

Section 1. Short Answer Questions

- (10 points) Derive the equation for a simple naive bayes classifier for document classification. State in words the assumptions that were followed while deriving the final form of the equation.
- (10 points) Ascertain the class ( $c$  or  $j$ ) predicted for the Test document: *Chinese Chinese Chinese Tokyo Japan* using a naive Bayes model trained using the data shown in Figure 1.

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	$c$
	2	Chinese Chinese Shanghai	$c$
	3	Chinese Macao	$c$
	4	Tokyo Japan Chinese	$j$

Figure 1:  $c$  = Chinese;  $j$  = Japanese

Please show all your computations in detail to substantiate the final prediction.

- (10 points) Figure 2 shows the training data (D) consisting of labelled tuples associated with an event occurring on a Saturday morning given certain environmental conditions. Based on this data, answer the following questions:

	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	P
4	rain	mild	high	false	P
5	rain	cool	normal	false	P
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	P
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
10	rain	mild	normal	false	P
11	sunny	mild	normal	true	P
12	overcast	mild	high	true	P
13	overcast	hot	normal	false	P
14	rain	mild	high	true	N

Figure 2: Class labels:  $P$  is positive and  $N$  is negative

- (1 points) What are the attributes and values of this dataset?
- (1 point) Compute the entropy (or expected information) of the entire data set.
- (8 points) Estimate the information gain associated with each attribute in the dataset.

Please show all your computations in detail with steps.