Question 1

a. hard and fast rules

Firstly, extract the data from the table 1 in accordance with buy = yes, we get No.1 No.3 No.7 No.8 No.9 No.11 No.12 No.13 No.15 No.16 No.18.

Then we set No.1 as most specific hypothesis:
<garden=yes , good_nbhood=no , parking=no , expensive=yes , interested = no , young=yes , rich = yes>

After check each of the data we got before, finally we may get most specific hypothesis :
<garden=yes , ? , ? , ? , ? , young=yes , ?>

b. naïve Bayes

Firstly, extract the data from the table 1 and calculate the probability of each attributes as the table below:

B=no	**B=yes**
9/20	**11/20**

P(E|B)	**E=no**	**E=yes**
B=no	**5/9**	**4/9**
B=yes	**5/11**	**6/11**

P(I|B)	**I=no**	**I=yes**
B=no	**1/9**	**8/9**
B=yes	**7/11**	**4/11**

P(R|B)	**R=no**	**R=yes**
B=no	**8/9**	**1/9**
B=yes	**8/11**	**3/11**

Then we used Bayes' rule to determine the most likely label of a new example < v1,v2,v3> :

arg max e∈{0,1} P(B = e |E = v1, I = v2,R = v3)
= arg max e∈{0,1} P(E = v1, I = v2,R = v3|E = e) P(B = e)/ P(E = v1, I = v2,R = v3)
= arg max e∈{0,1} P(E = v1, I = v2,R = v3|E = e) P(B = e)
= arg max e∈{0,1} P(E = v1 |B = e) P(I = v2 |B = e)P(R = v3 |B = e)P(E = e)

The maximum likelihood parameter estimate:

When B=yes,

arg max e∈{0,1} P(E = v1 |B = e) P(I = v2 |B = e)P(R = v3 |B = e)P(E = e)
= E=yes, I=no, R=no
<Expensive = yes, Interested=no, Rich=no>

When B=no,

arg max e∈{0,1} P(E = v1 |B = e) P(I = v2 |B = e)P(R = v3 |B = e)P(E = e)
= E=no, I=yes, R=no
<Expensive = no, Interested=yes, Rich=no>

Question 2

Attributes:

Buy - whether the client will buy the house or not
Seven attributes provided by the agency:
Garden - whether the house has a garden or not
Parking - whether the house has private parking or not
Good neighborhood - whether the house is in a good neighborhood or not
Expensive - whether the house is expensive or not
Young - whether the client is young or not
Rich - whether the client is rich or not
Interested -whether the client is interested in the house or not

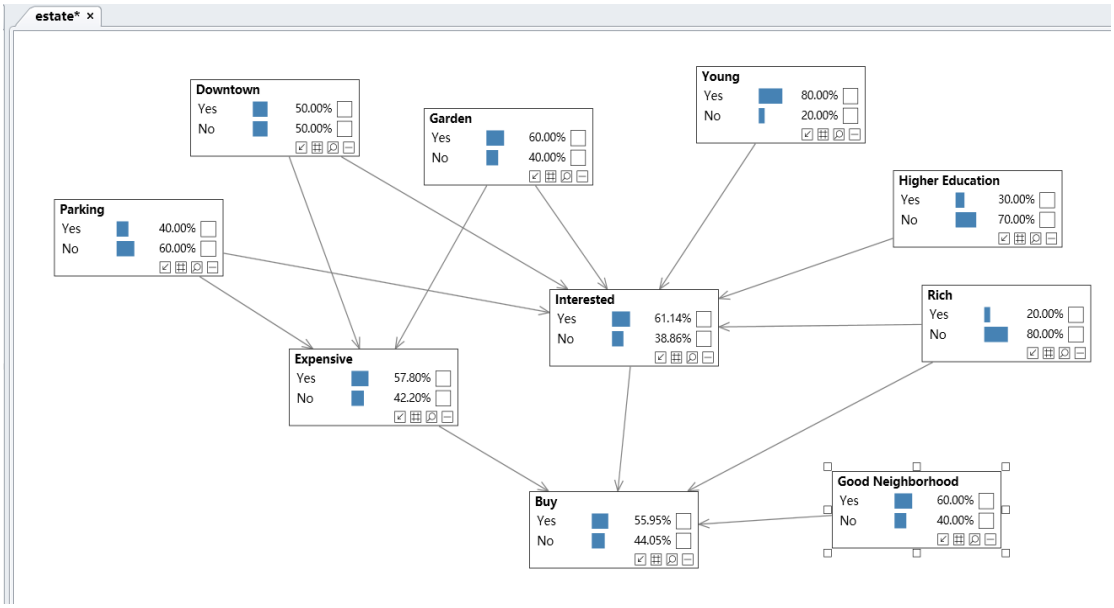
two attributes added:
Downtown - whether the house is at downtown
Whether the house is located in the downtown will lead to different lifestyles, so we think clients will carefully consider about it.
Higher Education - whether the client has received higher education
Customers with different academic qualifications will have different housing tastes, so we think there will be differences in demand for houses.

The distribution data of each note is from my previous knowledge about estate and some experience from question 1, and this Bayesian network will include 9 attributes.

Firstly ,we may define some distribution for each of basic attributes: 'rich' yes: no= 2:8, 'higher education' yes: no = 3:7, 'young' yes: no = 8:2, 'garden' yes: no = 6:4, 'downtown' yes: no = 5:5, 'parking' yes: no = 4:6, 'good neighborhood' yes: no = 6:4

Then, for attribute 'interested', it depends on 'rich', 'young', 'garden' and 'parking', normally rich and higher education client more like the house with garden and parking, and young people has more interested in a same house than elder people, and young people prefer to live in downtown because it's easy for them to go to work. The attribute 'expensive' depends on the basic features of house, like 'Garden', 'Parking'. And finally, for attribute 'buy', it most depends on the attribute 'interested', whether a client is interested in a house greatly influences whether he wants to buy it, and also depends on 'rich', 'young', 'good neighborhood' and 'expensive', and of course the rich client prefer to buy an expensive house, and a good neighborhood also makes people want to buy this house more.

After we apply distributions for each of the notes, we can get a specified Bayesian network as below:



Question 3

a. empirical error of each model

1. hard and fast rules

From table 2, we can see that No.1 No.7 No.8 No.10 cannot get expect answer
Therefore, **empirical error = 4/10 = 40%**

2. naïve Bayes models

From table 2, we can calculate that:

No.1
Predicting the label for <E=yes,I=no,R=no>
For B=no we get :4/9 * 1/9 * 8/9 * 9/20=0.02
For B=yes we get :6/11 * 7/11 * 8/11 * 11/20=0.139
Expect : yes Actually: no error

No.2
Predicting the label for <E=yes,I=yes,R=yes>
For B=no we get :4/9 * 8/9 * 1/9 * 9/20=0.02
For B=yes we get :6/11 * 4/11 * 3/11 * 11/20=0.03
Expect : yes Actually: yes

No.3
Predicting the label for <E=yes,I=no,R=no>
For B=no we get :4/9 * 1/9 * 8/9 * 9/20=0.02
For B=yes we get :6/11 * 7/11 * 8/11 * 11/20=0.139
Expect : yes Actually: no error

No.4
Predicting the label for <E=no,I=no,R=yes>
For B=no we get :5/9 * 1/9 * 1/9 * 9/20=0.003
For B=yes we get :5/11 * 7/11 * 3/11 * 11/20=0.043
Expect : yes Actually: no error

No.5
Predicting the label for <E=no,I=yes,R=no>
For B=no we get :5/9 * 8/9 * 8/9 * 9/20=0.198
For B=yes we get :5/11 * 4/11 * 8/11 * 11/20=0.066
Expect : no Actually: yes error

No.6
Predicting the label for <E=no,I=yes,R=no>
For B=no we get :5/9 * 8/9 * 8/9 * 9/20=0.198
For B=yes we get :5/11 * 4/11 * 8/11 * 11/20=0.066
Expect : no Actually: yes error

No.7
Predicting the label for <E=no,I=yes,R=no>
For B=no we get :5/9 * 8/9 * 8/9 * 9/20=0.198
For B=yes we get :5/11 * 4/11 * 8/11 * 11/20=0.066
Expect : no Actually: yes error

No.8
Predicting the label for <E=yes,I=no,R=no>
For B=no we get :4/9 * 1/9 * 8/9 * 9/20=0.02
For B=yes we get :6/11 * 7/11 * 8/11 * 11/20=0.139
Expect : yes Actually: no error

No.9
Predicting the label for <E=yes,I=yes,R=yes>
For B=no we get :4/9 * 8/9 * 1/9 * 9/20=0.02
For B=yes we get :6/11 * 4/11 * 3/11 * 11/20=0.03
Expect : yes Actually: yes

No.10
Predicting the label for <E=no,I=yes,R=no>
For B=no we get :5/9 * 8/9 * 8/9 * 9/20=0.198
For B=yes we get :5/11 * 4/11 * 8/11 * 11/20=0.066
Expect : no Actually: yes error

From above we can know No.1 No.3 No.4 No.5 No.6 No.7 No.8 No.10 has error therefore, **empirical error = 8/10 = 80%**

3. Bayesian network

After calculate each data from table 2 at the Bayesian network, we can get that:

1. Expect : no Actually: no
2. Expect : yes Actually: yes
3. Expect : no Actually: no
4. Expect : no Actually: no
5. Expect : yes Actually: yes
6. Expect : yes Actually: yes
7. Expect : no Actually: yes error
8. Expect : yes Actually: no error
9. Expect : yes Actually: yes
10. Expect : yes Actually: yes

From above we can know No.7 No.8 has error, therefore, **empirical error = 2/10 = 20%**

b. advantages & disadvantages

1.hard and fast rules

advantages:

- hard and fast rules are very easy to create and use, which don't take much time.
- Very intuitive and easy for people to understand quickly. In our estate case, we could directly see that garden and young are very relevant to whether the client will buy the house or not.

disadvantages:

- not suitable for large amount of data, easy to cause mistakes.

2.naïve Bayes models

advantages:

- Very fast when training and querying large amounts of data
- The logic is quite simple.
- And the model is more stable and robust very low false positive rate.
- Very intuitive and easy for people to understand.

disadvantages:

- Naive Bayes is only suitable for relatively independent datasets, otherwise the effect will be greatly reduced. In this estate model, we may see expensive and rich has some relationship thus this model is not very suitable for our case

3.Bayesian network

advantages:

- Bayesian networks use graphs to describe the interrelationships between data. The semantics are clear and easy for human to understand.
- Bayesian networks are easy to handle incomplete data sets, in our estate case, we don't know the data about new added attributes but it still works.
- Bayesian network is based on the qualitative structure of probability distributions; therefore, it could be used to make efficient inferences and decisions, which is very suitable for our case.

disadvantages:

- need too much work to train the network structure

c. Report

1 Introduction

With the continuous development of artificial intelligence in today's society, machine learning applications are becoming more and more widespread, and more applications are being used in the real estate industry. Here, we will use machine learning technology to make predictions for specific buyers and houses. Through detailed analysis of the three models of hard & fast, naive Bayes and Bayesian networks, the most suitable model will be selected, and make sure the profits of real estate companies get improved finally.

2 Model chosen

2.1 hard & fast

this model may be ready to use because choosing this model can save time and cost, the model can be trained and used as fast as possible, although some errors may occur if the dataset is too large.

2.2 naïve Bayes

This model may not be ready to use because Naive Bayes only can be used if the data sets quite independent, but in this real estate case, some data show strong correlations, so there will be great errors anyway, and the consumption is large either. It could be used if the dataset is not correlations, for example ‘good-neighboorhood’ , ‘young’, and ‘rich’

2.3 Bayesian network

This model is ready to use because the Bayesian network is a model based on mathematical probability. And its rigorous can make the error very small. Through the empirical error computed above, the error of this model is only 20%. Moreover, the intuitiveness of this model makes people easy to understand, and it can be well explained to clients of real estate companies

Through the above argument, we can find that both hard & fast and Bayesian network can be used. However, after careful consideration, we still choose Bayesian network, not only because of its rigorous, but also because of its accuracy.

3 Further Advice

In this practice, we finally selected the Bayesian network as the prediction model. In fact, in the process of constructing the network, we can try to use some more advanced algorithms to improve efficiency, such as feature clustering and feature mapping, which could greatly improve the time efficiency.