CISC-483-683: Assignment 5
Naive Bayes

Due Monday, October 7, 2019
60 points

1. (20 points) Top-Tier is a company that provides temporary employees to other companies. It has the following training data for building a model to classify an applicant as HIRE, REJECT, or MAYBE.

VET	AREA	YEARS	EDUCATION	DECISION
Yes	Tech	5	College	MAYBE
No	Finance	13	College	REJECT
Yes	Clerical	31	Secondary	REJECT
Yes	Finance	33	College	HIRE
Yes	Tech	23	Graduate	HIRE
No	Tech	18	College	MAYBE
Yes	Tech	11	Secondary	REJECT
Yes	Tech	15	College	HIRE
Yes	Tech	15	Graduate	REJECT
Yes	Sales	5	Secondary	REJECT
No	Sales	20	College	HIRE
No	Clerical	7	Secondary	MAYBE

The following computations should be done the same way Weka does to account for the possibility of zero occurrences of an attribute value.

   (a) Compute P(AREA | DECISION) for each value of predictor attribute AREA and each value of the class attribute DECISION.

   (b) Compute P(DECISION) for each value of the class attribute DECISION.

   (c) Compute P(YEARS=27 | DECISION) for each value of the class attribute DECISION.

   (d) Consider the following new test instance:

VET	AREA	YEARS	EDUCATION
Yes	Sales	27	Secondary

   How should this new test instance be classified? Make sure you show your work.

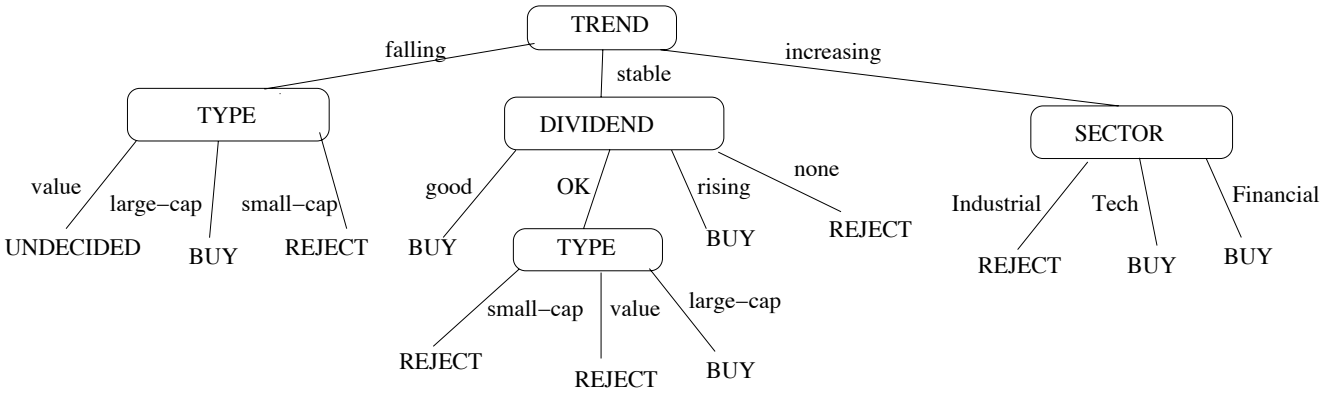
2. (15 points) You will be using dataset **credit-Bayes-19** that can be obtained from the Data-sets tab on the class web site at www.cis.udel.edu/ carberry/CISC-483-683

   One of the files, credit-data-all, is the full dataset; the second file, credit-Bayes-19, is a smaller file with fewer attributes. The attributes in these datasets have been altered to make them anonymous for privacy reasons, since the dataset was compiled by a credit granting agency on actual customers applying for credit. The predictor attributes for the full dataset are A1,...,A15 and the Class attribute is CREDIT which has values + and -. Some predictor attributes are numeric.

   Use NaiveBayesSimple to develop a model for the training data in **credit-Bayes-19**. Classify by hand the following instances; please be sure to show the details of your work. (The Weka results of training on the **credit-Bayes-19** dataset give you much of the information that you need for the calculations.)

b,34,w,v,1.10,t,f
b,45,i,o,0.635,f,t
a,?,i,v,0.100,f,f

3. (25 points) Suppose that you have the following decision tree, where **RECOMMENDATION?** is the class whose value we are trying to predict.



Invest-With-Us is a financial mangagement firm that has constructed a decision tree for recommending mutual funds. Suppose that this decision tree was constructed from the data shown below (and other instances not included in the table but which do not affect your answers to this problem.) Note that there are instances that are missing an attribute value. The class attribute is RECOMMENDATION.

SECTOR	TREND	TYPE	DIVIDEND	RECOMMENDATION
Tech	stable	small-cap		REJECT
Financial	stable	value		BUY
Industrial	stable	value		REJECT
Industrial	stable	small-cap	good	BUY
Tech	stable	small-cap	rising	BUY
Tech	stable	small-cap	rising	UNDECIDED
Tech	stable	large-cap	rising	UNDECIDED
Tech	stable	value	rising	BUY
Industrial	stable	value	rising	REJECT
Tech	stable	value	none	REJECT
Industrial	stable	small-cap	none	REJECT
Tech	stable	small-cap	none	REJECT
Tech	stable	value	none	REJECT
Tech	stable	value	none	REJECT
Industrial	stable	value	none	UNDECIDED
Financial	stable	small-cap	OK	UNDECIDED
Tech	stable	value	OK	REJECT
Industrial	stable	small-cap	OK	REJECT
Tech	stable	large-cap	OK	UNDECIDED
Industrial	stable	large-cap	OK	BUY
Tech	stable	large-cap	OK	BUY
Tech	stable	small-cap	OK	REJECT

(a) What is the entropy of the instances that arrive at the node that is labelled TYPE on the path TREND=stable, DIVIDEND=OK. Please show your work.

(b) The class value on the path TREND=stable, DIVIDEND=rising is BUY. Taking into account the training instances with a missing attribute value, show why BUY is the class value on this path in the decision tree.

(c) **CISC-483**: any two test cases (extra credit for doing all three test cases)
CISC-683: all three test cases Assume the following:
A. 22 training instances have TREND=stable (they are shown in the table on the preceding page)
B. 13 training instances have TREND=increasing
C. 15 training instances have TREND=falling

Suppose that you have the following three test instances:

Test-instance	SECTOR	TREND	TYPE	DIVIDEND
1.	Tech	falling	value	
2.	Industrial		small-cap	rising
3.	Financial	stable	small-cap	

For each test case, determine what class value the decision tree would assign and why — please make sure to show your work and computations.