CMT311 Coursework Example Solution

Question 1

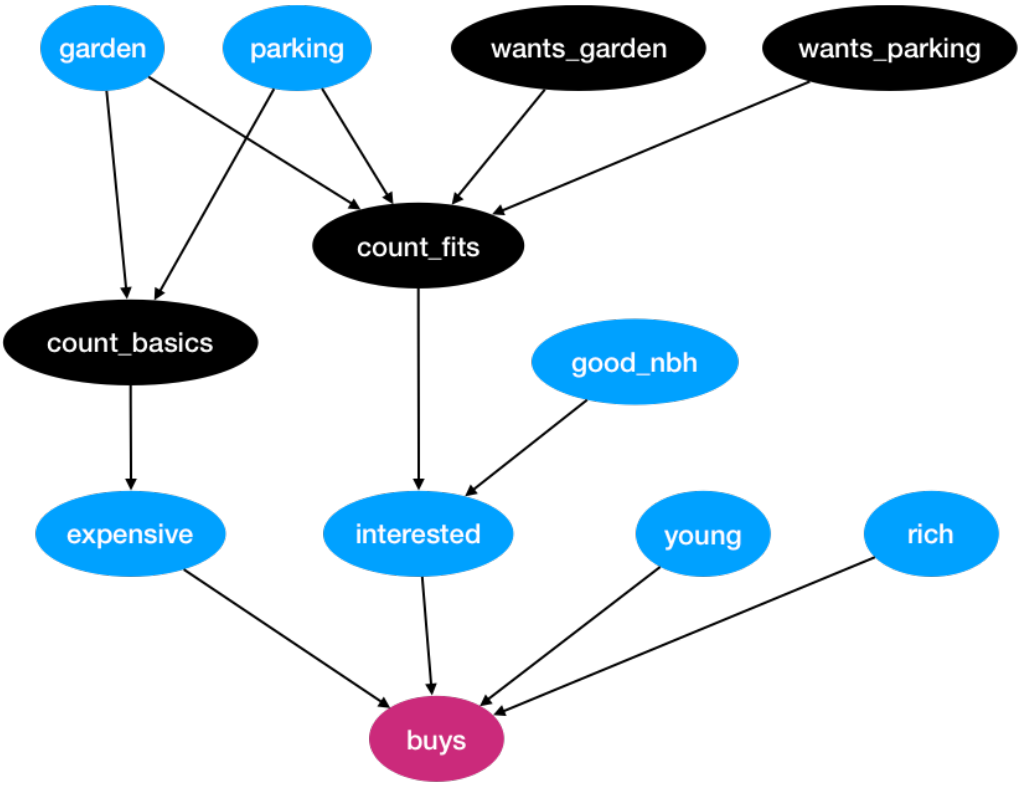
a) The most specific conjunction is “garden=yes & young=yes”. This correctly classifies all positive training examples, and any more general hypothesis in the class (i.e., dropping one or both of the conditions) incorrectly classifies negative training examples.

b) The maximum likelihood parameter estimate is given by the following probabilities:

p(buys=yes) = 11/20
p(expensive=yes | buys=yes) = 6/11
p(expensive=yes | buys=no) = 4/9
p(interested=yes | buys=yes) = 4/11
p(interested=yes | buys=no) = 8/9
p(rich=yes | buys=yes) = 3/11
p(rich=yes | buys=no) = 1/9

These are obtained by counting & computing relative frequencies on the training data (11 of the 20 training examples have buys=yes, 6 of the 11 examples with buys=yes have expensive-yes, etc)

Question 2



P(garden=yes)=13/20
P(garden=no)=7/20

P(parking=yes)=7/20
P(parking=no)=13/20

P(good_nbh=yes)=13/20
P(good_nbh=no)=7/20

P(young=yes)=16/20
P(young=no)=4/20

P(rich=yes)=4/20
P(rich=no)=16/20

P(wants_garden=yes)=0.5
P(wants_garden=no)=0.5

P(wants_parking=yes)=0.5
P(wants_parking=no)=0.5

count_basics = number of basic properties the house has, i.e.,
P(count_basics=2 | garden=yes,parking=yes)=1
P(count_basics=1 | garden=yes, parking=no)=1
P(count_basics=1 | garden=no, parking=yes)=1
P(count_basics=0 | garden=no, parking=no)=1

count_fits = number of basic properties that the customer wants and the house has, i.e.,
P(count_fits=2 | garden=yes,wants_garden=yes,parking=yes,wants_parking=yes) = 1
P(count_fits=1 | garden=yes,wants_garden=no,parking=yes,wants_parking=yes) = 1
P(count_fits=1 | garden=no,wants_garden=yes,parking=yes,wants_parking=yes) = 1
etc

P(buys=yes | rich=yes)=0.9
P(buys=yes | rich=no, interested=no) = 0.01
P(buys=yes | rich=no, interested=yes, expensive=no) = 0.8
P(buys=yes | rich=no, interested=yes, expensive=yes, young=yes) = 0.6
P(buys=yes | rich=no, interested=yes, expensive=yes, young=no) = 0.4

$P(expensive=yes \mid count\_basics=N) = 1-0.7^N$

P(interested=yes | count_fit=0, good_nbh=yes) = 0.1
P(interested=yes | count_fit=0, good_nbh=no) = 0.01
P(interested=yes | count_fit>0, good_nbh=yes) = 0.8
P(interested=yes | count_fit>0, good_nbh=no) = 0.3

We have designed the network structure to reflect what we believe to be the causal influences among the provided attributes (blue nodes) and towards the "buys" attribute the agency wants to predict (pink node). In the absence of more detailed information, we have opted to keep dependencies between the attributes as simple as possible, and we have obtained initial probability estimates for the root attributes from the training data where possible.
We have introduced the following additional attributes (black nodes). The new node count_basics is used to simplify modelling of the dependence of "expensive" on the set of basic features, where this new attribute counts the number of such features, and "expensive" then only depends on the aggregate: the more features a house has, the more likely it is to be expensive. Adding more basic attributes thus simply requires to incorporate these as new parents of the count, without changes to the specification of the probability for "expensive". The advantage is that this keeps the number of parameters to be specified when adding more attributes much more manageable than a direct tabular encoding using all such attributes as parents of "expensive".
Similarly, we introduce for each basic attribute X a new attribute wants_X, which should be easy to collect for future customer-house pairs, and which we initialize to uninformed distributions until such data is available. We use these variables to measure the fit between a house and a customer, again using counting, and use this count together with the good_nbhood attribute with an initial rule-of-thumbs estimate of probabilities.
We have opted to only have expensive, interesting, young and rich directly influence buys, as we believe that the other available attributes are better modelled as indirect influences, and have captured the direct influence in a rule-of-thumbs way as well.
We note that this "common sense" model should not be treated as a final proposal, but as an inspiration for further discussions with the experts in the agency, where we highlight some of the modelling options.

Question 3

a) The hypothesis from **Q1a** predicts examples 1,2,5,6,8,9 positive and the others negative, i.e., gets test examples 1, 7, 8 and 10 wrong, for an empirical error of **4/10**.

   For the hypothesis from Q1b, which uses a subset of attributes, the test data falls into three groups:
   G1: E=no & I=no & R=yes (example 4)
   G2: E=no & I=yes & R=no (examples 5,6,7,10)
   G3: E=yes & I=no & R=no (examples 1,3,8)
   G4: E=yes & I=yes & R=yes (examples 2,9)
   For these groups, the class probabilities are:

	Prob(yes)	Prob(no)	predict
G1	11/20*5/11*7/11*3/11 =105/2420=0.043388	9/20*5/9*1/9*1/9 =5/1620=0.003086	Yes
G2	11/20*5/11*4/11*8/11 =160/2420=0.066116	9/20*5/9*8/9*8/9 =320/1620=0.197531	No

G3	11/20*6/11*7/11*8/11 =336/2420=0.138843	9/20*4/9*1/9*8/9 =32/1620=0.019753	Yes
G4	11/20*6/11*4/11*3/11 =72/2420=0.029752	9/20*4/9*8/9*1/9 =32/1620=0.019753	yes

That is, the **NB hypothesis** makes errors on all test examples except 2 & 9, for an empirical error of **8/10**.

The empirical error of the hypothesis from **Q2 is 1/10**. As all parents of "buys" in our network are observed in the test data, prediction only requires the CPD for "buys". For examples 2, 4 & 9, we have rich=yes, and these are thus more likely (0.9 vs 1-0.9=0.1) predicted as buys=yes. This is correct in two of the three cases, but wrong for example 4.
For examples 1, 3 & 8, we have rich=no & interested=no, where the more likely prediction is buys=no, which is correct for all three.
For the remaining four examples, we have rich=no & interested=yes & expensive=no, which predicts buys=yes, which is again correct for all four.

b) Advantages and disadvantages:
- Conjunctions aren't appropriate in general: there is no single conjunctive reason why people buy houses that applies to every single person and house ("young people buy houses with gardens" doesn't tell us anything about old people or houses without garden).
- While the NB model is somewhat better in capturing variety and uncertainty, it should at least be trained on much more data to get more reliable parameter estimates; but even then, it cannot capture as much detail as the BN (though including more attributes could be a good alternative).
- The BN captures much more detail, but likely needs finetuning of parameters, either based on the expertise of people in the agency or automatically using a large enough training set that the agency would need to collect.

c) None of the models is ready to use in practice. Conjunctions are not appropriate in general, as they cannot handle uncertainty. A probabilistic model would be more appropriate, but both models provided here need to be refined further in collaboration with domain experts and/or based on more training data. For the naïve Bayes model, more training data is needed to get better probability estimates, and we should also consider adding more attributes to better capture the richness of the domain. For the Bayesian network, the rules-of-thumb elements used at the moment need to be refined to better reflect the actual situation observed by the agency, and again, we need better parameter estimates (from more data) and could consider adding more features. Any future developments of the model need to be tested against previously unseen test data, and of course, data collection and usage need to adhere to corresponding legislation.