Directions: Complete the following problems by hand. You may use a calculator.

1. (5 pts) We are interested in using decision trees to classify animals as mammals or not mammals. Consider the data given below:

	animal	temperature	birth	mammal
1.	human	W	Y	Y
2.	elephant	W	Y	Y
3.	turtle	C	N	N
4.	platypus	W	N	Y
5.	guppy	C	Y	N

W: warm-blooded
C: cold-blooded
birth: Does the animal give live birth?
Y: yes
N: no

Use Hunt's algorithm, classification error rate and the data given above to construct a decision tree for classifying mammals. What is the final classification error rate of your decision tree? Ans: 40%

2. (5 pts) Consider the data given below where columns X and Y are attributes and column C is the target. Use the Naive Bayes classifier to determine the probability $C = +$ given that the attribute values are $X = 1$ and $Y = 1$. You must show your calculations. Ans: 0.25

Note: There are three classes: $+$, $-$ and $*$.

	X	Y	C
1.	1	0	+
2.	0	1	−
3.	1	1	*
4.	1	0	−
5.	0	1	+

L17

2. $P(C=+ \mid X=1, Y=1) = \frac{P(1,1 \mid +)P(+)}{P(1,1)}$

$= \frac{P(X=1 \mid +)P(Y=1 \mid +)P(+)}{\frac{1}{5}}$

$= \frac{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{5}}{\frac{2}{5}} = \frac{1}{4} = 0.25$

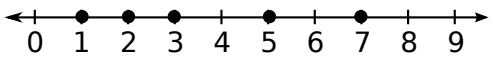
$P(1,1) = P(1,1 \mid +)P(+) + P(1,1 \mid -)P(-) + P(1,1 \mid *)P(*)$

$= P(X=1 \mid +)P(Y=1 \mid +)P(+) + P(X=1 \mid -)P(Y=1 \mid -)P(-) + P(X=1 \mid *)P(Y=1 \mid *)P(*)$

$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{5} + 1 \cdot 1 \cdot \frac{1}{5}$

$= \frac{2}{10} + \frac{2}{10} = \frac{4}{10} = \frac{2}{5}$

3. (5 pts) KMeans clustering is applied to the five points shown below with the number of clusters set equal to 2. The KMeans algorithm has not converged yet and currently has formed two clusters: {1,2} and {3,5,7}.



3. a) $\frac{1+2}{2} = 1.5$ $\frac{3+5+7}{3} = 5$

b) 2 iteration 2 : $\{1,2,3\}$ $\{5,7\}$
2 6

(a) Compute the centroids of the two clusters given above. Ans: 1.5, 5

c) C_1: $(1-2)^2 + (2-2)^2 + (3-2)^2 = 2$

C_2: $(5-6)^2 + (7-6)^2 = 2$

$2+2=4$

(b) Compute the two clusters that the KMeans algorithm will converge to. Hint: It converges very quickly. Ans: $C_1 = \{1,2,3\}$, $C_2 = \{5,7\}$

L22 (c) Compute the total SSE for the final set of clusters the KMeans algorithm converges to. Ans: 4

4. (5 pts) Consider the sentences listed below. Fill in the tables below after removing stop words.

- Sort words in alphabetical order.
- Leave zero values blank.
- Round all numbers in the tables to 1 decimal place.

0: Buy chocolate.
1: Chocolate can melt.
2: Fudge contains chocolate.
3: Buy fudge and chocolate.

L23, 25

$$\text{IDF} = \ln\left(\frac{\text{total number of documents+1}}{\text{number of documents containing term (word)+1}}\right) + 1$$

(a) Word Counts:

words	Buy	chocolate	contains	Fudge	melt
document 0	1	1			
1		1			1
2		1	1	1	
3	1	1		1	

(b) TF (Term Frequency): 出现在document的次数 / document里word总数(Xstopword)

words	Buy	chocolate	contains	Fudge	melt
document 0	0.5	0.5			
1		0.5			0.5
2		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	
3	$\frac{1}{3}$	$\frac{1}{3}$		$\frac{1}{3}$	

(c) IDF (Inverse Document Frequency) $\ln\left(\frac{\#\text{doc}+1}{\#\text{doc contains word}+1}\right)+1$

words	Buy	chocolate	contains	Fudge	melt
doc	2	4	1	2	1
IDF	1.51	1	1.916	1.51	1.916

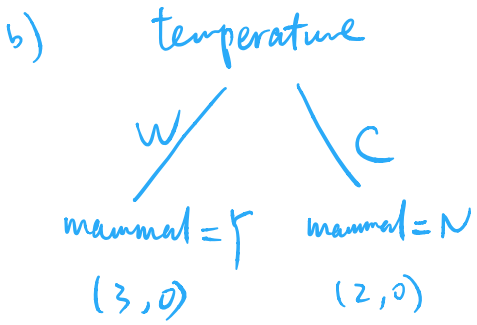
(d) TF-IDF:

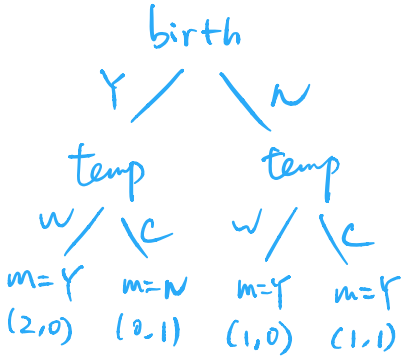
words	buy	chocolate	contains	Fudge	melt
document 0	0.8	0.5			
1		0.5			1
2		0.3	0.6	0.5	
3	0.5	0.3		0.5	

1. a) mammal = Y

(3, 2)

b)





$\frac{1}{5} = 20\%$?