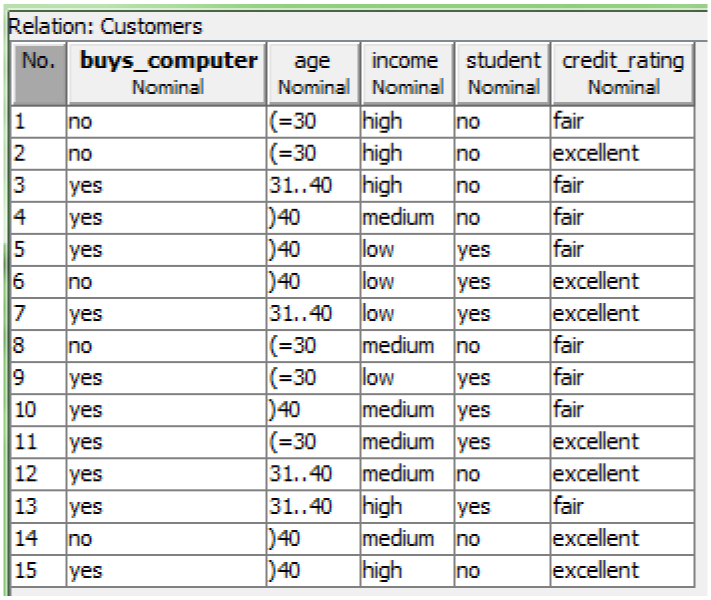
Data Mining, CSCI 347, Fall 2019
Homework 1, Naïve Bayes, due Sept. 23

Consider the following database with the class value being “buys_computer”. (In this image the open parenthesis represents a less-than sign, so the “(=30” value for age representes “<=30”.)



Relation: Customers

No.	buys_computer Nominal	age Nominal	income Nominal	student Nominal	credit_rating Nominal
1	no	(=30	high	no	fair
2	no	(=30	high	no	excellent
3	yes	31..40	high	no	fair
4	yes	)40	medium	no	fair
5	yes	)40	low	yes	fair
6	no	)40	low	yes	excellent
7	yes	31..40	low	yes	excellent
8	no	(=30	medium	no	fair
9	yes	(=30	low	yes	fair
10	yes	)40	medium	yes	fair
11	yes	(=30	medium	yes	excellent
12	yes	31..40	medium	no	excellent
13	yes	31..40	high	yes	fair
14	no	)40	medium	no	excellent
15	yes	)40	high	no	excellent

1. Write the formula to predict if a 25 year old student with a low income and a fair credit rating is likely to purchase a computer using Naïve Bayes Theorem. That is, write formulas for the following.

Pr[buys_computer=’yes’ | age is<=30 & income=’low’ & student=’yes’ & credit_rating=’fair’]

= (Pr[age is’<=30’ | buys_computer=’yes’] *
Pr[income=’low’ | buys_computer=’yes’] *
Pr[student=’yes’ | buys_computer=’yes’] *
Pr[credit_rating=’fair’ | buys_computer=’yes’] *
Pr[buys_computer=’yes’) /
Pr[age is ‘<=30’ & income=’low’& student=’yes’ & credit_rating=’fair’]

Pr[buys_computer=’no’ | age is<=30 & income=’low’ & student=’yes’ & credit_rating=’fair’]

= (Pr[age is<=30 | buys_computer=’no’] *
Pr[income=’low’ | buys_computer=’no’] *
Pr[student=’yes’ | buys_computer=’no’] *
Pr[credit_rating=’fair’ | buys_computer=’no’] *
Pr[buys_computer=’no’]) /
Pr[age is ‘<=30’ & income=’low’& student=’yes’ & credit_rating=’fair’]

2. Since this dataset is small, determine the values to use in the above formulas, counting instance by hand. Apply a Laplace estimator of 1, to avoid probabilities of 0.

age			income			student			credit_rating			buys_computer	
	yes	no		yes	no		yes	no		yes	no	Yes	no
<=30	~~2~~ 3	~~3~~ 4	high	~~3~~ 4	~~2~~ 3	yes	~~6~~ 7	~~1~~ 2	fair	~~6~~ 7	~~2~~ 3	~~10~~ 11	~~5~~ 6
31..40	~~4~~ 5	~~0~~ 1	medium	~~4~~ 5	~~2~~ 3	no	~~4~~ 5	~~4~~ 5	excellent	~~4~~ 5	~~3~~ 4		
>40	~~4~~ 5	~~2~~ 3	low	~~3~~ 4	~~1~~ 2								
<=30	3/13	4/8	high	4/13	3/8	yes	7/12	2/7	fair	7/12	3/7	11/17	6/17
31..40	5/13	1/8	medium	5/13	3/8	no	5/12	5/7	excellent	5/12	4/7		
>40	5/13	3/8	low	4/13	2/8								

3. Calculate the values, ignoring the denominators.

Pr[buys_computer='yes' | age is<=30 & income='low' & student='yes' & credit_rating='fair']

Pr[buys_computer='yes' | E]
= (Pr[age is<=30 | buys_computer='yes'] *
Pr[income='low' | buys_computer='yes'] *
Pr[student='yes' | buys_computer='yes'] *
Pr[credit_rating='fair' | buys_computer='yes'] *
Pr[buys_computer='yes']) / Pr[E]
= (3/13 * 4/13 * 7/12 * 7/12 * 11/17) \ Pr[E]
~ 0.016 / Pr[E]

Pr[buys_computer='no' | age is<=30 & income='low' & student='yes' & credit_rating='fair']

Pr[buys_computer='no' | E]
= (Pr[age is<=30 | buys_computer='no'] *
Pr[income='low' | buys_computer='no'] *
Pr[student='yes' | buys_computer='no'] *
Pr[credit_rating='fair' | buys_computer='no'] *
Pr[buys_computer='no']) / Pr[E]
= (4/8 * 2/8 * 2/7 * 3/7 * 6/17) / Pr[E]
~ 0.005 / Pr[E]

4. Normalize the results.

Pr[buys_computer='yes' | E] = 0.016/(0.016+0.005) = 76%
Pr[buys_computer='no' | E] = 0.005/(0.016+0.005) = 24%

The actual numbers give 74.3% and 25.7%

5. What would be predicted?

Yes would be predicted.

6. Use Weka to check your results.
   a. Go to the classify tab and make “buys_computer” the class attribute (can be set using the drop-down box that appears above the “Start” button)
   b. Run the classifer weka.classifiers.bayes,NaiveBayes on the dataset (you can use cross-validation or use the training set for testing because the classifier model uses the full training set)
   c. Compare Weka’s statistical values with yours.

Weka results match:

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

              Class
Attribute          yes     no
             (0.65) (0.35)
==============================
age
  <=30             3.0    4.0
  31..40           5.0    1.0
  >40              5.0    3.0
  [total]         13.0    8.0

income
  high             4.0    3.0
  medium             5.0    3.0
  low              4.0    2.0
  [total]         13.0    8.0

student
  yes              7.0    2.0
  no               5.0    5.0
  [total]         12.0    7.0

credit_rating
  fair             7.0    3.0
  excellent         5.0    4.0
  [total]         12.0    7.0
```

6. Use Weka to classify the above instance.

buys_computer	**age**	**income**	**student**	**credit_rating**
?	<=30	low	yes	fair

In order to do this do the following:

a. Create a new dataset which is like the original, only it just contains the above instance.

   Your dataset can be:

```
% Test dataset consisting of a single instance

@relation Customers

@attribute buys_computer {yes, no}
@attribute age {<=30, 31..40, >40}
@attribute income {high, medium, low}
@attribute student {yes, no}
@attribute credit_rating {fair excellent}


@data
?,<=30, low, yes, fair
```

b. On the Classify tab, in the Test options area, chose “Supplied test set” and select the new dataset with the single instance.
c. Click the “more options…” button and check the “Output predictions” box.
d. Run the Naïve Bayes classifier to see how Weka would classify this instance.
e. Do the results match what you expected.

   Yes.