancer (`Cancer` = '`No`'), and we have that there is 90% chance that the test proves that (`Test` returns '`Negative`'), it means that there is 9% chance (1 – 0.9 = 0.1) that the test detects cancer (returns '`positive`' results) when it is actually not there! We call this a **False Positive** (wrong positive), and again you will learn more about it later in lectures when we talk about evaluations.

So the filled table would be as follow:

Cancer	Test	Probability
Yes	Positive	80%
Yes	Negative	20%
No	Positive	10%
No	Negative	90%

2. Based on the results in question 1, calculate the **marginal probability** of 'positive' results in a Mammogram Screening Test.

According to the law of total probability, we know that

$$P(A) = \sum_{n} P(A|B_n)\, P(B_n)$$

[1] Remember these numbers are not accurate and simplified to ease the calculations in this question.

So, to calculate the probability of 'positive' result for Test, we will have:

$$P(Test = 'positive') = P(Test = 'positive'|Cancer = 'no').P(Cancer = 'no) + P(Test = 'positive'|Cancer = 'yes').P(Cancer = 'yes')$$

Based on the question definition, we know that the chance of having a breast cancer (for females aged between 40 and 50) is 1% . So `P(Cancer = 'yes') = 0.01` and `P(Cancer = 'no') = 0.99`.

From question 1, we know that the probability of a `positive` test result is 80% for a patient with cancer (`P(Test = 'positive'|Cancer= 'yes')=0.8`) and the probability of a `positive` test result is 10% for a patient with no cancer (`P(Test='positive'|Cancer='yes')=0.1`).

So, we have:

$$P(Test = 'positive') = 0.1 \times 0.99 + 0.8 \times 0.01 = 0.107$$

We can show all these in the **Joint Probability Distribution** table as follow.

		Test		Total
		Positive	Negative	
Cancer	Yes	0.01 x 0.8 = 0.008	0.01 x 0.2 = 0.002	0.01
	No	0.99 x 0.1 = 0.099	0.99 x 0.9 = 0.891	0.99
Total		0.107	0.893	1

We call the totals (row and column) the **Marginal Probability**, because they are in the margin!

3. Based on the results in question 1, calculate P(Cancer | Positive), using the Bayes Rule.

According to the Bayesian Rule we know that we can calculate the probability that a person actually has a breast cancer given that her mammography test results return positive, using the following formula:

$$P(Cancer = 'yes'|Test = 'positive') = \frac{P(Test = 'positive'|Cancer = 'yes').P(Cancer = 'yes')}{P(Test = 'positive')}$$

Based on the given information in the question text, we now that "80% of mammogram screening tests detect breast cancer when it is there", so `P(Test = 'positive' | Cancer = 'yes')` is `0.8 (80%)`.

Also, there `1%` chance of having breast cancer (for females aged between 40 and 50). So `P(Cancer = 'yes') = 0.01`.

Also from Question2, we have the `P(Test = 'positive') = 0.107` (the expectation of 'positive' results for a mammogram test).

So we can easily calculate the `P(Cancer= 'yes' | Test = 'positive')`:

$$P(Cancer = 'yes'|Test = 'positive') = \frac{0.8 \times 0.01}{0.107} \cong 0.075 = 7.5\%$$

This result shows that even if a mammography test results returns positive, there is only 7.5% chance that the person actually has Cancer! ☺

4. 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	?

(i). Explain which of the given instances are 'test' instances and which ones are the 'train' instances.

Based on the definition, the train instance in a supervised Learning model are the ones that have a 'label' (belong to a certain Class). So the instances A,B,C,D,E and F are our train instances (or observations) in this dataset. And the instances G and H that do not have a label are our test instances.

Usually, the goal of a Supervised Learning Model is to be able to predict correct labels for the test instances.

(ii). Build a Naïve Bayes model for the given training instances.

A Naïve Bayes model is probabilistic classification Model. All we need for building a Naive Bayes model is to calculate the right probabilities (Prior and Conditional).

For this dataset, our class (or label or variable we trying to predict) is `PLAY`. So, we need the probability of each label (the prior probabilities):

$P(Play = Y) = \frac{1}{2}$ $P(Play = N) = \frac{1}{2}$

We also need to identify all the conditional probabilities between the labels of class (`PLAY`) and all the other attribute values such as `s`, `o`, `r` (for `Outlook`) or `h`, `m`, `c` (for `Temp`) and so on:

$P(Outl = s \mid N) = \frac{2}{3}$ $P(Outl = o \mid N) = 0$ $P(Outl = r \mid N) = \frac{1}{3}$

$P(Outl = s \mid Y) = 0$ $P(Outl = o \mid Y) = \frac{1}{3}$ $P(Outl = r \mid Y) = \frac{2}{3}$

$P(Temp = h \mid N) = \frac{2}{3}$ $P(Temp = m \mid N) = 0$ $P(Temp = c \mid N) = \frac{1}{3}$

$P(Temp = h \mid Y) = \frac{1}{3}$ $P(Temp = m \mid Y) = \frac{1}{3}$ $P(Temp = c \mid Y) = \frac{1}{3}$

$P(Humi = n \mid N) = \frac{2}{3}$ $P(Humi = h \mid N) = \frac{1}{3}$

$P(Humi = n \mid Y) = \frac{1}{3}$ $P(Humi = h \mid Y) = \frac{2}{3}$

$P(Wind = T \mid N) = \frac{2}{3}$ $P(Wind = F \mid N) = \frac{1}{3}$

$P(Wind = T \mid Y) = 0$ $P(Wind = F \mid Y) = 1$

5. Using the Naïve Bayes model that you developed in question 4, classify the given test instances.

(i). No smoothing.

For instance, **G** we have the following:

N: $P(N) \times P(Outl = o \mid N)\, P(Temp = m \mid N) P(Humi = n \mid N)\, P(Wind = T \mid N)$

$$= \frac{1}{2} \times 0 \times 0 \times \frac{2}{3} \times \frac{2}{3} = 0$$

Y: $P(Y) \times P(Outl = o \mid Y)\, P(Temp = m \mid Y) P(Humi = n \mid Y)\, P(Wind = T \mid Y)$

$$= \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times 0 = 0$$

To find the label we need to compare the results for the two tested labels (`Y` and `N`) and find the one that has a higher likelihood (Maximum Likelihood Estimation).

$$\hat{y} = \operatorname{argmax}_{y \in \{Y,N\}} P(y|T = G)$$

However, based on these calculations we find that both values are 0! So, our model is unable to predict any label for test instance G.

The fact is as long as there is a single 0 in our probabilities, none of the other probabilities in the product really matter.

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:

N: $P(N) \times P(Outl = ? \mid N)\, P(Temp = h \mid N) P(Humi = ? \mid N)\, P(Wind = F \mid N)$

$$\approx P(N) \times P(Temp = h \mid N) P\, P(Wind = F \mid N)$$

$$= \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} = \frac{1}{9}$$

Y: $P(Y) \times P(Outl = ? \mid Y)\, P(Temp = h \mid Y) P(Humi = ? \mid Y)\, P(Wind = F \mid Y)$

$$\approx P(N) \times P(Temp = h \mid Y) P\, P(Wind = F \mid Y)$$

$$= \frac{1}{2} \times \frac{1}{3} \times 1 = \frac{1}{6}$$

Therefore, the result of our argmax function for the test instance **H** is **Y**.

$$\operatorname{argmax}_{y \in \{Y,N\}} P(y|T = H) = Y$$

(ii). Using the "epsilon" smoothing method.

For test instance G, using the 'epsilon' smoothing method, we can simply replace the 0 values with a small positive constant (like 10^{-6}), that we call ε. So, we'll have:

N: $$= \frac{1}{2} \times \varepsilon \times \varepsilon \times \frac{2}{3} \times \frac{2}{3} = \frac{2\varepsilon^2}{9}$$

$$Y: \qquad = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \varepsilon = \frac{\varepsilon}{54}$$

By smoothing, we can sensibly compare the values. Because of the convention of ε being very small (it should be less than $\frac{1}{12}$ (why?)), Y has the greater score (higher likelihood). So Y is the output of our **argmax** function and **G is 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.

For **H**, we don't have any zero probability, so the calculations are similar to when we had no smoothing:

$$N: \qquad P(N) \times P(Temp = h \mid N)\, P(Wind = F \mid N)$$

$$\approx P(N) \times P(Temp = h \mid N) P\, P(Wind = F \mid N)$$

$$= \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} = \frac{1}{9} \cong 0.1$$

$$Y: \qquad P(Y) \times P(Temp = h \mid Y)\, P(Wind = F \mid Y)$$

$$\approx P(N) \times P(Temp = h \mid Y) P\, P(Wind = F \mid Y)$$

$$= \frac{1}{2} \times \frac{1}{3} \times \frac{3}{3} = \frac{1}{6} \cong 0.16$$

Therefore, the result of our argmax function for the test instance **H** is **Y**.

$$argmax_{y \in \{Y,N\}}\, P(y|T = H) = Y$$

(iii). Using "Laplace" smoothing ($\alpha = 1$)

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:

$$P_i = \frac{x_i + \alpha}{N + \alpha d}$$

In this method we add α, which is 1 here, to all possible event (seen and unseen) for each attribute. So all unseen event (that currently have the probability of 0) will receive a count of 1 and the count for all seen events will be increased by 1 to ensure that the monocity is maintained.

For example, for the attribute `Outl` that have 3 different values (`s`, `o`, and `r`). Before, we estimated *P(*`Outl` *=*`o`*|*`Y`*)* = $\frac{1}{3}$ before; now, we add 1 to the numerator (add 1 to the count of `o`), and 3 to the denominator (1 (for `o`) + 1 (for `r`)+ 1 (for `s`)). So now *P(*`Outl` *=* `o`*|* `Y`*)* have the estimate of $\frac{1+1}{3+3} = \frac{2}{6}$.

In another example, *P(*`Wind` *=* `T`*|* `Y`*)* is not presented (unseen) in our training dataset (*P(*`Wind` *=* `T`*|* `Y`*)* = $\frac{0}{3}$). Using the Laplace smoothing (α =1), we add 1 to the count of `Wind` = `T` (given `Play` = `Y`) and 1 to the count of `Wind` = `F` (given `Play` = `Y`) and so now we have *P(*`Wind` *=* `T`*|* `Y`*)* = $\frac{0+1}{3+2} = \frac{1}{5}$.

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:

$$N: \quad P(N) \times P(Outl = o \mid N)\, P(Temp = m \mid N) P(Humi = n \mid N)\, P(Wind = T \mid N)$$

$$= \frac{1}{2} \times \frac{0+1}{3+3} \times \frac{0+1}{3+3} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2}$$

$$= \frac{1}{2} \times \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} \times \frac{3}{5} = 0.005$$

$$Y: \quad P(Y) \times P(Outl = o \mid Y)\, P(Temp = m \mid Y) P(Humi = n \mid Y)\, P(Wind = T \mid Y)$$

$$= \frac{1}{2} \times \frac{1+1}{3+3} \times \frac{1+1}{3+3} \times \frac{1+1}{3+2} \times \frac{0+1}{3+2}$$

$$= \frac{1}{2} \times \frac{2}{6} \times \frac{2}{6} \times \frac{2}{5} \times \frac{1}{5} \cong 0.0044$$

Unlike with the epsilon procedure, N has the greater score — even though there are two attribute values that have never occurred with N.

For H:

$$N: \quad = \frac{1}{2} \times \frac{2+1}{3+3} \times \frac{1+1}{3+2} = 0.1$$

$$Y: \quad = \frac{1}{2} \times \frac{1+1}{3+3} \times \frac{3+1}{3+2} \cong 0.13$$

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.