School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 2, 2020)

Sample Solutions: Week 3

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 is **NOT** there[1]. Based on this 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 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.

Based on 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 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 a **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 = 'Yes' | Test = 'Positive')*, using the Bayes Rule.

Based on the Bayesian Rule, we know that we can calculate the probability that a person actually has 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 know 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 return positive, there is only a 7.5% chance that the person actually has Cancer! ☺

4. What is optimisation? What is a “loss function” ?

In the context of Machine Learning, optimisation means finding the **optimal parameters** of the model that give us the most accurate results (predictions).

To find the best possible results, optimisation usually involves minimising (the error) or maximising (the correct answers). Again, in the context of Machine Learning, most of the optimisation problems are described in terms of **cost** (i.e. error). We want to minimise undesirable outcomes (errors). To do so, we define a function that best describes our *undesirable outcomes* for each model. This function is called a *cost function* or a *loss function*.

5. For the following set of classification problems, design a Naive Bayes classification model. Answer the following questions for each problem: (1) what are the instances, what are the features (and values)? (2) explain which distributions you would choose to model the observations, and (3) explain the significance of the Naive Bayes assumption.

(i). You want to classify a set of images of animals in to 'cats', 'dogs', and 'others'.

(1) Here the images are the instances, and the features are the pixels of the image. Each pixel can have values such as pixel intensity or colour code or shade. The important notice here is that these values (in the context of image processing) are continues.

(2) Since our features are continuous, the Gaussian (or normal) distribution is most appropriate (assuming that our feature values are (roughly) Gaussian distributed). The Gaussian distribution has a bell shape curve and useful features that make the calculations fairly easy.

(3) The Naïve Bayes assumption tells us that given each class (‘cat’, ‘dog’, ‘others’), we treat all features as independent. But the reality is that this assumption is not true at all. In fact clearly the {intensity, colour, ...} of neighbouring pixels depend on one another. However, we can still use Naïve Bayes for developing a model and predicting the labels.

(ii). You want to classify whether each customer will purchase a product, given all the products (s)he has bought previously.

(1) In this problem, each customer can be used as an instance. The features can be the products (or types of products) in the catalogue. The value for these features can be 0 and 1 as an indicator of whether the customer has purchased the product; values = 0/1 (did or did not purchase), or perhaps counts of how many times the customer bought a specific (type of) product.

(2) In this setting, the features are discrete. If we assume count-based features, we define a Multinomial distribution over K dimensions (K=number of products) and values are the counts of purchases of that particular customer of each product; we can use essentially the same approach using binary indicators (leading to the Binomial distribution). The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent Boolean experiments (with the probability of p for success in each experiment).

(3) Here the NB assumptions tell us that given the label (‘purchase’, ‘not purchase’) all previous purchases are treated as independent features. But clearly, this is not the case (e.g., if a customer purchased Game of Thrones seasons 1-5, it should influence the probability of the customer also having purchased Game of Thrones season 6).