COMP9417 - Machine Learning
Tutorial: Probabalistic Classification

Question 1 (Bayes Rule)

Assume that the probability of a certain disease is 0.01. The probability of testing positive given that a person is infected with the disease is 0.95, and the probability of testing positive given that the person is not infected with the disease is 0.05.

(a) Calculate the probability of testing positive.

(b) Calculate the probability of being infected with the disease, given that the test is positive.

(c) Now assume that you test the individual a second time, and the test comes back positive (so two tests, two positives). Assume that conditional on having the disease, the outcomes of the two tests are independent, what is the probability that the individual has the disease? (note, conditional independence in this case means that $P(TT|D) = P(T|D)P(T|D)$, and not $P(TT) = P(T)P(T)$.)

Question 2 (Lecture Review)

In this question, we will review some important ideas from the lecture.

(a) What is probabalistic classification? How does it differ from non-probabalistic classification methods?

(b) What is the Naive Bayes assumption and why do we need it?

(c) Consider the problem from lectures of classifying emails as spam or ham, with training data summarised below: Each row represents an email, and each email is a combination of words taken from the set $\{a, b, c, d, e\}$. We treat the words d, e as stop words - these are words that are not useful for classification purposes, for example, the word 'the' is too common to be useful for classifying documents as spam or ham. We therefore define our vocabulary as $V = \{a, b, c\}$. Note that in this case we have two classes, so $k = 2$, and we will assume a uniform prior, that is:

e_1	b	d	e	b	b	d	e			
e_2	b	c	e	b	b	d	d	e	c	c
e_3	a	d	a	d	e	a	e	e		
e_4	b	a	d	b	e	d	a	b		
e_5	a	b	a	b	a	b	a	e	d	
e_6	a	c	a	c	a	c	a	e	d	
e_7	e	a	e	d	a	e	a			
e_8	d	e	d	e	d					

$$p(c_+) = p(c_-) = \frac{1}{2},$$

where c_+ = spam, c_- = ham. Review the multivariate Bernoulli Naive Bayes set-up and classify the test example: assume we get a new email that we want to classify: $e_\star = abbdebb$

(d) Next, review Smoothing for the multivariate Bernoulli case. Why do we need smoothing What happens to our previous classification under the smoothed multvariate Bernoulli model?

(e) Redo the previous analysis for the Multinomial Naive Bayes model without smoothing. Use the following test email: $e_\star = abbdebbcc$

(f) Repeat the analysis for the smoothed Multinomial Naive Bayes model.

Question 3 (More Naive Bayes Practice)

Refer to the training data in the table below, which shows the number of times each of four words A, B, C and D occurs in each of eight documents, four in the positive class, and four negative.

Document No.	A	B	C	D	Class
1	2	0	4	4	+
2	0	3	3	0	+
3	3	0	0	2	+
4	0	0	2	0	+
5	0	0	0	1	-
6	3	0	0	0	-
7	4	3	0	0	-
8	4	0	0	1	-

Consider a new document $x_\star$ with 1 occurence of A, no occurences of B, no occurence of C, and 2 occurences of D. Recall that if $X = (X_1, X_2, X_3, X_4) \sim \text{Multinomial}(p_1, p_2, p_3, p_4)$, then

$$P(X = (x_1, x_2, x_3, x_4)) = \frac{\left(\sum_{i=1}^{4} x_i\right)!}{\prod_{i=1}^{4} x_i!} \prod_{i=1}^{4} p_i^{x_i}.$$

(a) Construct a Naive Bayes classifier for the problem of predicting whether a document should be classified as positive or negative using the multinomial distribution to model the probability of a word occurring or not in a class. Under your model, what is the value of $P(+|x_\star)$? Do not use smoothing.

(b) Now, apply (add-1) smoothing to your probability estimates. Under the smoothed probabilities, what is $P(-|x_\star)$ under the multinomial model?

(c) Recall that if $X \sim \text{Bernoulli}(p)$ then

$$P(X = x) = p^x(1-p)^{1-x}.$$

Construct a Naive Bayes classifier for the problem of predicting whether a document should be classified as positive or negative using the multivariate Bernoulli distribution to model the probability of a word occurring or not in a class. What is $p(+|x_\star)$ now? Do not use smoothing.

(d) Now, apply (add-1) smoothing to your probability estimates. Under the smoothed probabilities, what is $P(-|x_\star)$ under the multivariate Bernoulli model?