CS57300
PURDUE UNIVERSITY

SEPTEMBER 13, 2021

DATA MINING

ELEMENTS OF DATA MINING ALGORITHMS

OVERVIEW

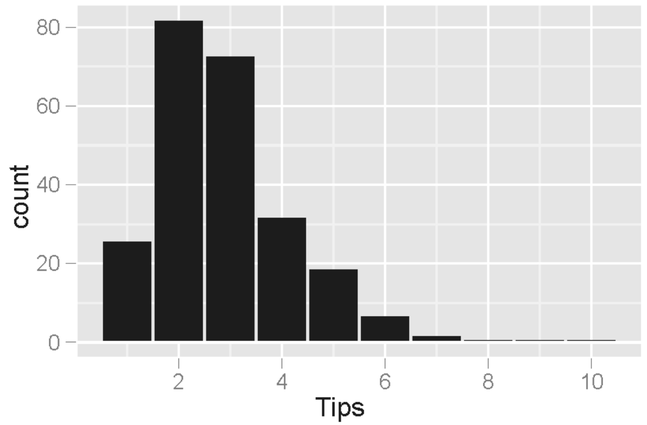
- **Task specification**
- Knowledge representation
- Learning technique
  - Search + scoring
- Prediction and/or interpretation

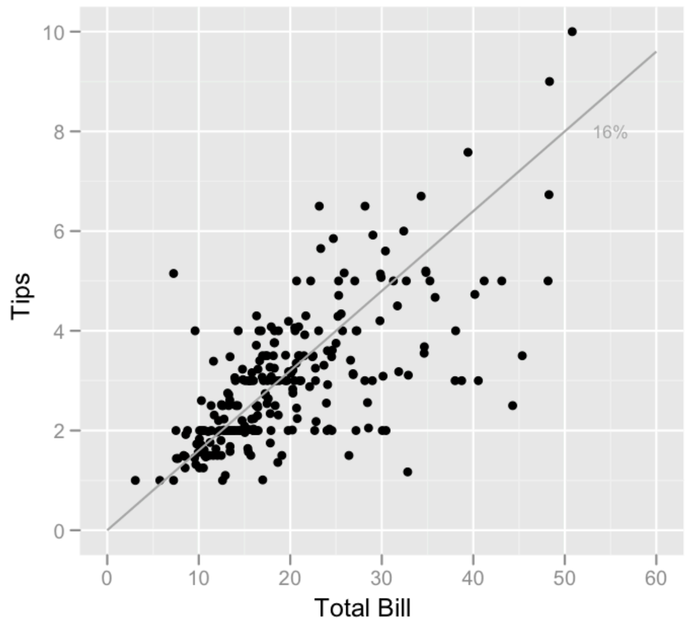
EXPLORATORY DATA ANALYSIS

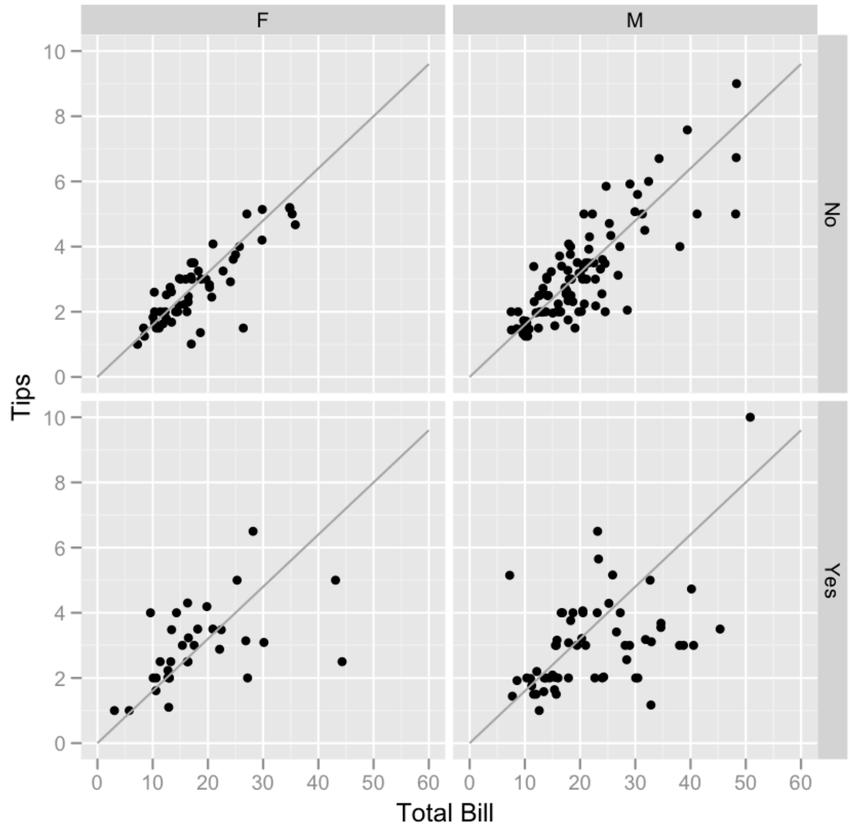
- Goal
  - Interact with data without clear objective
  - Summarize the main characteristics of the data
- Techniques
  - Mostly visualization

EXPLORATORY DATA ANALYSIS EXAMPLE

- What influences the amount of tip that a dining party will give to the waiter?







Cook, D. And Dwayne, D. F. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi

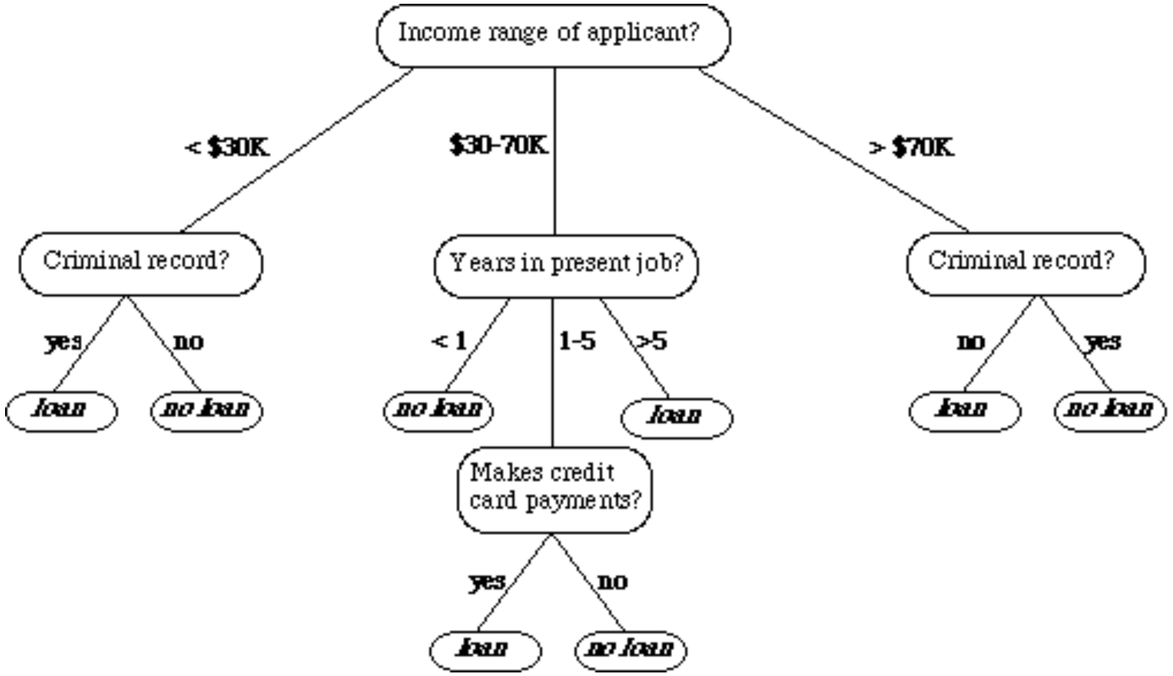
PREDICTIVE MODELING

- Goal
  - Learn model to predict the unknown value of a variable of interest given observed attribute values
- Techniques
  - Classification, regression

Also known as: **supervised** learning

PREDICTIVE MODELING EXAMPLE

- Zestimate: House sales price prediction!
- Predicting loan repayment (and thus decide whether to provide a loan)



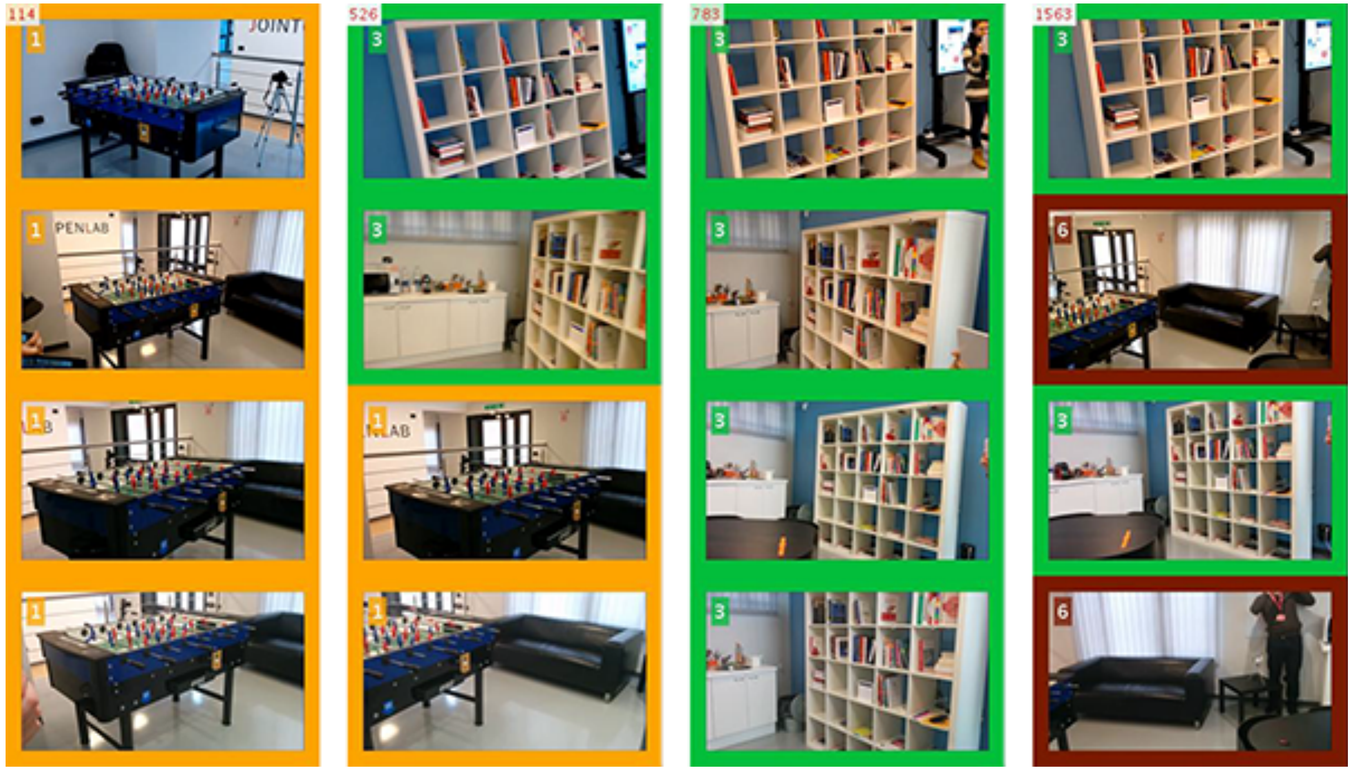
DESCRIPTIVE MODELING

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

Also known as: **unsupervised** learning

DESCRIPTIVE MODELING EXAMPLE

- Video/scene clustering



Milotta et al. RECfusion: Automatic Scene Clustering and Tracking in Videos from Multiple Sources

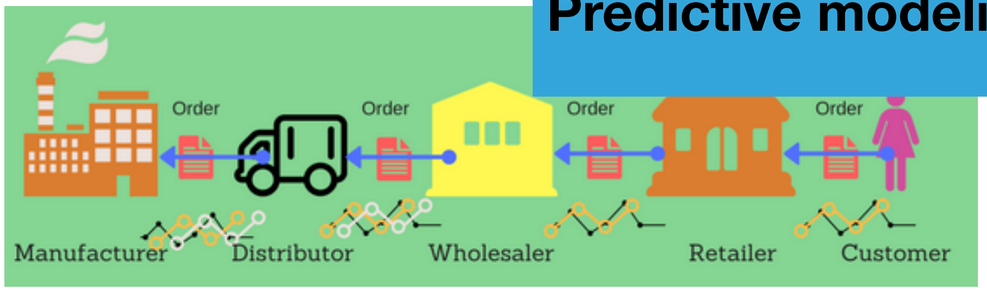
PATTERN DISCOVERY

- Goal
  - Detect patterns and rules that describe subsets of examples
- Techniques
  - Association rules, anomaly detection, etc.

Model: global summary of a data set
Pattern: local to a subset of the data

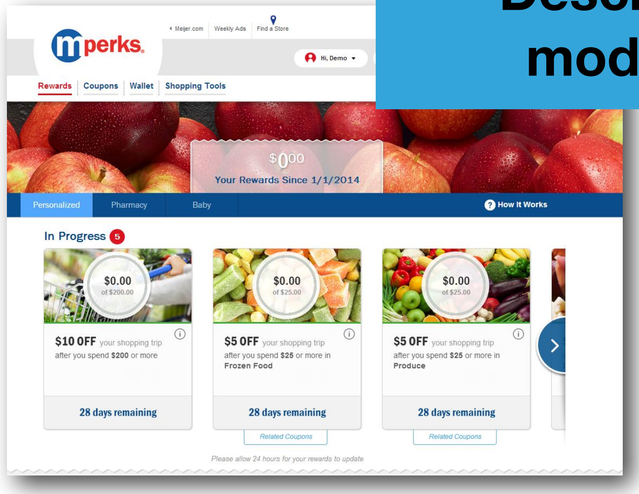
WHAT TASKS ARE THEY?

Predictive modeling

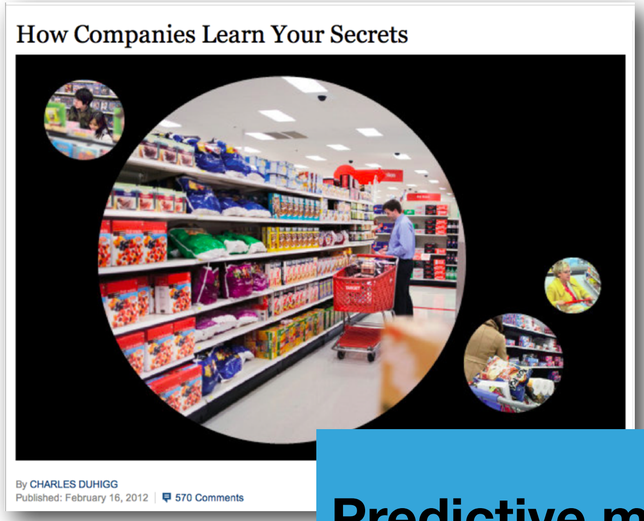


Sales and inventory forecast

Descriptive modeling



Customer segmentation



Predictive modeling

Pregnant custo

Beer & Dia

Pattern Discovery

OVERVIEW

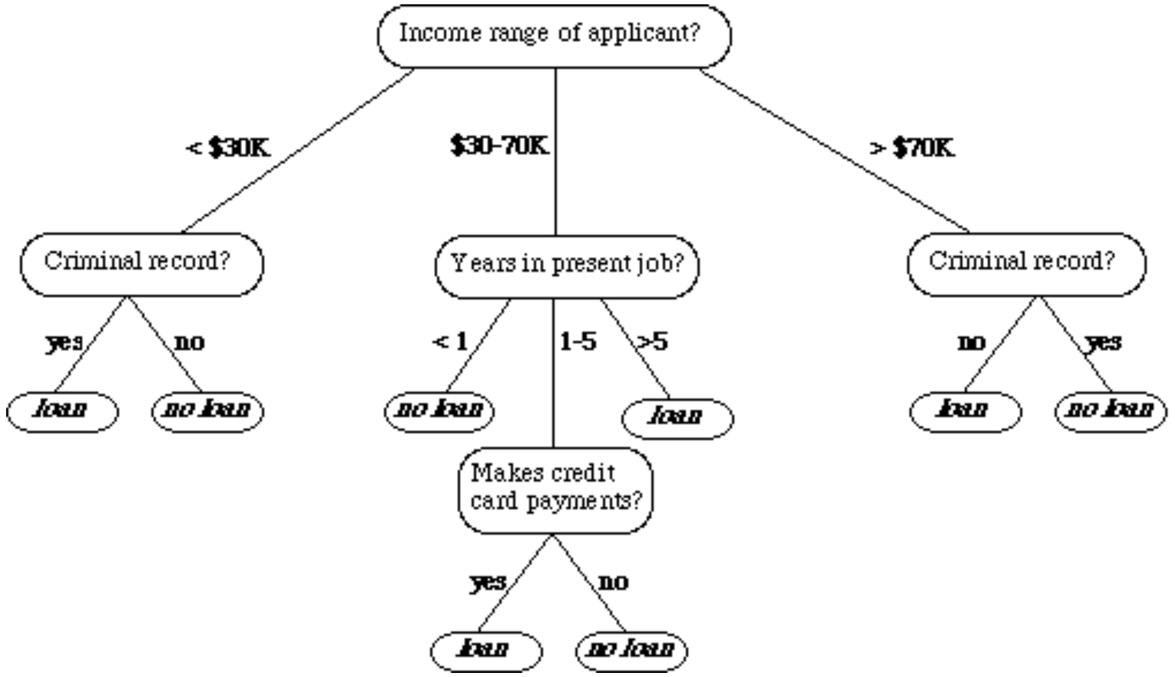
- Task specification
- **Knowledge representation**
- Learning technique
  - Search + scoring
- Prediction and/or interpretation

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 space of possible models/patterns for algorithm to search over

KNOWLEDGE REPRESENTATION EXAMPLE: PREDICTIVE MODELING

- If-then rule
  - *If* (personal income > $70k) AND (criminal record = 'no'), *then* loan=yes
- Decision tree
  - Each node corresponds to an attribute
  - Each leaf is a class label

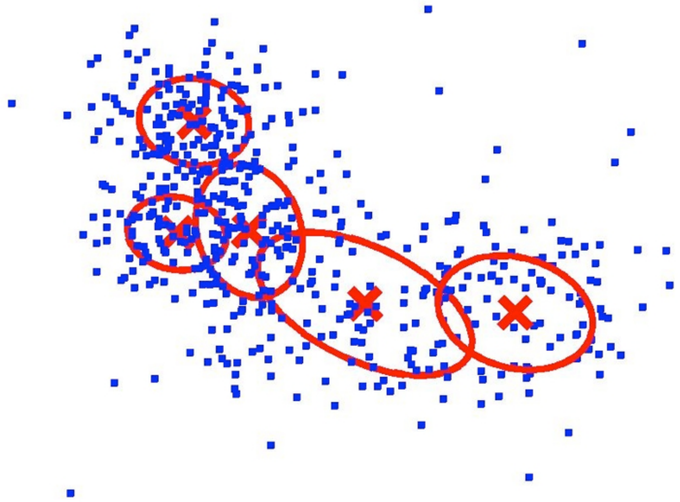


KNOWLEDGE REPRESENTATION EXAMPLE: PREDICTIVE MODELING

- Conditional probability distributions (i.e., P($Y \mid \boldsymbol{X}$))
  - Logistic regression: $log \frac{P(Y=1|\boldsymbol{x})}{1-P(Y=1|\boldsymbol{x})} = \beta_0 + \boldsymbol{\beta x}$
  - Model the log-odds as a linear combination of predictors
- Linear regression
  - $y = \beta_1 x_1 + \beta_2 x_2 ... + \beta_0$
  - y is the response variable, $\boldsymbol{x}$ is the predictor variable

KNOWLEDGE REPRESENTATION EXAMPLE: DESCRIPTIVE MODELING

- Mixture model: Instances represented as a weighted combination of mixture distributions



$$f(x) = \sum_{k=1}^{K} w_k f_k(x;\theta)$$

probability of observing x

likelihood of x being generated from cluster k

likelihood of point belonging to cluster k

KNOWLEDGE REPRESENTATION EXAMPLE: PATTERN DISCOVERY

- Association rules
  - $I = \{i_1, i_2, \ldots, i_n\}$ is a set of n items
  - $T = \{t_1, t_2, \ldots, t_m\}$ is a set of m transactions
  - An association rule has the form $X \rightarrow Y$, where X and Y are subsets of I, which means if items in X appear in a transaction, then items in Y are likely to appear in that transaction
    - E.g., {beer} → {diaper}; {bread} → {milk}

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

OVERVIEW

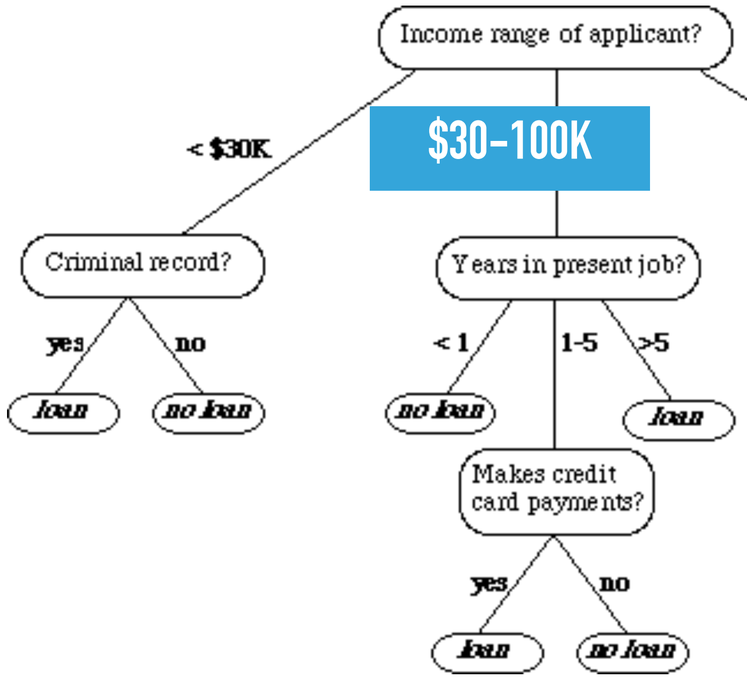
- Task specification
- Knowledge representation
- **Learning technique**
  - Search + scoring
- Prediction and/or interpretation

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, determining their score, and identifying the ones with the “best” score

MODEL SPACE

- Defined by the choice of knowledge representation
- Decision tree:



What values to split on?

Loan

What attributes to include?

MODEL PARAMETERS AND STRUCTURE

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

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

- **Parameters:**
  - Feature values in classification tree
  - Coefficients in regression model
  - Probability estimates in graphical model
- **Structure:**
  - Nodes in classification tree
  - Variables in regression model
  - Edges in graphical model

Search: Convex/smooth optimization techniques

Search: Heuristic approaches for combinatorial optimization

EXAMPLE LEARNING PROBLEM

Task: Devise a rule to classify items based on the attribute ***x***

Knowledge representation:

If-then rules

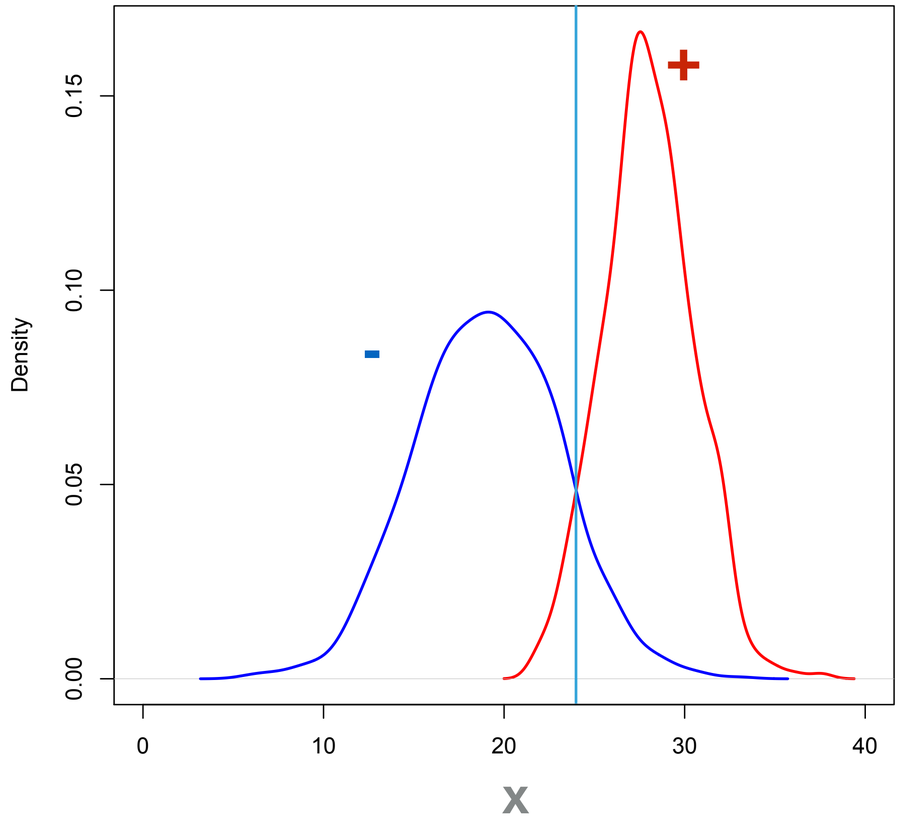
Example rule:

If x > 24 then **+**

Else **-**

What is the model space?

All possible thresholds



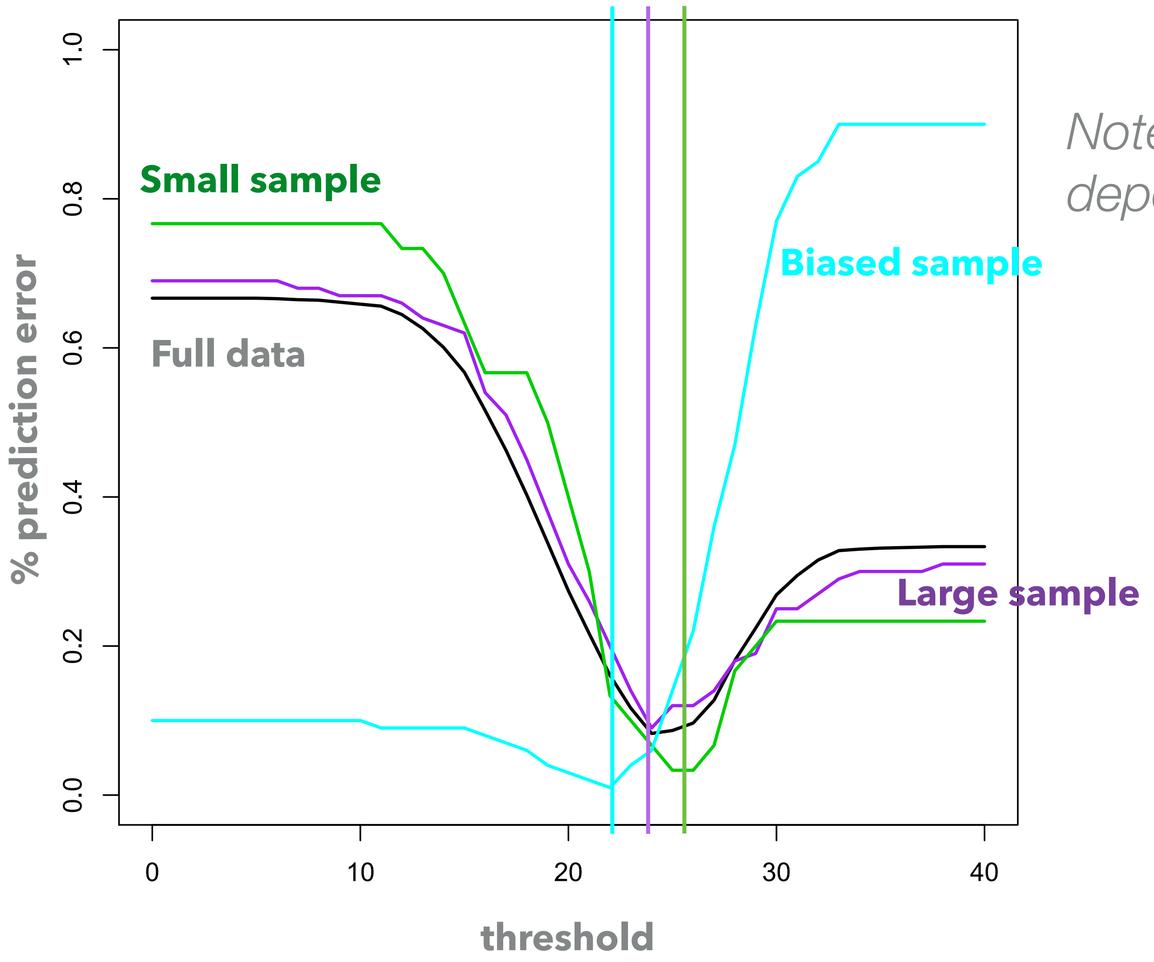
What score function?

Prediction error rate

SCORE FUNCTION OVER MODEL SPACE

Search procedure?

Try all thresholds, select one with lowest score



*Note: learning result depends on **data***

OVERVIEW

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

INFERENCE AND INTERPRETATION

- Prediction technique
  - Method to apply learned model to new data for prediction/analysis
  - Only applicable for predictive and some descriptive models
  - Prediction is often used during **learning** (i.e., search) to determine value of scoring function
- Interpretation of results
  - Objective: significance measures
  - Subjective: importance, interestingness, novelty

EXAMPLE: IDENTIFYING EMAIL SPAM

- Task
  - Design automatic spam detector that can differentiate between labeled emails
- Data
  - Table of relative word/ punctuation frequencies
- Knowledge representation
  - If/then rules with conjunctions of features
- Learning technique
  - **Search** over set of rules, **select** rule with maximum accuracy on training data

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

if (%george < 0.6) & (%you > 1.5) then spam
else email.

EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

- Data analysis approach that employs a number of (mostly graphical) techniques to:
  - Maximize insight into data
  - Uncover underlying structure
  - Identify important variables
  - Detect outliers and anomalies
  - Test underlying modeling assumptions
  - Develop parsimonious models
  - Generate hypotheses from data

DATA VISUALIZATION

VISUALIZATION

- Human eye/brain have evolved powerful methods to detect structure in nature
- Display data in ways that exploit human pattern recognition abilities
- Limitation: Can be difficult to apply if data size (number of dimensions or instances) is large

VISUALIZING/SUMMARIZING DATA

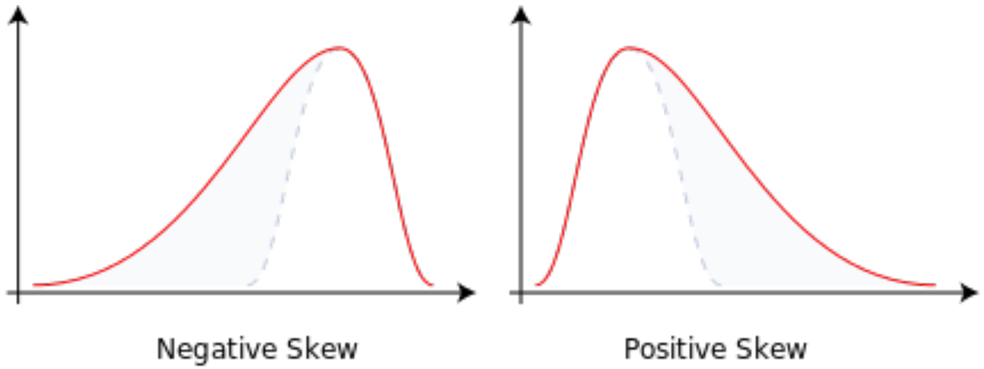
- Low-dimensional data
  - Summarizing data with simple statistics
  - Plotting raw data (1D, 2D, 3D)
- Higher-dimensional data
  - Dimensionality reduction: e.g., principal component analysis

DATA SUMMARIZATION

- Measures of location
  - Mean: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x(i)$
  - Median: value with 50% of points above and below
  - Quartile: value with 25% (75%) points below
  - Mode: most common value

DATA SUMMARIZATION

- Measures of dispersion or variability
  - Variance: $\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2$
  - Standard deviation: $\hat{\sigma}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2}$
  - Range: difference between max and min point
  - Interquartile range: difference between 1st and 3rd Q
  - Skew: $\frac{\sum_{i=1}^{n}(x(i)-\hat{\mu})^3}{(\sum_{i=1}^{n}(x(i)-\hat{\mu})^2)^{\frac{3}{2}}}$



DATA VISUALIZATION

- Serve for different purposes
  - Composition: e.g., see for a discrete dimension x_i, the fraction of each values
  - Distribution: e.g., see the distribution of a continuous dimension of data x_i
  - Comparison:
    - Compare values of two continuous dimensions of the data, x_i and x_j
    - Given discrete x_i, compare the values of x_j when x_i takes different values.
  - Relationship: e.g., examine the relationship between x_i and x_j

COMPOSITION

- Pie charts
- Stacked bars
  - Temporal trends
  - Compare across groups