Probabilistic decision support systems: homework

In which, we overview goals of the homework, earlier topics, special domains with workgroup options, the default (mandatory) and the midterm (optional) parts, and requirements, expected content, formats, scoring.

Goals

Causal Bayesian networks (CBNs) have a unique position in artificial intelligence:

- As a probabilistic logic knowledge base, a CBN provides a coherent framework to represent beliefs (see Bayesian interpretation of probabilities).
- As a decision network, it provides a coherent framework to represent preferences for actions.
- As a dependency map, it explicitly represents the system of conditional independencies in a given domain.
- As a causal map, it explicitly represents the system of causal relations in a given domain.
- As a decomposable probabilistic graphical model, it parsimoniously represents the quantitative stochastic dependencies (the joint distribution) of a domain, and it allows efficient observational inference.
- As an uncertain causal model, it parsimoniously represents the quantitative, stochastic, autonomous mechanisms in a domain and it allows efficient interventional and counterfactual inference.

The goal of the homework is to (1) deepen theoretical knowledge, (2) develop engineering know-how, and (3) provide a personal experience about this central, multifaceted element of AI through a complete workflow.

Major steps

Major steps in the workflow of the homework are as follows:

- Default part:
  - Construct a causal Bayesian network.
  - Test it by inference, sensitivity/perturbation/bootstrap analysis.
  - Demonstrate observational, causal, and counterfactual inference.
  - Extend into a decision network and infer optimal actions.
- Midterm part
  - Scale-up and perform these steps using a programming environment.

Domain selection

Try to select a balanced domain, in which you can model both the causes and the consequences of the central part in the domain.

Because of the crisis, two special domains with workgroup option are offered:

- COVID-19 diagnostics
  - early, differential diagnostics at home
  - Kaggle COVID-19 Open Research Dataset Challenge (CORD-19)
- Distance education
  - student characteristics
  - external conditions (home environment)
  - characteristics of education (material, lecturer)
  - outcome measures
    - knowledge transfer
    - scientific interest
    - industrial skills
    - societal aspects

Workgroups ideally have 3-7 participants and in case of distance education they accomplish a comprehensive intelligent data analysis study: study design, data collection from course participants, data engineering, model construction, model refinement using data, model evaluation and (limited) interpretation.

Earlier homework topics

The causes-phenomenon-effects structures are the following in domains of earlier homeworks.

- Biomed: clinical diagnostics
  - risk and protective factors
  - diseases
  - symptoms
- Tech: fault discovery (PC, mobile, software)
  - faults
  - mechanisms
  - observable errors
- Commercial: recommendation of devices or items (laptop, mobile, movies)
  - preliminary context (needs)
  - preferences
  - consumer satisfaction
- Travel: how to travel to the university
  - external factors, obstacles
  - options
  - consequences
- Education: personal performance
  - influencing factors
  - goals
  - outcomes

A detailed description of the default and midterm parts

Default part

- Select a domain and sketch the structure of a Bayesian network model.
- Consult it.
- Quantify your BN model.
- Test it with
  - global inference, i. e., test overarching inferences with distant query and evidence nodes/variables,
  - „information sensitivity of inference” analysis.
- Check it by relearning it from self-generated data.
  - Generate a data set from your model.
  - Learn a model from your data.
  - Compare the structural and parametric differences between the two models
- Demonstrate observational, causal, and counterfactual inference in the model.
- Extend your BN model to a decision network.
- Investigate the value of further information
  - select values for some “evidence” variables (E=e),
  - using BayesCube calculate the current expected loss/utility EU(D|e),
  - select a variable “I” as potential “further” information,
  - using BayesCube calculate the conditional probabilities of potential further observations (i.e. the conditional probabilities of potential values of this “further information” variable, p(I=i|E=e)),
  - using BayesCube calculate the expected losses/utilities corresponding to these potential further observations EU(D|e,i),
  - calculate the (expected) value of (perfect) information corresponding to this variable “I”, Σi p(i|e)*EU(D|e,i)- EU(D|e).

Midterm part

- The suggested software environment is pomegranate, to explore other environments, see the MI Almanach option.
- The expected form of documentation is a notebook (Google colab or Azure notebook). The structure should follow the pomegranate colab notebook syllabus with additional subtasks.
- Write a formal specification for your model with test cases, i. e., expectations about
  - structural aspects
    - causal relations: distant but important cause-effect relations
    - (in)dependency relations: multivariate (ir)relevance statements, e.g., independent causes, Markov boundaries, confounded effects,
    - context-sensitive independencies
  - parametric aspects
    - qualitative expectations
      - about distant aggregates
      - monotonicity relations („signs”) in local models
    - interval estimates
      - credible regions
  - sample size estimates
    - estimation of sample size corresponding to a priori estimates.

- Perform ALL(!) the subtasks in the default homework using pomegranate.
- Perform and document additional steps either in BayesCube or pomegranate
  - Analyse estimation biases in multiple scenarios
    - underconfidence
    - overconfidence
  - Investigate the effect of model uncertainty and sample size on learning
    - vary the strength of dependency in the model (increase underconfidence to decrease information content) and sample size and see their effect on learning.

Documentation

Default part:

The homework should be summarized in a document, whose structure follows and describes the steps of the workflow.

The length is expected around 5-10 pages.

Midterm part:

The expected form of documentation is a notebook (Google colab or Azure notebook). The structure should follow the pomegranate colab notebook syllabus with additional subtasks.

Consultation

The preliminary approval of your planned homework is mandatory!

Submission

After the consultation, the model XML with its documentation, and optionally the notebook for the midterm part, in a single ZIP file should be uploaded to the homework submission site (belated homework can be submitted in the first week after the semester, but please try to accomplish it by the 13th week).

Tools

The software system BayesCube with manual is available at

http://bioinfo.mit.bme.hu/

Syllabus for the pomegranate environment is available at

https://colab.research.google.com/drive/1HO5rmSq35YiN8R39Nu_N8c8OwLbozGnY

http://www.mit.bme.hu/system/files/oktatas/targyak/9892/PDSS_nagyHF_colab_v0_1.pdf

Hints

1. Prefer causality, i.e. temporal direction and mechanisms (easier estimation of conditionals).
2. Do not use variables with more values than 5 (binary variables usually suffice).
3. Do not use aggregate, semantic variables (with semantic relations).
4. Save and version your models.

Scoring

Each subtask will get a mark and their average will be used to compute the final grade.

Reference

Russel-Norvig: Artificial intelligence: a modern approach (2nd edition or above)

- Chapter 14.,16. (optional chapters 13-16, 18-20)