Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CMT311
Module Title: Principles of Machine Learning
Lecturer: Dr. Angelika Kimmig
Assessment Title: Probabilistic Modelling Case Study
Assessment Number:
Date Set: 29.11.19
Submission Date and Time: 28.02.20 at 9:30am
Return Date: 27.03.20

This assignment is worth 30% of the total marks available for this module. If coursework is submitted late (and where there are no extenuating circumstances):

1 If the assessment is submitted no later than 24 hours after the deadline, the mark for the assessment will be capped at the minimum pass mark;
2 If the assessment is submitted more than 24 hours after the deadline, a mark of 0 will be given for the assessment.

Your submission must include the official Coursework Submission Cover sheet, which can be found here:

https://docs.cs.cf.ac.uk/downloads/coursework/Coversheet.pdf

Submission Instructions

Submit the following two files on learning central:

Description		Type	Name
Cover sheet	**Compulsory**	One PDF (.pdf) file	[student number].pdf
Answers	**Compulsory**	One PDF (.pdf) file	cmt311_[student number].pdf

Any deviation from the submission instructions above (including the number and types of files submitted) will result in a mark of zero for the assessment.

Staff reserve the right to invite students to a meeting to discuss coursework submissions

---

Assignment

This coursework considers the following hypothetical scenario. A real estate agency would like to use artificial intelligence to better predict whether a certain customer will buy a specific house, so they can focus their efforts on promising potential sales. Specifically, they want to label pairs of customers and houses according to whether they belong to the target class *buys* or not. The agency has selected seven attributes, each taking values from {yes,no}, namely

- Basic features of the house:
  - *garden*: whether the house has a garden or not
  - *parking*: whether the house has private parking or not
- *good_nbhood*: whether the house is in a good neighborhood or not
- *expensive*: whether the house is expensive or not
- characteristics of the client:
  - *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

They also have some preliminary ideas about the kind of models they are interested in, and have collected a small dataset for machine learning. Your task is to help them understand their options better and to recommend next steps towards realizing their goal, by answering the following questions.

Question 1

Select two baseline models based on the training data in Table 1 as follows:

a) The first class of hypotheses the agency is interested in are "hard and fast rules" for when a customer buys a house, i.e., conjunctions of attribute tests for attributes they have collected in the training data, such as "*rich*=yes & *parking*=no & *garden*=yes & *young*=no" or "*interested*=yes". Provide the most specific hypothesis in this class that is consistent with the training data in Table 1. Briefly justify your answer.

b) The second class of hypotheses the agency is interested in is the class of naïve Bayes models using the three attributes *expensive*, *interested* and *rich*. Provide the maximum likelihood parameter estimate for such a model given the training data in Table 1. Briefly justify your answer.

Question 2

As an alternative to the two models above, the agency asks you to design a custom Bayesian network for them. Their minimal requirements for this are as follows:

- The Bayesian network must use all seven attributes provided by the agency.

- You may include as many new attributes as you like, as long as they add valuable information and it is realistic for the agency to have access to them.
- Whether a house is *expensive* has to depend on all basic features the house has, and it should be easy to incorporate additional such features into the model.
- As not all customers are equally likely to be interested in the same house, the Bayesian network must explicitly model what makes **this** customer *interested* in **this** house, i.e., it has to relate information about the house to information about the customer.

Provide a fully specified Bayesian network based on this brief. Justify your design choices and explain how additional features of a house would be included in your model, using at most one page of text.

Question 3

The final task is a recommendation of next steps, based on a critical assessment of the three models considered above. Specifically:

a) Compute the empirical error for each of your models from questions 1a, 1b and 2 on the test data provided in Table 2. Assume that the probabilistic models predict the more likely class and break ties in favor of *buys*=yes.

b) Provide a bullet point list of the relative advantages and disadvantages of the three models.

c) Provide a report of at most one page answering these questions from the agency:
    - For each of the three models:
        - Is the model ready to use in practice?
        - If yes, why?
        - If no, why not, and what can we do to get a model of that type that can be used in practice?
    - If more than one model is ready for use in practice, which one should we use, and why?
    - Do you have any further advice that could help us improve our use of machine learning for this problem?

Table 1:

	garden	good_nbhood	parking	expensive	interested	young	rich	**buys**
1	yes	no	no	yes	no	yes	yes	**yes**
2	no	yes	yes	yes	yes	yes	no	**no**
3	yes	yes	no	no	yes	yes	no	**yes**
4	no	yes	no	yes	yes	yes	no	**no**
5	no	no	no	no	yes	no	no	**no**
6	yes	yes	yes	yes	no	no	no	**no**
7	yes	yes	yes	no	yes	yes	no	**yes**
8	yes	no	no	yes	no	yes	yes	**yes**
9	yes	no	no	no	no	yes	no	**yes**
10	no	yes	no	yes	yes	yes	yes	**no**
11	yes	no	yes	yes	no	yes	yes	**yes**
12	yes	yes	yes	yes	no	yes	no	**yes**
13	yes	yes	no	no	yes	yes	no	**yes**
14	yes	no	yes	no	yes	no	no	**no**
15	yes	yes	no	no	no	yes	no	**yes**
16	yes	yes	no	yes	yes	yes	no	**yes**
17	no	yes	no	no	yes	yes	no	**no**
18	yes	yes	no	yes	no	yes	no	**yes**
19	no	no	no	no	yes	no	no	**no**
20	no	yes	yes	no	yes	yes	no	**no**

Table 2:

	garden	good_nbhood	parking	expensive	interested	young	rich	**buys**
1	yes	yes	no	yes	no	yes	no	**no**
2	yes	no	no	yes	yes	yes	yes	**yes**
3	yes	yes	no	yes	no	no	no	**no**
4	no	yes	no	no	no	yes	yes	**no**
5	yes	no	no	no	yes	yes	no	**yes**
6	yes	yes	no	no	yes	yes	no	**yes**
7	yes	no	yes	no	yes	no	no	**yes**
8	yes	yes	no	yes	no	yes	no	**no**
9	yes	no	yes	yes	yes	yes	yes	**yes**
10	no	yes	no	no	yes	yes	no	**yes**

Learning Outcomes Assessed

L1: Reflect on and explain the role of and need for uncertainty in AI systems.
L2: Assess the key concepts and algorithms underlying probabilistic graphical models.
L3: Apply basic algorithms to toy examples.
L5: Explain how reasoning about uncertainty can be combined with reasoning and learning in relational domains.

Criteria for assessment

Credit will be awarded against the following criteria, globally across questions:

- Correctness of technical answers (Q1, Q3a) [30%]
- Quality and appropriateness of Bayesian network model (Q2) [30%]
- Quality, clarity and conciseness of textual answers (Q2, Q3b, Q3c) [40%]

Indicative levels of attainment:

Distinction (70-100%): excellent understanding of all relevant concepts; no technical mistakes; outstanding Bayesian network; clear and concise explanations and discussions; well rounded assessment of pros and cons; well justified recommendation --- *the agency got the professional advice they were looking for*
Merit (60-69%): good understanding of all relevant concepts; no technical mistakes; Bayesian network satisfies all minimal requirements; explanations and discussions address key points, but are missing detail; appropriate recommendation --- *solid advice for the agency, but some open questions remaining*
Pass (50-59%): sufficient understanding of all relevant concepts; no or only minor technical errors; limited explanations; limited range of critical assessment making recommendation hard to justify --- *a reasonable starting point for further discussion with the agency, but many open questions*
Fail (0-49%): big gaps; major errors; missing or unclear explanations and discussions; inappropriate recommendation --- *the agency did not get what they wanted and will have to ask another expert*

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned on 27.03.20 via learning central.

Feedback from this assignment will be useful for the exam.