Coursework 2 specification for 2022

Data Analytics ECS648U/ ECS784U/ ECS784P
Version 1.31, Revised on 03/04/2022 by Dr Anthony Constantinou.

1. Important Dates

- Release date: Week 9, **Monday 21st March** 2022 at 10:00 AM.
- Submission deadline: Week 13, **Wednesday 20th April** 2022 at 10:00 AM.
- Late submission deadline (penalty applies): Within 7 days after deadline.

General guidelines:

i. Occasionally someone will upload their submission and forget to hit the submit button. When this happens, the submission will be marked as 'Draft'. Make sure you fully complete the submission process.

ii. A penalty will be applied automatically by the system for late submissions.
   a. Lecturers cannot grant extensions and cannot remove the penalty.
   b. Extensions can only be granted through approval of an Extenuating Circumstances (EC) claim. Details can be found on your Student Support page, including how to submit an EC claim along with deadline dates.
   c. Penalties are automatically applied by the system can only be challenged via submission of an EC.
   d. If you submit an EC form, your case will be reviewed by the EC panel. When the panel reaches a decision, they will inform both you and the Module Organiser.
   e. If you miss both the submission deadline and the late submission deadline, you will automatically receive a score of 0.

iii. Submissions via e-mail are not accepted.

iv. For details about submission regulations, please refer to your student handbook.

2. Coursework overview

Coursework 2 involves applying causal structure learning to a data set of your choice. You will have to complete a series of tasks, and then answer a set of questions.

- This coursework is based on the lecture material covered between Weeks 6 and 12, and on the lab material covered between Weeks 9 and 11.

- The coursework must be completed individually.

- Submission should be a single file (Word or PDF) containing your answers to each of the questions. Ensure you clearly indicate which answer corresponds to what question. Data sets and other relevant files are not needed, but do save them in case we ask to have a look at them.

- To complete the coursework, follow the tasks below and answer ALL questions enumerated in Section 3. It is recommended that you read this document in full *before* you start completing Task 1.

- You can start going through Task 1 as early as you want. If you face difficulties, we recommend that you try again *after* you attend the Week 9 lab, scheduled for Friday 25th March.

- You can start working on your answers as early as you want, but keep in mind that you need to attend up to Week's 11 material to gain the knowledge needed to answer all the questions.

TASK 1: Set up and reading

a) Visit http://bayesian-ai.eecs.qmul.ac.uk/bayesys/
b) Download the Bayesys user manual.
c) Set up the NetBeans project by following the steps in Section 1 of the manual.
d) Read Sections 2, 3 and 4 of the manual.
e) *Skip* Section 5.
f) Read Section 6 and repeat the example.
   i. *Skip* subsections 6.3 and 6.4.
g) Read Section 7 and repeat the example.
h) *Skip* Sections 8, 9 and 10.
i) Read Section 11.
   i. *Skip* subsection 11.6.

TASK 2: Determine research area and prepare data set

You are free to choose or collate your own data set. As with Coursework 1, we recommend that you address a data-related problem in your professional field or a field you are interested in. If you are motivated in the subject matter, the project will be more fun for you, and you will likely perform better. For your convenience, we have copied the data sources from Coursework 1 into this document (refer to Section 4).

Data requirements:

- **Number of variables:** The data set must contain a minimum of 8 variables (penalty applies below 8 variables).

  While there is no upper bound restriction on the number of the variables, we strongly recommend using considerably less than 100 variables for the purposes of the coursework (but there is no penalty for using more variables). For example, using around 30 or less variables will make it much easier for you to visualise the causal graph, but will also improve learning speed significantly (some algorithms might take hours to complete when given 100 variables!). You do not need to use a special technique for feature selection – it is up to you to decide which variables to keep. We will *not* be assessing feature selection decisions.

- **Re-use data from CW1:** You are allowed to reuse the data set you have prepared for Coursework 1, as long as: a) you consider that data set to be suitable for causal structure learning (refer to Q1 in Section 3), and b) it contains at least 8 variables.

- **Sample size:** There are no restrictions. If algorithms take a long time to complete, you are also free to use part of the samples to speed up the learning process (you do not need permission from the lecturer to do this).

- **Bayesys repository:** You are *not* allowed to use any of the data sets available in the Bayesys repository for this coursework.

TASK 3: Pre-process your data set for structure learning

- Bayesys assumes the input data are categorical or discrete; e.g., $\{\text{"low"}, \text{"medium"}, \text{"high"}\}$, $\{\text{"yellow"}, \text{"blue"}, \text{"green"}\}$, $\{\text{"} < 10\text{"}, \text{"10-20"}, \text{"20} + \text{"}\}$ etc, rather than a continuous range of numbers. If your data set contains continuous variables, Bayesys will consider each of the variable values as a different category. This will cause problems with model dimensionality, leading to poor accuracy and high runtime (if this is not clear why, refer to the Conditional Probability Tables (CPTs) covered in the lectures).

  To address this issue, you should discretise all continuous variables to reduce the number of states to reasonable levels. For example, a variable with continuous values ranging from 1 to 100 (e.g., $\{\text{"14.34"}, \text{"78.56"}, \text{"89.23"}\}$) can be discretised into categories such as $\{\text{"1to20"}, \text{"21to40"}, \text{"41to60"}, \text{"61to80"}, \text{"81to100"}\}$. Because Coursework 2 is not concerned with data pre-processing, you are free to follow any approach you wish to discretise continuous variables. You could discretise the variables manually as discussed in the above example, or even use K-means, but you will not be rewarded extra marks for choosing a more sophisticated approach. We simply need the input data set to be categorical, and it is up to you how you accomplish this.

- Your data set must not contain missing values/empty cells. If it does, replace ALL empty cells with a new category value called *missing* (or use a different relevant name). This will force the algorithms to consider missing values as an additional state. If missing data follow a pattern, this may or may not help the algorithm to produce a more accurate graph, but this is out of the scope of Coursework 2.

Once you ensure your data set is consistent with what has been stated above, rename your data set to *trainingData.csv* and place it in folder *Input*.

TASK 4: Draw out your knowledge-based graph

1. Use your personal knowledge to produce a knowledge-based causal graph based on the variables you decide to keep in your data set. Remember that this graph is based on *your* knowledge, and it is not necessarily correct or incorrect.

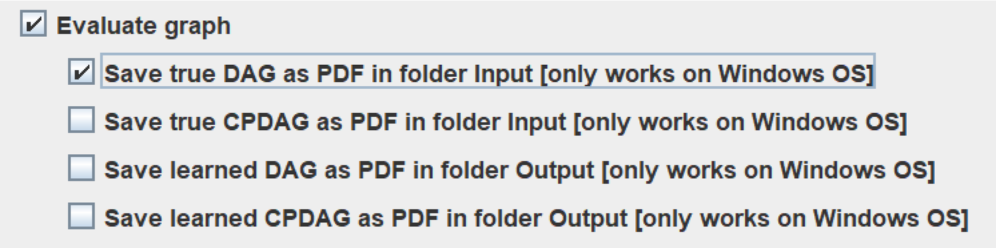
   You may find it easier if you start drawing the graph by hand, and then recording the directed relationships in the *DAGtrue.csv* file. In creating your *DAGtrue.csv* file, we recommend that you edit one of the sample files that come with Bayesys; e.g., you can create a copy of the *DAGtrue_ASIA.csv* file available in the directory *Sample input files/Structure learning*, then rename the file to *DAGtrue.csv*, and then replace the directed relationships with those present in your knowledge graph.

2. Once you are happy with the graph you have prepared, ensure the file is called *DAGtrue.csv* and placed in folder *Input*.

   NOTE: If your OS is not showing the file extensions (e.g., ".csv", ".pdf"), name your file *DAGtrue* and not *DAGtrue.csv*; otherwise, the file might end up being called *DAGtrue.csv.csv* unintentionally (when the file extension is not visible). If this happens, Bayesys will be unable to locate the file.

3. Make a copy of the *DAGtrue.csv* file, and rename this copy into *DAGlearned.csv* and place it in folder *Output*. You can discard the copied file once you complete Task 4.

4. Ensure that your *DAGtrue.csv* and *trainingData.csv* files (as per Task 3) are in folder *Input*, and the *DAGlearned.csv* file is in folder *Output*. Run Bayesys in NetBeans. Under tab *Main*, select *Evaluate graph* and then click on the first subprocess as shown below. Then hit the *Run* button found at the bottom of tab *Main*.

   ☑ Evaluate graph

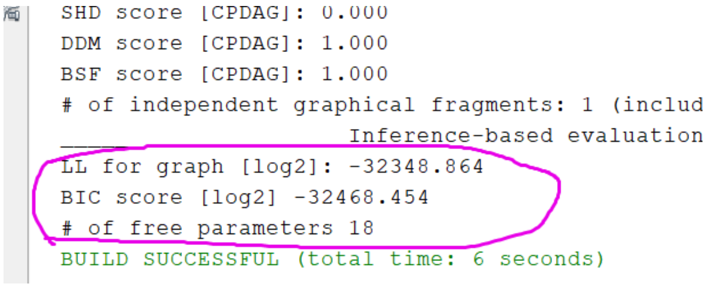
   ☑ Save true DAG as PDF in folder Input [only works on Windows OS]

   □ Save true CPDAG as PDF in folder Input [only works on Windows OS]

   □ Save learned DAG as PDF in folder Output [only works on Windows OS]

   □ Save learned CPDAG as PDF in folder Output [only works on Windows OS]

   The above process will generate output information in the terminal window of NetBeans. Save the last three lines, as highlighted in the Fig below; you will need this information later when answering some of the questions in Section 3.



```
SHD score [CPDAG]: 0.000
DDM score [CPDAG]: 1.000
BSF score [CPDAG]: 1.000
# of independent graphical fragments: 1 (includ
______                        Inference-based evaluation
LL for graph [log2]: -32348.864
BIC score [log2] -32468.454
# of free parameters 18
BUILD SUCCESSFUL (total time: 6 seconds)
```

Additionally, the above process should have generated one PDF files in folder *Input* called *DAGtrue.pdf*. Save this file as you will need it for later.

This only concerns MAC/Linux users: The above process might return an error while creating the PDF file, due to compatibility issues. Even if the system completes the process without errors, the PDF files generated may be corrupted and won`t open on MAC/Linux. If this happens, you should use the online GraphViz editor to produce your graphs, available here: https://edotor.net/ , which converts text into a visual drawing. As an example, copy the code shown below in the web editor:

```
digraph {
    Earthquake -> Alarm
    Burglar -> Alarm
    Alarm -> Call
}
```

You can then edit the above code to be consistent with your *DAGtrue.csv*. You can actually copy-and-paste the variable relationships (e.g., Earthquake → Alarm) directly from *DAGtrue.csv* into the code editor, since they share the same format.

TASK 5: Perform structure learning

1. Run Bayesys. Under tab *Main*, select *Structure learning* and algorithm *HC_CPDAG* (default selection). Select *Evaluate graph* and then click on the last two (out of four) options so that you also generate the learned DAG and CPDAG in PDF files, in addition to the *DAGlearned.csv* file which is generated by default. Then, hit the *Run* button.

2. Once the above process completes, you should see:

   i. Relevant text generated in the terminal window of NetBeans.

   ii. The files *DAGlearned.csv*, *DAGlearned.pdf* and *CPDAGlearned.pdf* generated in folder *Output*. As stated in Task 4, the PDF files may be corrupted on MAC/Linux, and you will have to use the online GraphViz editor to produce the graph corresponding to *DAGlearned.csv* (simply copy the relationships from the CSV file into the editor as discussed in Task 4).

3. Repeat the above process for the other five algorithms; i.e., HC, TABU_DAG, TABU_CPDAG, SaiyanH, and MAHC. Save the same output information and files that each algorithm produces (ensure you first read the NOTE below).

   NOTE: As stated in the manual, Bayesys overwrites the output files every time it runs. You need to remember to either rename or move the output files to another folder ***before*** running the next algorithm.

   Similarly, if you happen to have one of the output files open – for example, viewing the *DAGlearned.pdf* in Adobe Reader while running structure learning - Bayesys will *fail* to replace the PDF file, and the output file will not reflect the latest iteration. Therefore, ensure you close all output files before running structure learning.

3. Questions

This coursework involves applying six different structure learning algorithms to your data set. We do not expect you to have a detailed understanding of how the algorithms operate. None of the Questions focuses on the algorithms and hence, your answers should *not* focus on discussing differences between algorithms.

- You should answer ALL questions.
- You should answer the questions in your own words.
- Do *not* exceed the maximum number of words specified for each question. If a question restricts the answer to, say 100 words, only the first 100 words will be considered when marking the answer.
- Marking is out of 100.

QUESTION 1: Discuss the research area and the data set you have prepared, along with pointers to your data sources. Screen-capture part of your final version of your data set and present it here as a Figure. For example, if your data set contains 15 variables and 1,000 samples, you could present the first 10 columns and a small part of the sample size. Explain why you considered this data set to be suitable for structure learning, and what questions you expect a structure learning algorithm to answer.

Maximum number of words: 150

Marks: 10

QUESTION 2: Present your knowledge-based DAG (i.e., *DAGtrue.pdf* or the corresponding *DAGtrue.csv* graph visualised through the web editor), and briefly describe the information you have considered to produce this graph. For example, did you refer to the literature to obtain the necessary knowledge, or did you consider your own knowledge to be sufficient for this problem? If you referred to the literature to obtain additional information, provide references and very briefly describe the knowledge gained from each paper. If you did not refer to the literature, justify why you considered your own knowledge to be sufficient in determining the knowledge-based graph.

NOTE: It is possible to obtain maximum marks without referring to the literature, as long as you clearly justify why you considered your personal knowledge alone to be sufficient. Any references provided will *not* be counted towards the word limit.

Maximum number of words: 200

Marks: 10

QUESTION 3: Complete Table Q3 below with the results you have obtained by applying each of the algorithms to your data set during Task 5. Compare your CPDAG scores produced by F1, SHD and BSF with the corresponding CPDAG scores shown in Table 2.1 (page 12) in the Bayesys manual.

Specifically, are your scores mostly lower, similar, or higher compared to those shown in Table 2.1 in the manual? Why do you think this is? Is this the result you expected? **Explain why**.

Table Q3. The scores of the six algorithms when applied to your data set.

Algorithm	CPDAG scores			Log-Likelihood (LL) score	BIC score	# free parameters	Structure learning elapsed time
	BSF	SHD	F1				
HC_CPDAG							
HC_DAG							
TABU_CPDAG							
TABU_DAG							
SaiyanH							
MAHC							

Maximum number of words: 250

Marks: 15

QUESTION 4: Present the CPDAG generated by HC_CPDAG (i.e., *CPDAGlearned.pdf* or the corresponding *CPDAGlearned.csv* graph visualised through the web editor). Highlight the three causal classes in the CPDAG. You only need to highlight *one* example for each causal class. If a causal class is not present in the CPDAG, explain why this might be the case.

Maximum number of words: 200

Marks: 10

QUESTION 5: Rank the six algorithms by score, as determined by each of the three metrics specified in Table Q5. Are your rankings consistent with the rankings shown under the column "*Rankings according to the Bayesys manual*" found in Table Q5 below? Is this the result you expected? **Explain why**.

Table Q5. Rankings of the algorithms based on your data set, versus ranking of the algorithms based on the results shown in Table 2.1 in the Bayesys manual.

	Your rankings			Rankings according to the Bayesys manual		
Rank	BSF [single score]	SHD [single score]	F1 [single score]	BSF [average score]	SHD [av. normalised[1] score]	F1 [average score]
1				TABU_CPDAG [0.533]	MAHC [0.481]	SaiyanH [0.576]
2				SaiyanH [0.515]	TABU_CPDAG [0.44]	TABU_CPDAG [0.564]
3				HC_CPDAG [0.506]	SaiyanH [0.438]	MAHC [0.562]
4				MAHC [0.499]	HC_CPDAG [0.402]	HC_CPDAG [0.537]
5				TABU_DAG [0.484]	TABU_DAG [0.397]	TABU_DAG [0.53]
6				HC_DAG [0.438]	HC_DAG [0.314]	HC_DAG [0.479]

Maximum number of words: 200

Marks: 10

QUESTION 6: Refer to your elapsed structure learning runtimes and compare them to the runtimes shown in Table 2.1 in the Bayesys manual. Indicate whether your results are consistent or not with the results shown in Table 2.1. **Explain why**.

Maximum number of words: 100

Marks: 10

[1] The SHD score is normalised between 0 and 1 in Table 2.1. This is because SHD is sensitive to both the number of the nodes and the number of the edges of the true graph, and this means that unnormalised SHD score comparisons between different networks will be biased. You do *not* need to normalise the SHD score in your case, since the comparisons will be based on a single network/same data set.

QUESTION 7: Compare the BIC score, the Log-Likelihood (LL) score, and the number of free parameters generated in Task 4, against the same values produced by the six algorithms in Task 5. What do you understand from the difference between those three scores? Are these the results you expected? **Explain why**.

Table Q7. The BIC scores, Log-Likelihood (LL) scores, and number of free parameters generated by each of the six algorithms during Task 4 and Task 5.

Algorithm	Your Task 4 results			Algorithm	Your Task 5 results		
	BIC score	Log-Likelihood	Free parameters		BIC score	Log-Likelihood	Free parameters
Your knowledge-based graph				HC_CPDAG			
				HC_DAG			
				TABU_CPDAG			
				TABU_DAG			
				SaiyanH			
				MAHC			

Maximum number of words: 200

Marks: 15

QUESTION 8: Select TWO knowledge approaches from those covered in Week 11 Lecture and Lab; i.e., any two of the following: a) Directed, b) Undirected, c) Forbidden, d) Temporal, e) Initial graph, f) Variables are relevant, and g) Target nodes. Apply each of the two approaches to the structure learning process of HC_CPDAG, separately (i.e. only use one knowledge approach at a time). It is up to you to decide how many constraints to specify for each approach. Then, complete Table Q8 and explain the differences in scores produced *before* and *after* incorporating knowledge. Are these the results you expected? **Explain why**.

Remember to clarify which two knowledge approaches you have selected from those listed between (a) and (g) above, and show in a separate/new table the constraints you have specified for each approach. These constraints must come from your knowledge graph you have produced in Task 4. Note that knowledge approach (f) does not require any constraints; but yes, you can still use this as one of your two selections.

Table Q8. The scores of HC_CPDAG applied to your data, with and without knowledge.

Knowledge approach	CPDAG scores			LL	BIC	Free parameters	Number of edges	Runtime
	BSF	SHD	F1					
Without knowledge								
With knowledge (List 1st knowledge approach here)								
With knowledge (List 2nd knowledge approach here)								

Maximum number of words: 300

Marks: 20

4. Data sources

Using public data is the most common choice. If you have access to private data, that is also an option, though you will have to be careful about what results you can release to us. Some sources of publicly available data are listed below (you don`t have to use these sources).

- **Kaggle**
  https://www.kaggle.com/
  Over 50,000 public data sets for machine learning.

- **UK Covid Data**
  https://coronavirus.data.gov.uk/
  Official UK COVID data

- **Data.gov**
  http://data.gov
  This is the resource for most government-related data.

- **Socrata**
  http://www.socrata.com/resources/
  Socrata is a good place to explore government-related data. Furthermore, it provides some visualization tools for exploring data.

- **UN3ta**
  https://data.un.org/
  UN data is an Internet-based data service which brings UN statistical databases.

- **European Union Open Data Portal**
  http://open-data.europa.eu/en/data/
  This site provides a lot of data from European Union institutions.

- **Data.gov.uk**
  http://data.gov.uk/
  This site of the UK Government includes the British National Bibliography: metadata on all UK books and publications since 1950.

- **The CIA World Factbook**
  https://www.cia.gov/library/publications/the-world-factbook/
  This site of the Central Intelligence Agency provides a lot of information on history, population, economy, government, infrastructure, and military of 267 countries.

- **US Census Bureau**
  http://www.census.gov/data.html
  This site provides information about US citizens covering population data, geographic data, and education.

- **Health Data**
  Healthdata.gov
  https://www.healthdata.gov/
  This site provides medical data about epidemiology and population statistics.

- **NHS Health and Social Care Information Centre**
  http://www.hscic.gov.uk/home
  Health datasets from the UK National Health Service.

- **Social Data**
  Facebook Graph
  https://developers.facebook.com/docs/graph-api
  Facebook provides this API which allows you to query the huge amount of information that users are sharing with the world.

- **Topsy**
  http://topsy.com/
  Topsy provides a searchable database of public tweets going back to 2006 as well as several tools to analyze the conversations.

- **Google Trends**
  http://www.google.com/trends/explore
  Statistics on search volume (as a proportion of total search) for any given term, since 2004.

- **Likebutton**
  http://likebutton.com/
  Mines Facebook's public data--globally and from your own network--to give an overview of what people "Like" at the moment.

- **Amazon Web Services public datasets**
  http://aws.amazon.com/datasets
  The public data sets on Amazon Web Services provide a centralized repository of public data sets. An interesting dataset is the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information. Also a NASA database of satellite imagery of Earth is available.

- **DBPedia**
  http://wiki.dbpedia.org
  Wikipedia contains millions of pieces of data, structured and unstructured, on every subject. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.

- **Freebase**
  http://www.freebase.com/
  This community database provides information about several topics, with over 45 million entries.

- **Gapminder**
  http://www.gapminder.org/data/
  This site provides data coming from the World Health Organization and World Bank covering economic, medical, and social statistics from around the world.

- **Google Finance**
  https://www.google.com/finance
  Forty years' worth of stock market data, updated in real time.

- **National Climatic Data Center**
  http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim
  Huge collection of environmental, meteorological, and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.

- **WeatherBase**
  http://www.weatherbase.com/
  This site provides climate averages, forecasts, and current conditions for over 40,000 cities worldwide.

- **Wunderground**
  http://www.wunderground.com/
  This site provides climatic data from satellites and weather stations, allowing you to get all information about the temperature, wind, and other climatic measurements.

- **Football datasets**
  http://www.football-data.co.uk/
  This site provides historical data for football matches around the world.

- **Pro-Football-Reference**
  http://www.pro-football-reference.com/
  This site provides data about football and several other sports.

- **New York Times**
  http://developer.nytimes.com/docs
  Searchable, indexed archive of news articles going back to 1851.

- **Google Books Ngrams**
  http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
  This source searches and analyses the full text of any of the millions of books digitized as part of the Google Books project.

- **Million Song Data Set**
  http://aws.amazon.com/datasets/6468931156960467
  Metadata on over a million songs and pieces of music. Part of Amazon Web Services.