PURDUE UNIVERSITY

CS34800
Information Systems

Data Mining and Data Analysis

Prof. Chris Clifton

28 November 2016

Thanks to Prof. Jiawei Han and others for some of this material

Indiana Center for Database Systems ™

Decision Support Systems

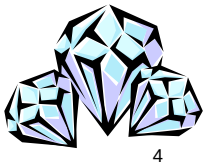
- **Decision-support systems** are used to make business decisions, often based on data collected by on-line transaction-processing systems.
- Examples of business decisions:
  - What items to stock?
  - What insurance premium to change?
  - To whom to send advertisements?
- Examples of data used for making decisions
  - Retail sales transaction details
  - Customer profiles (income, age, gender, etc.)

Decision-Support Systems: Overview

- **Data analysis** tasks are simplified by specialized tools and SQL extensions
  - Example tasks
    - For each product category and each region, what were the total sales in the last quarter and how do they compare with the same quarter last year
    - As above, for each product category and each customer category
- **Statistical analysis** packages (e.g., : S++) can be interfaced with databases
  - Statistical analysis is a large field, but not covered here
- **Data mining** seeks to discover knowledge automatically in the form of statistical rules and patterns from large databases.
- A **data warehouse** archives information gathered from multiple sources, and stores it under a unified schema, at a single site.
  - Important for large businesses that generate data from multiple divisions, possibly at multiple sites
  - Data may also be purchased externally

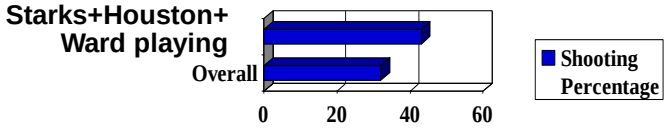
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs



What is Data Mining?
Real Example from the NBA

- Play-by-play information recorded by teams
  - Who is on the court
  - Who shoots
  - Results
- Coaches want to know what works best
  - Plays that work well against a given team
  - Good/bad player matchups
- Advanced Scout (from IBM Research) is a data mining tool to answer these questions

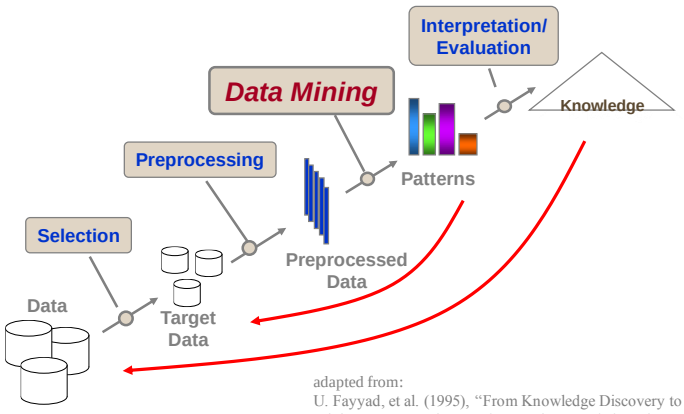


http://www.nba.com/news_feat/beyond/0126.html

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis

Knowledge Discovery in Databases: Process



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

See also: http://www.crisp-dm.org

Data Mining: History of the Field

- Knowledge Discovery in Databases workshops started '89
  - Now a conference under the auspices of ACM SIGKDD
  - IEEE conference series started 2001
- Key founders / technology contributors:
  - Usama Fayyad, JPL (and a variety of positions since then, including a data mining startup)
  - Gregory Piatetsky-Shapiro, GTE (now runs Kdnuggets, a data mining website)
  - Rakesh Agrawal, IBM Research (a variety of positions since then, now running a startp, Data Insights Laboratories)

The term "data mining" has been around since at least 1983 – as a pejorative term in the statistics community

What Can Data Mining Do?

- Cluster
- Classify
  - Categorical, Regression
- Summarize
  - Summary statistics, Summary rules
- Link Analysis / Model Dependencies
  - Association rules
- Sequence analysis
  - Time-series analysis, Sequential associations
- Detect Deviations

Clustering

- Find groups of similar data items
- Statistical techniques require some definition of "distance" (e.g. between travel profiles) while conceptual techniques use background concepts and logical descriptions

Uses:

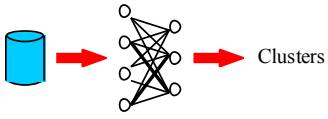
- Demographic analysis

Technologies:

- Self-Organizing Maps
- Probability Densities
- Conceptual Clustering

"Group people with similar travel profiles"

- George, Patricia
- Jeff, Evelyn, Chris
- Rob



Classification

- Find ways to separate data items into pre-defined groups
  - We know X and Y belong together, find other things in same group
- Requires "training data": Data items where group is known

Uses:

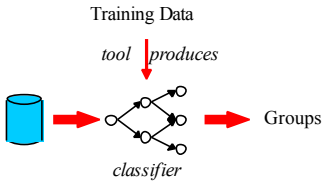
- Profiling

Technologies:

- Generate decision trees (results are human understandable)
- Neural Nets

"Route documents to most likely interested parties"

- English or non-english?
- Domestic or Foreign?



Association Rules

- Identify dependencies in the data:
  - X makes Y likely
- Indicate significance of each dependency
- Bayesian methods

Uses:

- Targeted marketing

"Find groups of items commonly purchased together"

- People who purchase fish are extraordinarily likely to purchase wine
- People who purchase turkey are extraordinarily likely to purchase cranberries

Date/Time/Register	Fish	Turkey	Cranberries	Wine	…
12/6 13:15 2	N	Y	Y	Y	…
12/6 13:16 3	Y	N	N	Y	…

Technologies:

- AIS, SETM, Hugin, TETRAD II

Sequential Associations

- Find event sequences that are unusually likely
- Requires "training" event list, known "interesting" events
- Must be robust in the face of additional "noise" events

Uses:

- Failure analysis and prediction

Technologies:

- Dynamic programming (Dynamic time warping)
- "Custom" algorithms

"Find common sequences of warnings/faults within 10 minute periods"

- Warn 2 on Switch C preceded by Fault 21 on Switch B
- Fault 17 on any switch preceded by Warn 2 on any switch

Time	Switch	Event
21:10	B	Fault 21
21:11	A	Warn 2
21:13	C	Warn 2
21:20	A	Fault 17

Deviation Detection

- Find unexpected values, outliers

Uses:

- Failure analysis
- Anomaly discovery for analysis

Technologies:

- clustering/classification methods
- Statistical techniques
- visualization

- "Find unusual occurrences in IBM stock prices"

Sample date	*Event*	*Occurrences*
58/07/04	Market closed	317 times
59/01/06	2.5% dividend	2 times
59/04/04	50% stock split	7 times
73/10/09	not traded	1 time

Date	Close	Volume	Spread
58/07/02	369.50	314.08	.022561
58/07/03	369.25	313.87	.022561
58/07/04	Market Closed		
58/07/07	370.00	314.50	.022561

War Stories: Warehouse Product Allocation

The second project, identified as "Warehouse Product Allocation," was also initiated in late 1995 by RS Components' IS and Operations Departments. In addition to their warehouse in Corby, the company was in the process of opening another 500,000-square-foot site in the Midlands region of the U.K. To efficiently ship product from these two locations, it was essential that RS Components know in advance what products should be allocated to which warehouse. For this project, the team used IBM Intelligent Miner and additional optimization logic to split RS Components' product sets between these two sites so that the number of partial orders and split shipments would be minimized.

Parker says that the Warehouse Product Allocation project has directly contributed to a significant savings in the number of parcels shipped, and therefore in shipping costs. In addition, he says that the Opportunity Selling project not only increased the level of service, but also made it easier to provide new subsidiaries with the value-added knowledge that enables them to quickly ramp-up sales.

"By using the data mining tools and some additional optimization logic, IBM helped us produce a solution which heavily outperformed the best solution that we could have arrived at by conventional techniques," said Parker. "The IBM group tracked historical order data and conclusively demonstrated that data mining produced increased revenue that will give us a return on investment 10 times greater than the amount we spent on the first project."

http://direct.boulder.ibm.com/dss/customer/rscomp.html

War Stories: Inventory Forecasting

American Entertainment Company

Forecasting demand for inventory is a central problem for any distributor. Ship too much and the distributor incurs the cost of restocking unsold products; ship too little and sales opportunities are lost.

IBM Data Mining Solutions assisted this customer by providing an inventory forecasting model, using segmentation and predictive modeling. This new model has proven to be considerably more accurate than any prior forecasting model.

More war stories (many humorous) starting with slide 21 of:
http://robotics.stanford.edu/~ronnyk/chasm.pdf

Data Mining Complications

- Volume of Data
  - Clever algorithms needed for reasonable performance
- Interest measures
  - How do we ensure algorithms select “interesting” results?
- “Knowledge Discovery Process” skill required
  - How to select tool, prepare data?
- Data Quality
  - How do we interpret results in light of low quality data?
- Data Source Heterogeneity
  - How do we combine data from multiple sources?

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Can We Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering
- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Bayesian Classification
- Instance Based Methods
- Classification by decision tree induction
- Classification by Neural Networks
- Classification by Support Vector Machines (SVM)
- Prediction
- Classification accuracy
- Summary

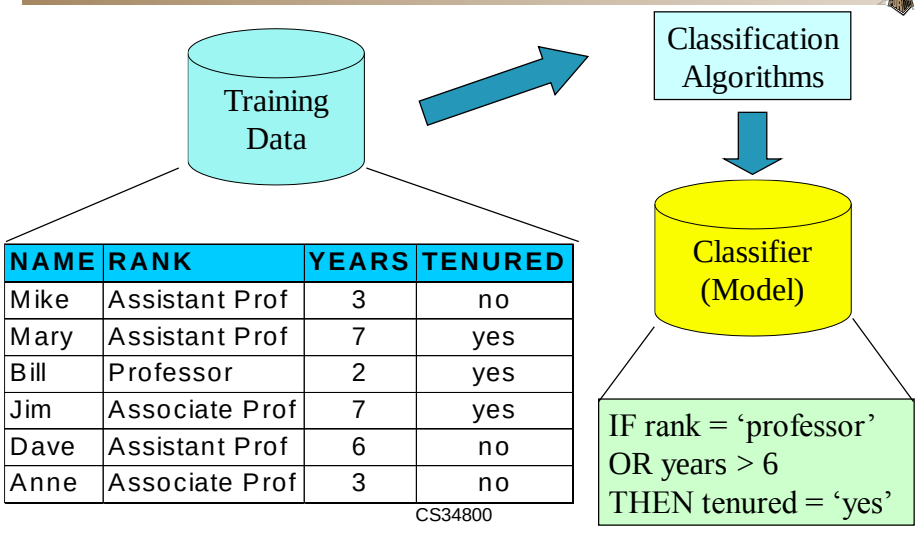
Classification vs. Prediction

- Classification:
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction:
  - models continuous-valued functions, i.e., predicts unknown or missing values
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

Classification—A Two-Step Process

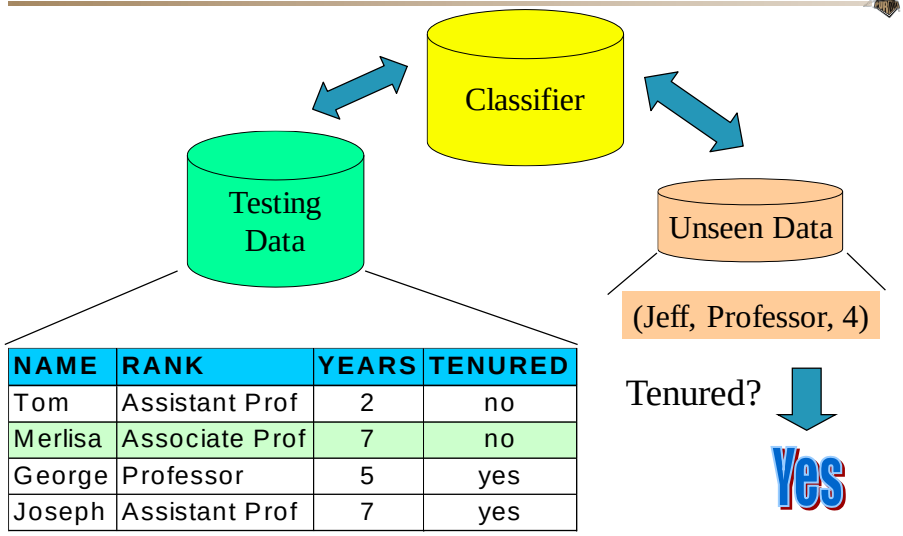
- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

Classification Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Classification Process (2): Use the Model in Prediction

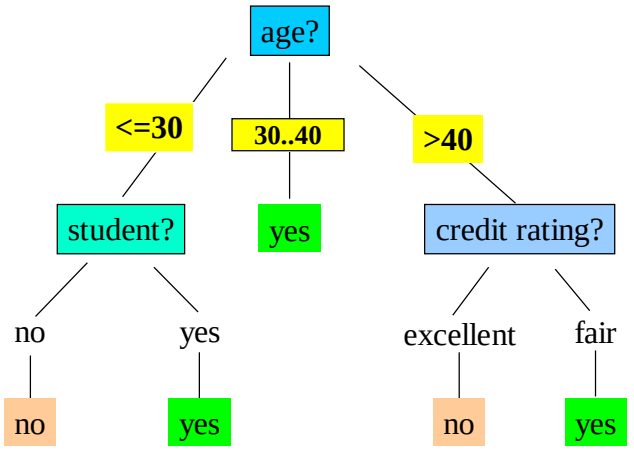


NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	
31…40	high	yes	fair	
>40	medium	no	excellent	

A Decision Tree for *"buys_computer"*



PURDUE UNIVERSITY

CS34800
Information Systems

Data Mining and Data Analysis

Prof. Chris Clifton

30 November 2016

Thanks to Prof. Jiawei Han and others for some of this material

Indiana Center for Database Systems ™

Classification and Prediction

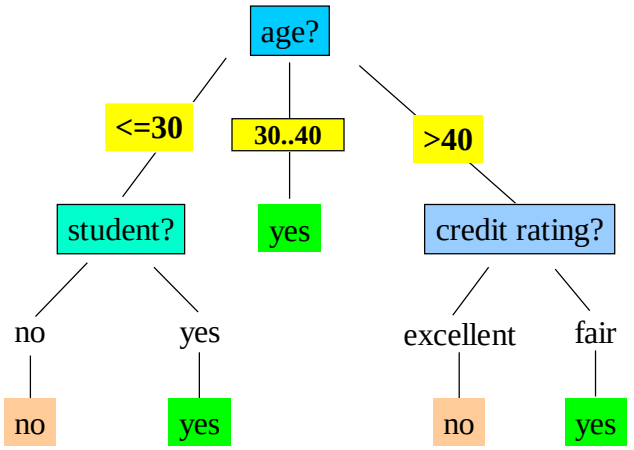
- What is classification? What is prediction?
- Issues regarding classification and prediction
- Bayesian Classification
- Instance Based Methods
- Classification by decision tree induction
- Classification by Neural Networks
- Classification by Support Vector Machines (SVM)
- Prediction
- Classification accuracy
- Summary

Training Dataset

This follows an example from Quinlan's ID3

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Output: A Decision Tree for "*buys_computer*"



Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- S contains s_i tuples of class C_i for i = {1, …, m}
- information measures info required to classify any arbitrary tuple

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

- entropy of attribute A with values $\{a_1,a_2,\ldots,a_v\}$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$$

- information gained by branching on attribute A

$$Gain(A) = I(s_1,s_2,\ldots,s_m) - E(A)$$

Attribute Selection by Information Gain Computation

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"
- I(p, n) = I(9, 5) =0.940
- Compute the entropy for *age*:

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
30…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$E(age) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes and 3 no. Hence

$$Gain(age) = I(p,n) - E(age) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

Other Attribute Selection Measures

- Gini index (CART, IBM IntelligentMiner)
  - All attributes are assumed continuous-valued
  - Assume there exist several possible split values for each attribute
  - May need other tools, such as clustering, to get the possible split values
  - Can be modified for categorical attributes

Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Bayesian Classification
- Instance Based Methods
- Classification by decision tree induction
- Classification by Neural Networks
- Classification by Support Vector Machines (SVM)
- Prediction
- Classification accuracy
- Summary

Bayes' Theorem: Basics

- Let X be a data sample whose class label is unknown
- Let H be a hypothesis that X belongs to class C
- For classification problems, determine P(H|X): the probability that the hypothesis holds given the observed data sample X
- P(H): prior probability of hypothesis H (i.e. the initial probability before we observe any data, reflects the background knowledge)
- P(X): probability that sample data is observed
- P(X|H) : probability of observing the sample X, given that the hypothesis holds

Bayes' Theorem

- Given training data *X, posteriori probability of a hypothesis H, P(H|X)* follows the Bayes theorem

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

- Informally, this can be written as

  posterior =likelihood x prior / evidence

- MAP (maximum posteriori) hypothesis

$$h_{MAP} \equiv \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} P(D \mid h)P(h).$$

- Practical difficulty: require initial knowledge of many probabilities, significant computational cost

Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

- The product of occurrence of say 2 elements x_1 and x_2, given the current class is C, is the product of the probabilities of each element taken separately, given the same class $P([y_1,y_2],C) = P(y_1,C) * P(y_2,C)$
- No dependence relation between attributes
- Greatly reduces the computation cost, only count the class distribution.
- Once the probability $P(X|C_i)$ is known, assign X to the class with maximum $P(X|C_i)*P(C_i)$

Training dataset

Class:
C1:buys_computer= 'yes'
C2:buys_computer= 'no'

Data sample
X =(age<=30,
Income=medium,
Student=yes
Credit_rating=
Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayesian Classifier: Example

- Compute P(X/Ci) for each class
  P(age="<30" | buys_computer="yes") = 2/9=0.222
  P(age="<30" | buys_computer="no") = 3/5 =0.6
  P(income="medium" | buys_computer="yes")= 4/9 =0.444
  P(income="medium" | buys_computer="no") = 2/5 = 0.4
  P(student="yes" | buys_computer="yes)= 6/9 =0.667
  P(student="yes" | buys_computer="no")= 1/5=0.2
  P(credit_rating="fair" | buys_computer="yes")=6/9=0.667
  P(credit_rating="fair" | buys_computer="no")=2/5=0.4

X=(age<=30 ,income =medium, student=yes,credit_rating=fair)

P(X|Ci) : P(X|buys_computer="yes")= 0.222 x 0.444 x 0.667 x 0.0.667 =0.044
P(X|buys_computer="no")= 0.6 x 0.4 x 0.2 x 0.4 =0.019

P(X|Ci)*P(Ci) : P(X|buys_computer="yes") * P(buys_computer="yes")=0.028
P(X|buys_computer="no") * P(buys_computer="no")=0.007

X belongs to class "buys_computer=yes"

Naïve Bayes Classifier: Comments

- Advantages :
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence , therefore loss of accuracy
  - Practically, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history etc
    Symptoms: fever, cough etc., Disease: lung cancer, diabetes etc
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier
- How to deal with these dependencies?
  - Bayesian Belief Networks

Data Mining applied to Aviation Safety Records (*Eric Bloedorn*)

- Many groups record data regarding aviation safety including the National Transportation Safety Board (NTSB) and the Federal Aviation Administration (FAA)
- Integrating data from different sources as well as mining for patterns from a mix of both structured fields and free text is a difficult task
- The goal of our initial analysis is to determine how data mining can be used to improve airline safety by finding patterns that predict safety problems

Aircraft Accident Report

- This data mining effort is an extension of the FAA Office of System Safety's Flight Crew Accident and Incident Human Factors Project
- In this previous approach two database-specific human error models were developed based on general research into human factors
  - FAA's Pilot Deviation database (PDS)
  - NTSB's accident and incident database
- These error models check for certain values in specific fields
- Result
  - Classification of some accidents caused by human mistakes and slips.

Problem

- Current model cannot classify a large number of records
- A large percentage of cases are labeled 'unclassified' by current model
  - ~58,000 in the NTSB database (90% of the events identified as involving people)
  - ~5,400 in the PDS database (93% of the events)
- Approximately 80,000 NTSB events are currently labeled 'unknown'
- Classification into meaningful human error classes is low because the explicit fields and values required for the models to fire are not being used
- Models must be adjusted to better describe data

Data mining Approach

- Use information from text fields to supplement current structured fields by extracting features from text in accident reports
- Build a human-error classifier directly from data
  - Use expert to provide class labels for events of interest such as 'slips', 'mistakes' and 'other'
  - Use data-mining tools to build comprehensible rules describing each of these classes

Example Rule

- Sample Decision rule using current model features and text features

  If (person_code_1b= 5150,4105,5100,4100) and
  ((crew-subject-of-intentional-verb = true) or
  (modifier_code_1b = 3114))
  Then
  mistake

- "If pilot or copilot is involved and either the narrative, or the modifier code for 1b describes the crew as intentionally performing some action then this is a mistake"

Data Mining Ideas: Logistics

- Delivery delays
  - Debatable what data mining will do here; best match would be related to "quality analysis": given lots of data about deliveries, try to find common threads in "problem" deliveries
- Predicting item needs
  - Seasonal
    - Looking for cycles, related to similarity search in time series data
    - Look for similar cycles between products, even if not repeated
  - Event-related
    - Sequential association between event and product order (probably weak)