NAME:
UTD ID:

NAÏVE BAYES QUESTIONS

1. Given the following car theft training examples, construct a Naïve Bayes classifier:

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Given that a car is a red, SUV, and of domestic origin, is it more likely to be stolen than not?

a. First use basic naïve Bayes classifier without Laplace smoothing

b. In the second attempt, use Laplace smoothing formula as follows:

$$P(X = x_i | Y = y_j) = \frac{n_c + mp}{n + m}$$

where n_c = number of cases for which Y = y_j and X = x_i
n = number of cases for which Y = y_j
p = apriori estimate for $P(X = x_i | Y = y_j)$ i.e. 1/ number of distinct values of X
m = equivalent sample size, assume 3 for this problem

NAME:
UTD ID:

2. For the playing tennis example discussed in class, find out the most likely class i.e. Yes or No for the following data instance:

Outlook = Sunny; Temperature = mild; Humidity = Normal; Wind= weak

NAME:
UTD ID:

3. A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases. Furthermore, only 0.008 of the entire population has this disease.

1. What is the probability that this patient has cancer?
2. What is the probability that he does not have cancer?
3. What is the diagnosis?

NAME:
UTD ID:

4.

RID	age	income	student	credit	C_i: buy
1	youth	high	no	fair	C_2: no
2	youth	high	no	excellent	C_2: no
3	middle-aged	high	no	fair	C_1: yes
4	senior	medium	no	fair	C_1: yes
5	senior	low	yes	fair	C_1: yes
6	senior	low	yes	excellent	C_2: no
7	middle-aged	low	yes	excellent	C_1: yes
8	youth	medium	no	fair	C_2: no
9	youth	low	yes	fair	C_1: yes
10	senior	medium	yes	fair	C_1: yes
11	youth	medium	yes	excellent	C_1: yes
12	middle-aged	medium	no	excellent	C_1: yes
13	middle-aged	high	yes	fair	C_1: yes
14	senior	medium	no	excellent	C_2: no

The data samples are described by attributes *age*, *income*, *student*, and *credit*. The class label attribute, *buy*, tells whether the person buys a computer, has two distinct values, *yes* (class C_1) and *no* (class C_2)

Find classification for the following example:

$$\mathbf{X} = (age = youth, income = medium, student = yes, credit = fair)$$

NAME:
UTD ID:

5.

Email classification: training data

E-mail	*a*?	*b*?	*c*?	Class
e_1	0	1	0	+
e_2	0	1	1	+
e_3	1	0	0	+
e_4	1	1	0	+
e_5	1	1	0	−
e_6	1	0	1	−
e_7	1	0	0	−
e_8	0	0	0	−

+ denotes spam
- denotes not spam

If you receive an email with feature set (1, 1, 1), which class is it likely to be?

NAME:
UTD ID:

6. Consider the dataset shown below which identifies whether a person evades on their taxes or not.

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Some data about the Taxable Income attribute is given below:
For class=No: sample mean=110 sample variance=2975
For class=Yes: sample mean=90 sample variance=25

Construct a naïve Bayes model and predict whether the result for the following test case:

X = (**Refund** = No, **Married**, **Income** =120K)