Week 5 Video 2

Relationship Mining

Causal Mining
Causal Data Mining

- These slides developed in partnership with Stephen Fancsali, Carnegie Learning, Inc.
Causal Data Mining

- Distinct from prediction or correlation mining

- The goal is not to figure out what predicts $X$,
- or to figure out what is correlated to $X$,
- but instead…
Causal Data Mining

find causal relationships in data.

- A causes B

Examples from Scheines (2007):

What features of student behavior cause learning?

What will happen when we make everyone take a reading quiz before each class?

What will happen when we program our tutor to intervene to give hints after an error?
Use graphs to represent causal structure

- Frequently directed graphs without cycles
  - (Bayesian networks – see week 4 slides)
  - Nodes represent variables
  - (Directed) edges represent causal relationships
Causal Data Mining

- Algorithms infer (classes of) causal graphs that explain dependencies in observed data
  - From observed data alone, often cannot infer a unique causal graph.
Finding Causal Structure

- Easy to determine if you intervene
  - Some experiments are impossible, too expensive, unethical, etc.

- Can you determine this from purely correlational data?
  - Spirtes, Glymour, and Scheines say: sometimes, yes!
Example

- Is repeatedly retrying quizzes harmful?
  - Does repeatedly retrying quizzes cause decreased learning?

- Suppose an investigator notices that repeatedly retrying quizzes and exam score are negatively associated (i.e., correlated).
A direct causal relationship could explain this correlation…
or the correlation of *retry quiz* and *exam* might arise from a common cause, e.g., prior knowledge.
(or both!)
Causal Graphs

- Suppose that when we control for pre-test, the correlation of retry quiz & exam disappears.
  - E.g., the partial correlation is not significantly different from zero.
Three causal graphs can explain this conditional independence equally well...
Causal Graphs

- but only one is compatible with background knowledge
  - *pre-test* is prior to behavior in a tutor and a final exam.
Big idea

- Infer class of graphs that can represent the full pattern of such (in)dependencies among measured variables.
Causal Data Mining

- TETRAD is a key software package used to study this

- http://www.phil.cmu.edu/projects/tetrad/
TETRAD

- Implements multiple algorithms for inferring causal structure from data
- Different algorithms are applicable given particular assumptions.
Assumptions guide algorithm choice

- Are there unmeasured common causes?
- Linear relationships between variables?
- Are underlying dynamics acyclic or cyclic?
- Distribution of variables: Gaussian vs. non-Gaussian
- See TETRAD User Guide for detailed discussion...
Math & Assumptions

- See


Examples in EDM
This example uses an algorithm that allows for unmeasured common causes of measured variables.

`pretest_score \rightarrow total_steps` can signify

1. `pretest_score` is a cause of `total_steps`;
2. `pretest_score` & `total_steps` share a common cause;
3. both!
Rau & Scheines (2012)
Rai et al. (2011)
Rai et al. (2011)
Rai et al. (2011)
Wait, what?
Use domain knowledge to constrain search.

The future can’t cause the past.

cf. example of *pre-test* being prior to *retry quiz & exam*. 
Important

- Important to use causal modeling algorithms correctly!
  - Which assumptions are reasonable?
  - The future can’t cause the past
    - Except in movies
Important

- Are variables good proxies for what we intend to study (especially if “latent”)?
  - Suppose pre-test isn’t an appropriate measure of prior knowledge.
  - pre-test might not “screen off” retry quiz & exam, so we might still think that retry quiz causes decreased learning (exam).
Causal Modeling

- A powerful tool
- But needs to be used carefully!
Next lecture

- Association rule mining