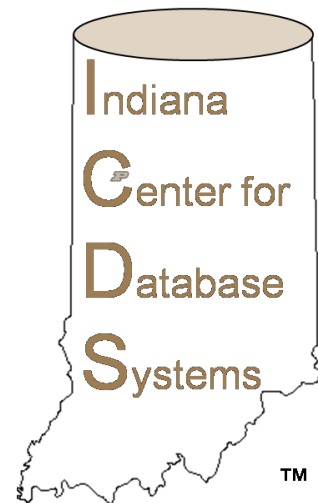




CS37300: Data Mining and Machine Learning

Profs. Tianyi Zhang & Rajiv Khanna

Aug 23, 2023



- Course logistics
- Intro to Data Mining Process

Example

These trains carry toxic chemicals.



These trains do not carry toxic chemicals.

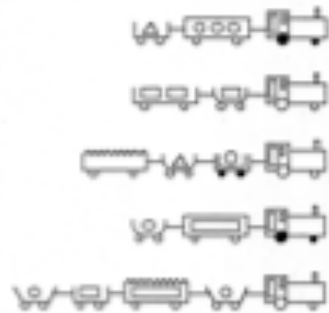


Does this train carry toxic chemicals?

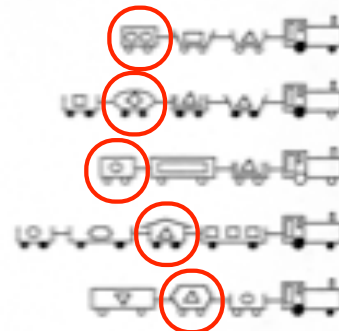


Example rule (1)

These trains carry toxic chemicals.



These trains do not carry toxic chemicals.

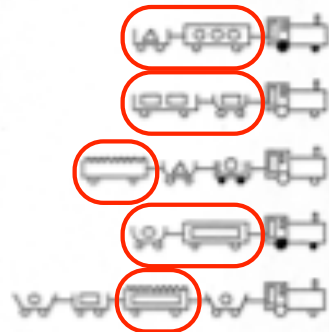


Does this train carry toxic chemicals?



Example rule (2)

These trains carry toxic chemicals.



These trains do not carry toxic chemicals.



Does this train carry toxic chemicals?



How did you devise rules?

- Look for characteristics of one set but not the other?
- Reject potential rules that didn't cover enough examples?
- Examine several potential rules independently and together?
- Consider simple rules first?

How did you devise rules?

- Look for characteristics of one set but not the other? – [class imbalance]
- Reject potential rules that didn't cover enough examples? – [low confidence in signal]
- Examine several potential rules independently and together? – [complexity of the model]
- Consider simple rules first? – [tractability]

A Data Mining System

- Data representation: Describe the data
- Task specification: Outline the goal(s)
- Knowledge representation: Describe the “rules”
- Learning technique:
 - Search: Identify the “best” rule
 - Evaluation function: Estimate confidence
- Prediction technique: Apply the rule

- Data size: vastly larger or changing rapidly
- Data representation: can affect ability to learn and interpret models
- Knowledge representation: needs to capture more subtle forms of probabilistic dependence
- Search space size
- Evaluation functions: difficult to assess confidence in model utility

Elements of Data Mining & Machine Learning Algorithms

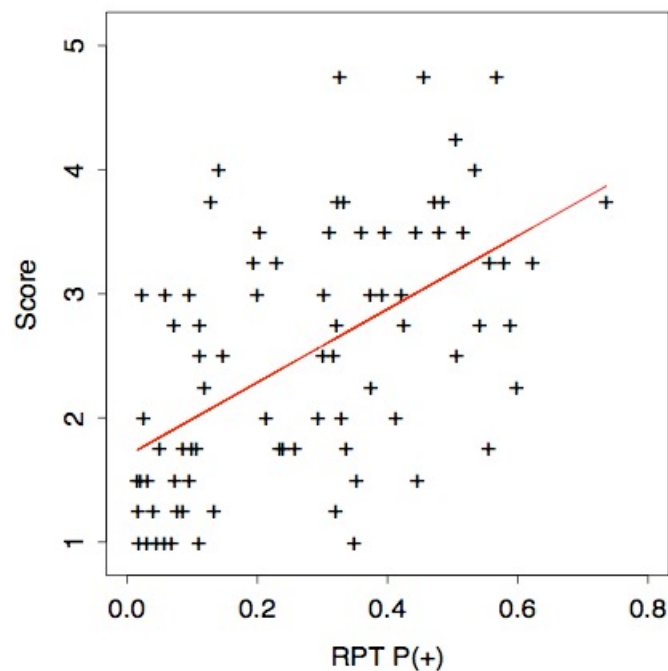
- **Task specification**
- Data representation
- Knowledge representation
- Learning technique
 - Search + scoring
- Prediction and/or interpretation

Task specification

- *Objective of the person who is analyzing the data*
- *Description of the characteristics of the analysis and desired result*
- Examples:
 - From a set of *labeled examples*, devise an *understandable model* that will *accurately predict* whether a stockbroker will commit fraud in the near future.
 - From a set of *unlabeled examples*, cluster stockbrokers into a *set of homogeneous groups* based on their demographic information

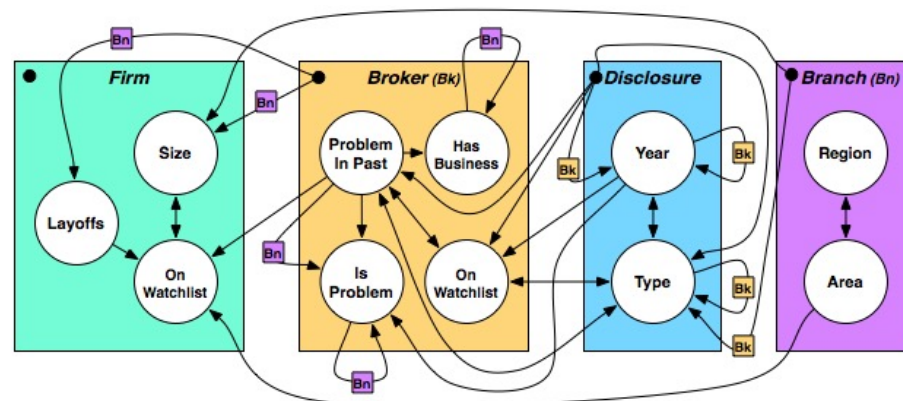
Exploratory data analysis

- Goal
 - Interact with data without clear objective
- Techniques
 - Visualization, adhoc modeling



Descriptive modeling

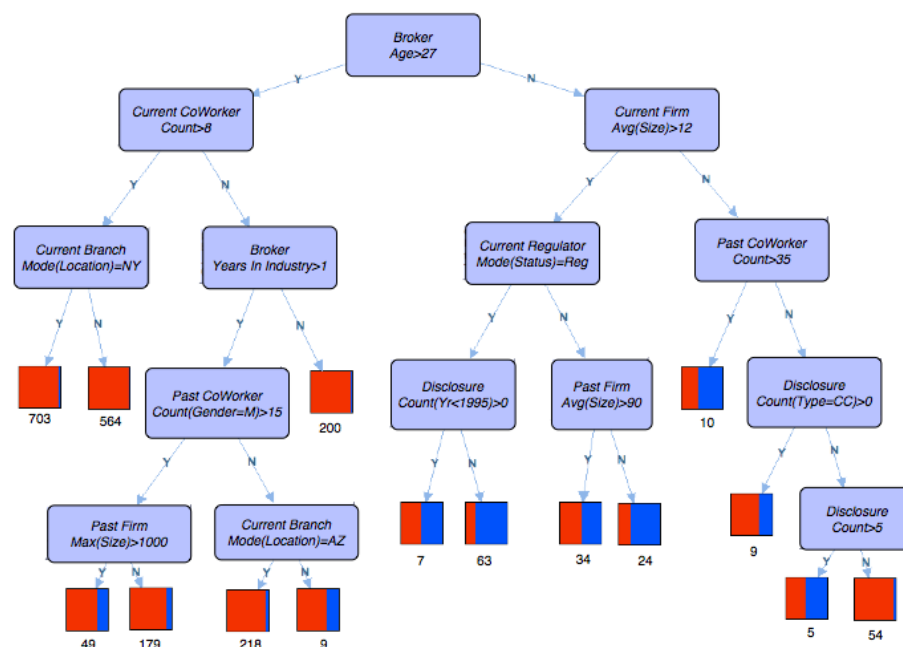
- Goal
 - Summarize the data or the underlying generative process
- Techniques
 - Density estimation, cluster analysis and segmentation



Also known as: **unsupervised** learning

Predictive modeling

- Goal
 - Learn model to predict unknown class label values given observed attribute values
- Techniques
 - Classification, regression



Also known as: **supervised** learning

- Task specification
- **Data representation**
- Knowledge representation
- Learning technique
 - Search + scoring
- Prediction and/or interpretation

Data representation

- Choice of **data structure** for representing individual and collections of measurements
- Individual measurements: single observations (e.g., person's date of birth, product price)
- Collections of measurements: sets of observations that describe an **instance** (e.g., person, product)
- Choice of representation determines applicability of algorithms and can impact modeling effectiveness
- Additional issues: data sampling, data cleaning, feature construction

Individual measurements

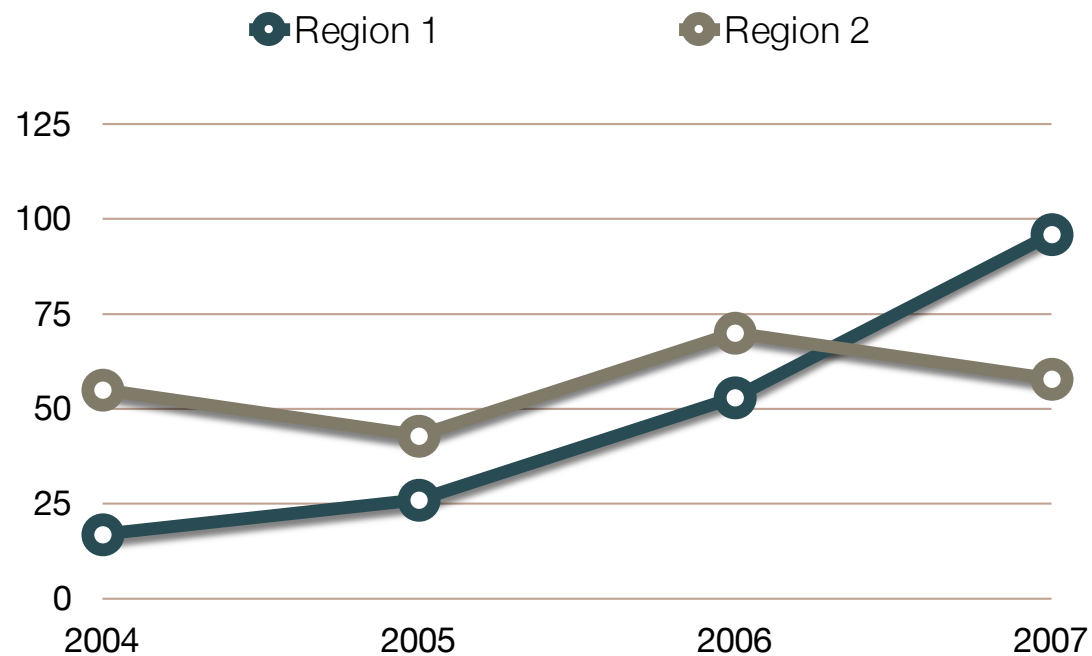
- Unit measurements:
 - Discrete values — categorical or ordinal variables
 - Continuous values — interval and ratio variables
- Compound measurements:
 - $\langle x, y \rangle$
 - $\langle \text{value}, \text{time} \rangle$

Data representation: Table/vectors

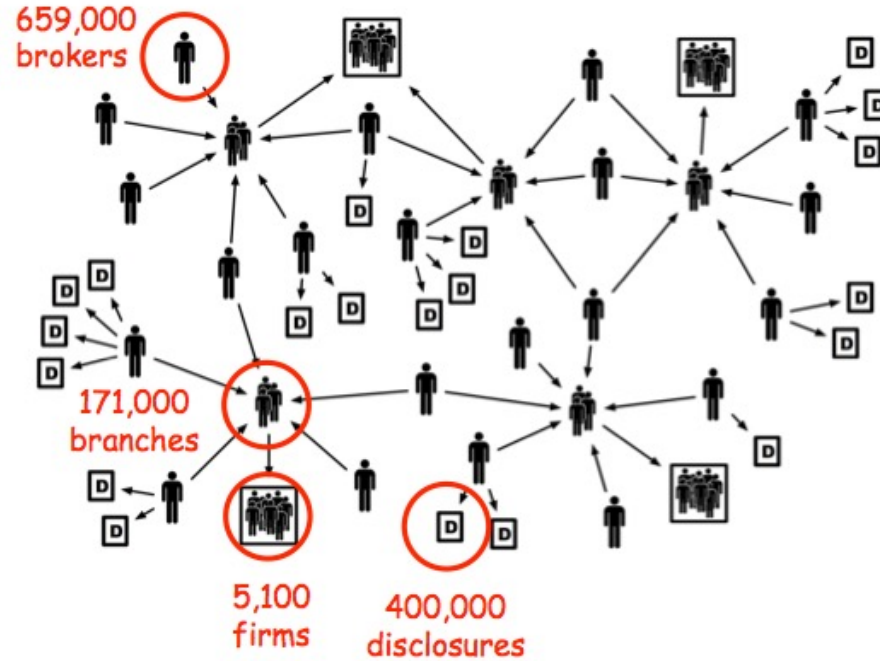
| Fraud | Age | Degree | StartYr | Series7 |
|-------|-----|--------|---------|---------|
| + | 22 | Y | 2005 | N |
| - | 25 | N | 2003 | Y |
| - | 31 | Y | 1995 | Y |
| - | 27 | Y | 1999 | Y |
| + | 24 | N | 2006 | N |
| - | 29 | N | 2003 | N |

N instances \times p attributes

Data representation: Time series/sequences



Data representation: Relational/graph data

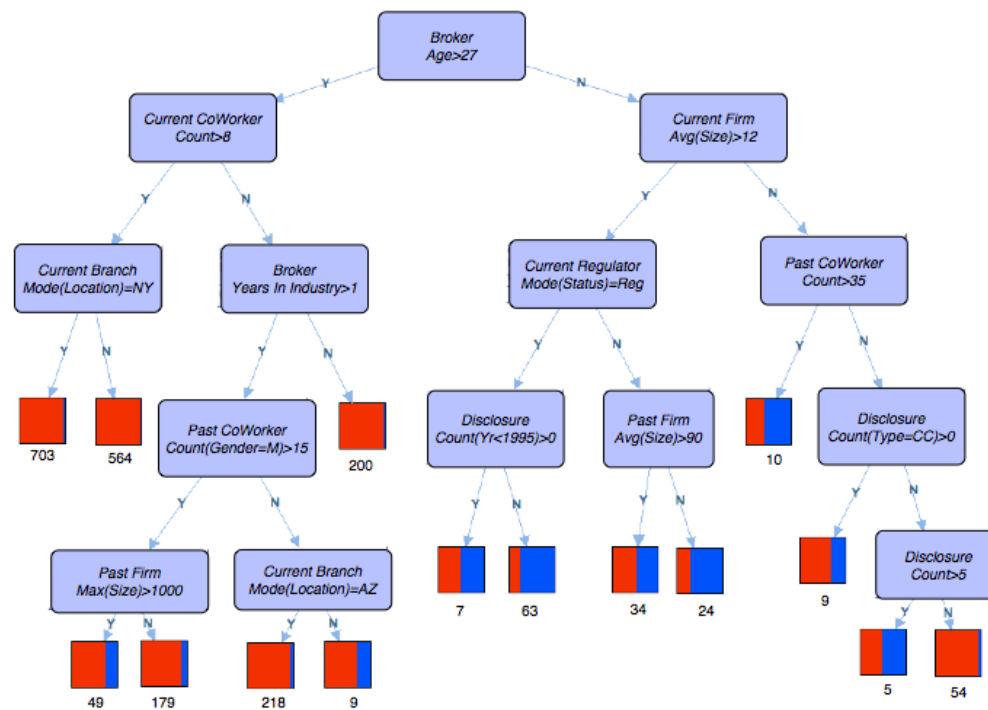


- Task specification
- Data representation
- **Knowledge representation**
- Learning technique
 - Search + scoring
- Prediction and/or interpretation

Knowledge representation

- *Underlying structure of the model or patterns that we seek from the data*
 - Specifies the models/patterns that could be returned as the results of the data mining algorithm
 - Defines the **model space** that algorithms search over (i.e., all possible models/patterns)
- Examples:
 - **If-then rule**
If short closed car **then** toxic chemicals
 - **Conditional probability distribution**
 $P(\text{fraud} \mid \text{age}, \text{degree}, \text{series7}, \text{startYr})$
 - **Decision tree**

Knowledge representation: Classification tree



Each node corresponds to a feature; each leaf a class label or probability distribution

Knowledge representation: Regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_0$$

- x_i are predictor variables
- y is response variable
- Example:
 - Predict number of disclosures given income and trading history

- Task specification
- Data representation
- Knowledge representation
- **Learning technique**
 - Search + scoring
- Prediction and/or interpretation

Learning technique

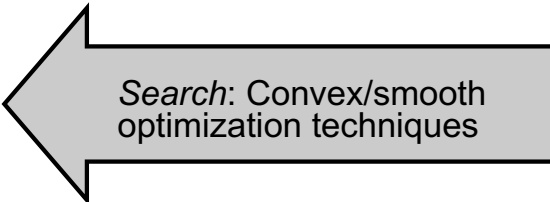
- Method to construct model or patterns from data
- **Model space**
 - Choice of knowledge representation defines a set of possible models or patterns
- **Scoring function**
 - Associates a numerical value (score) with each member of the set of models/patterns
- **Search technique**
 - Defines a method for generating members of the set of models/patterns and determining their score

Scoring function

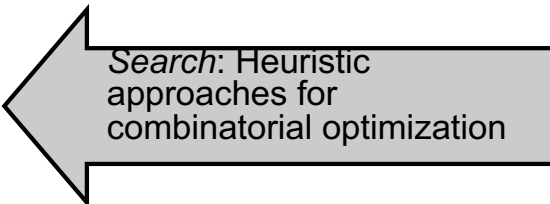
- *A numeric score assigned to each possible model in a search space, **given a reference/input dataset***
 - Used to judge the quality of a particular model for the domain
- Score function are **statistics**—estimates of a population parameter based on a sample of data
- Examples:
 - Misclassification
 - Squared error
 - Likelihood

Parameter estimation vs. structure learning

- Models have both **parameters** and **structure**
- Parameters:
 - Coefficients in regression model
 - Feature values in classification tree
 - Probability estimates in graphical model
- Structure:
 - Variables in regression model
 - Nodes in classification tree
 - Edges in graphical model



Search: Convex/smooth optimization techniques



Search: Heuristic approaches for combinatorial optimization

Example learning problem

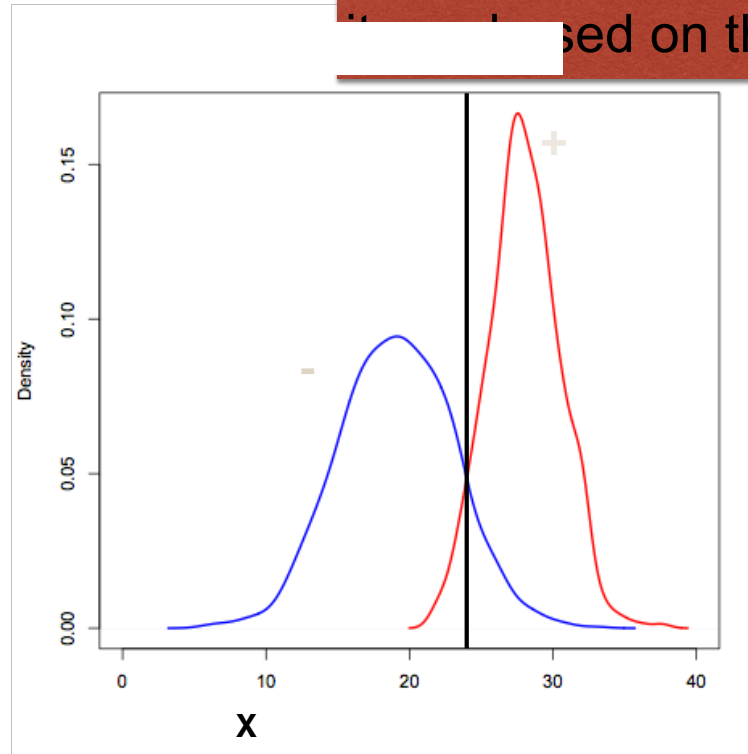
Task: Devise a rule to classify
data based on the attribute **X**

Knowledge representation:
If-then rules

Example rule:
If $x > 25$ then +
Else -

What is the model space?

All possible thresholds



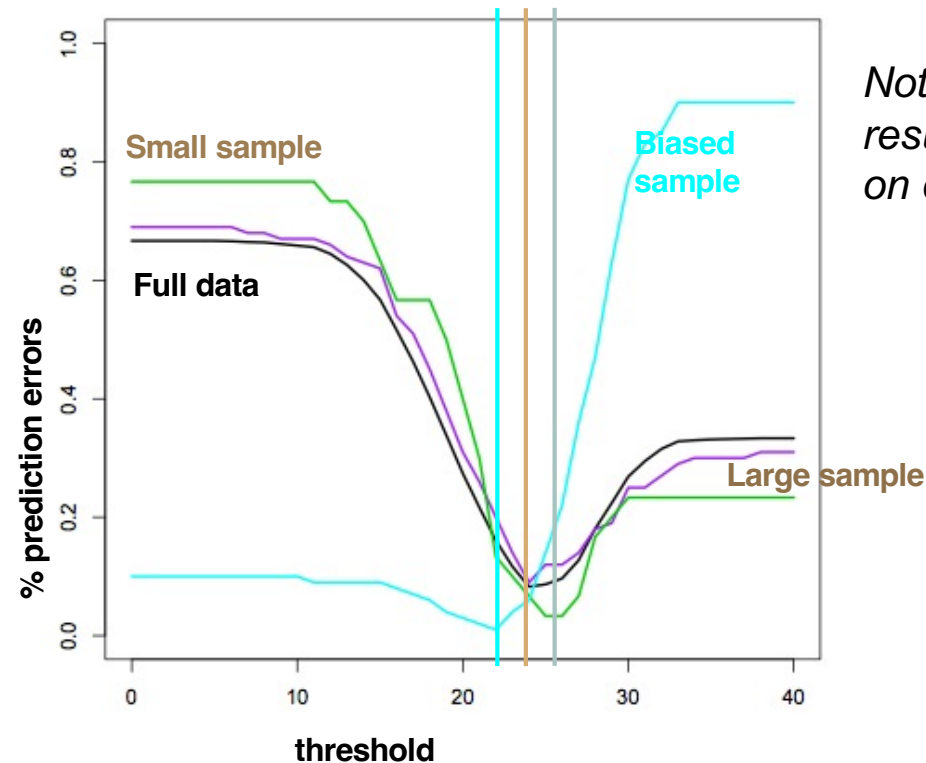
What score function?

Prediction error rate

Score function over model space

Search
procedure?

Try all
thresholds,
select one
with lowest
score



Note: learning
result depends
on **data**

Take-home quiz (Due Friday 9:00am in Gradescope)

Example *claims* from recent news articles that could be supported or disproved by data analysis:

- *The temperature of the planet is rising and the increase is due to human activities such as fossil fuel use and deforestation.*
- *Aspirin is effective in reducing cancer risk.*
- *Fathers who perform an equal share of household chores are more likely to have daughters who aspire to less traditionally feminine occupations.*

Task: Identify two specific claims in news articles in the last week

1. Briefly state the claim
2. Describe the data that is (or could be) used to support the claim
3. Include a reference to the article

Length: One paragraph per claim, max one page total.