Data Preparation

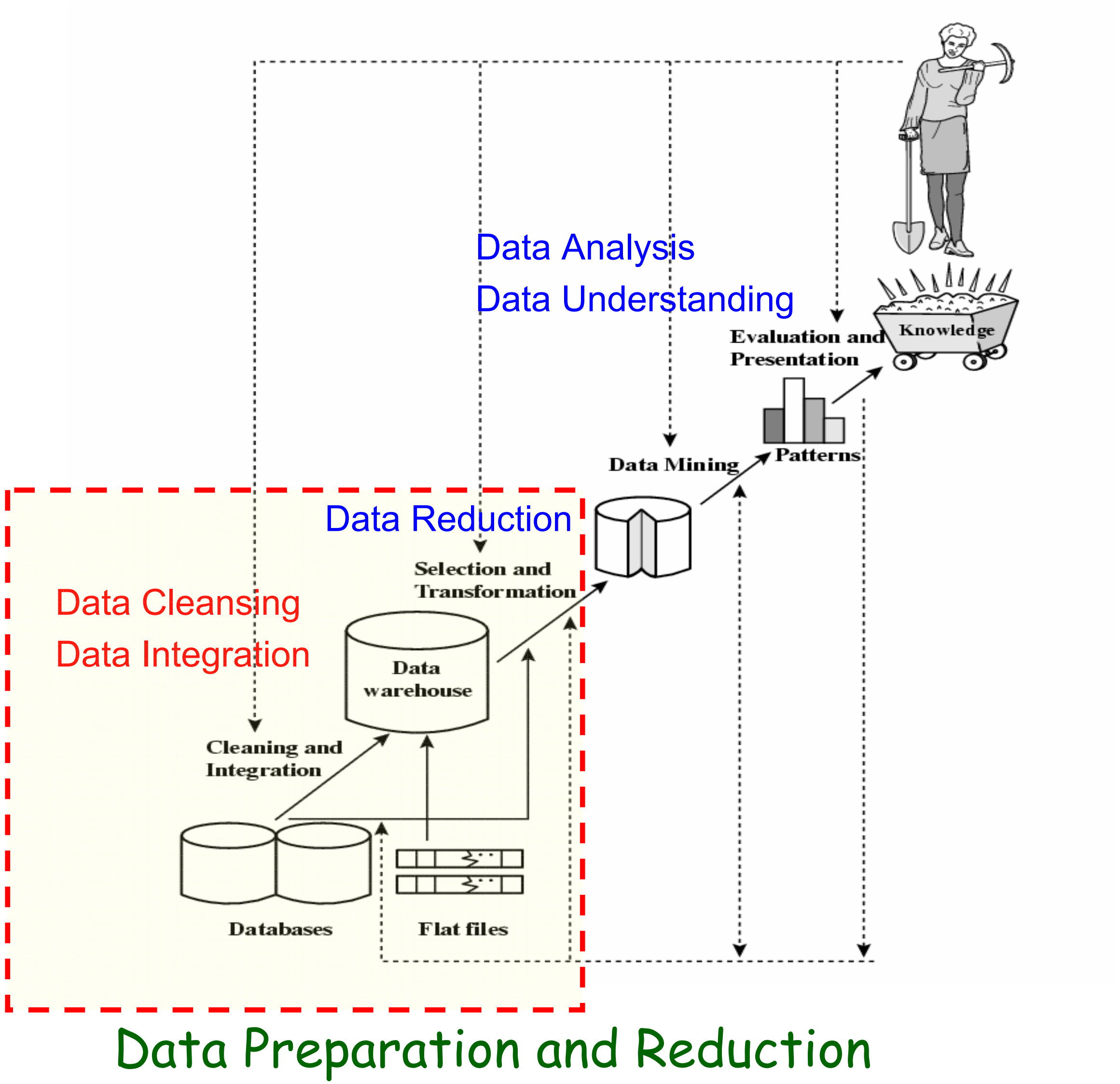
Berlin Chen 2005

References:

1. *Data Mining: Concepts, Models, Methods and Algorithms*, Chapters 2, 3

2. *Data Mining: Concepts and Techniques*, Chapters 3, 8

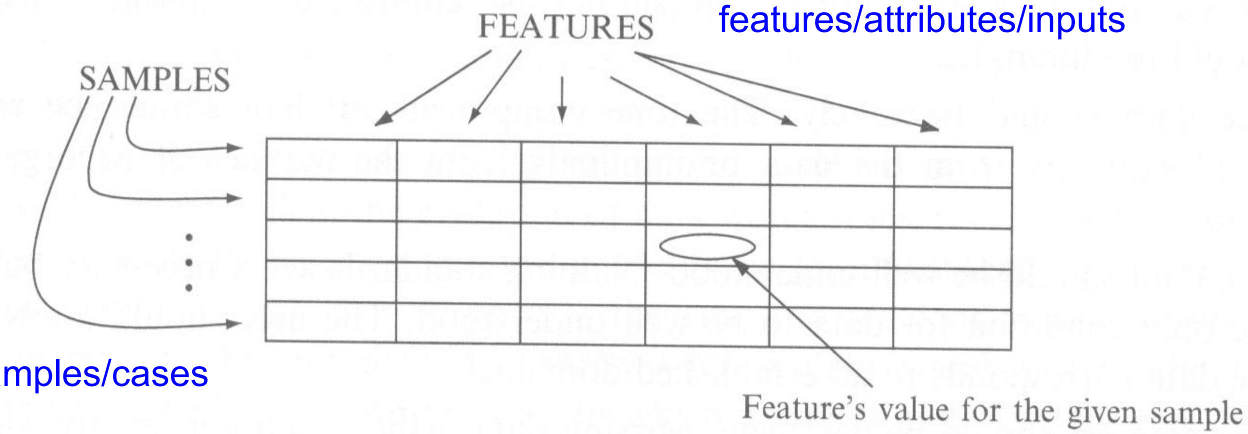
Where Are We Now ?



Data Samples

- Large amounts of samples with different types of features (attributes)
- Each sample is described with several features
  - Different types of values for every feature
    - Numeric: real-value or integer variables
      - Support "order" and "distance" relations
    - Categorical: symbolic variables
      - Support "equal" relation

features/attributes/inputs



samples/examples/cases

Data Samples (cont.)

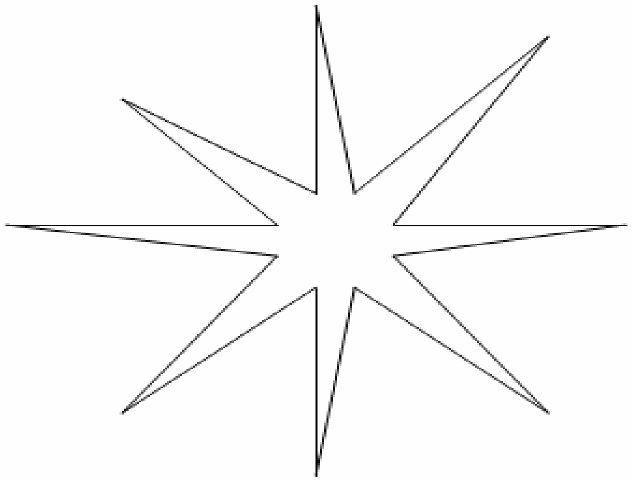
- Another way of classification of variables
  - Continuous variables
    - Also called *quantitative* or *metric* variables
    - Measured using interval or ratio scales
      - Interval: e.g., temperature scale
      - Ratio: e.g., height, length,.. (has an absolute zero point)
  - Discrete variables
    - Also called *qualitative* variables
    - Measured using nonmetric scales (nominal, ordinal)
      - Nominal: e.g., (A,B,C, ...), (1,2,3, ...)
      - Ordinal: e.g., (young, middle-aged, old), (low, middle-class, upper-middle-class, rich), …
    - A special class of discrete variable: periodic variables
      - Weekdays (Monday, Tuesday,..): distance relation exists

Data Samples (cont.)

- Time: one additional dimension of classification of data
  - Static data
    - Attribute values do not change with time
  - Dynamic (temporal) data
    - Attribute values change with time

Curse of Dimensionality

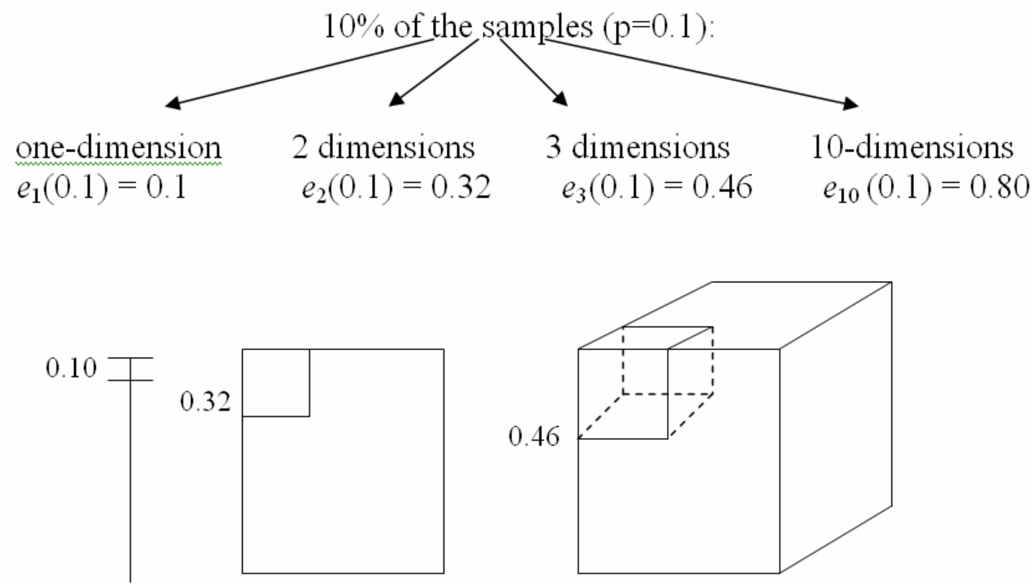
- Data samples are very often high dimensional
  - Extremely large number of measurable features
  - The properties of high dimensional spaces often appear counterintuitive
  - High dimensional spaces have a larger surface area for a given volume
  - Look like a porcupine after visualization



Curse of Dimensionality (cont.)

- Four important properties of high dimensional data
  1. The size of a data set yielding the same density of data points in an *n*-dimensional space increases exponentially with dimensions
  2. A large radius is needed to enclose a fraction of the data points in a high dimensional space

With the same density



$$p = 0.1$$
$$\Rightarrow e_2(0.1) = (0.1)^{1/2} = 0.32$$
$$\Rightarrow e_3(0.1) = (0.1)^{1/3} = 0.46$$
$$\dots$$
$$\Rightarrow e_{10}(0.1) = (0.1)^{1/10} = 0.80$$

radius

$$e_d(p) = p^{1/d}$$

dimensionality

fraction of samples

Curse of Dimensionality (cont.)

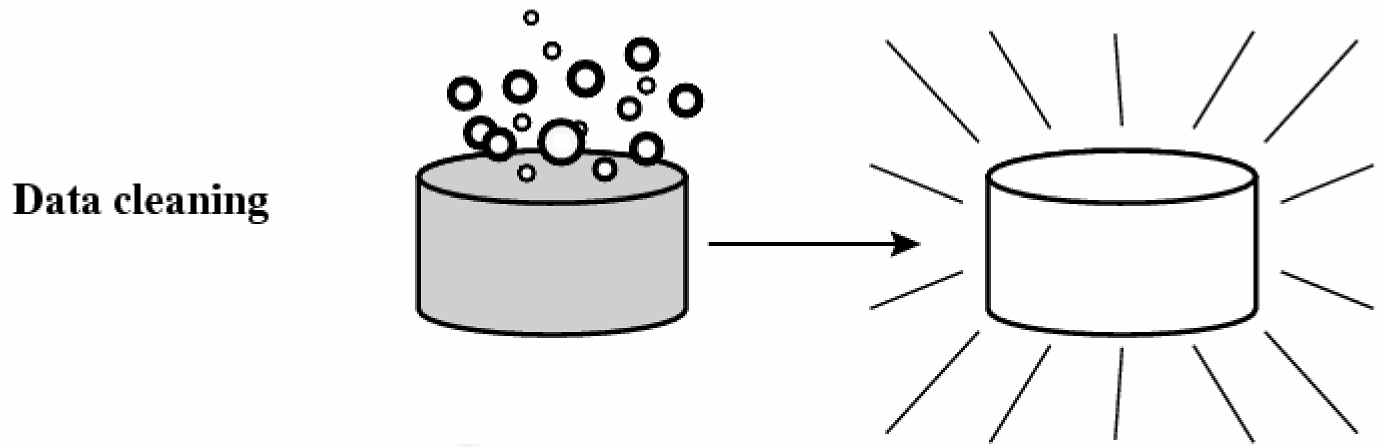
3. Almost every point is closer to an edge than to another sample point in a high dimensional space

4. Almost every point is an outlier. The distance between the prediction point and the center of the classified points increases

With the same number of samples

Central Tasks for Data Preparation

- Organize data into a standard form that is ready for processing by data-mining and other computer-based tools
- Prepare data set that lead to the best data-mining performances



Sources for Messy Data

- Missing Values
  - Values are unavailable
- Misrecording
  - Typically occurs when large volumes of data are processed
- Distortions
  - Interfered by noise when recording data
- Inadequate Sampling
  - Training/test examples are not representative
- ….

Transformation of Raw Data

- Data transformation can involve the following
  - Normalizations
  - Data Smoothing
  - Differences and Ratios (attribute/feature construction)
  - ….

Attention should be paid to data transformation, because relatively simple transformations can sometimes be far more effective for the final performance !

Normalizations

- For data mining methods with examples represented in an *n*-dimensional space and distance computation between points, data normalization may be needed
  – Scaled values to a specific range, e.g., [-1,1] or [0,1]
  – Avoid overweighting those features that have large values (especially for distance measures)

1. Decimal Scaling:
   – Move the decimal point but still preserve most of the original digital value

$$v'(i) = v(i)/10^k$$

$$\text{for small } k \text{ such that } \max(|v'|) < 1$$

The feature value might concentrate upon a small subinterval of the entire range

$$\left.\begin{array}{r}\text{largest} = 455 \\ \text{smallest} = -834\end{array}\right\} \Rightarrow k = 3$$

$$(-0.834 \sim 0.455)$$

$$\left.\begin{array}{r}\text{largest} = 150 \\ \text{smallest} = -10\end{array}\right\} \Rightarrow k = 3$$

$$(-0.01 \sim 0.15)$$

Normalizations (cont.)

2. Min-Max Normalization:

– Normalized to be in [0, 1]

$$v'(i) = \frac{v(i) - \min(v)}{(\max(v) - \min(v))}$$

– Normalized to be in [-1, 1]

$$v'(i) = 2\left[\frac{v(i) - \min(v)}{(\max(v) - \min(v))} - 0.5\right]$$

- The automatic computation of min and max value requires one additional search through the entire data set
- It may be dominated by the outliers
- It will encounter an "out of bounds" error !

Normalizations (cont.)

3. Standard Deviation Normalization

- Also called *z-score* or *zero-mean* normalization
- The values of an attribute are normalized based on the mean and standard deviation of it
- Mean and standard deviation are first computed for the entire data set

$$v'(i) = \frac{v(i) - mean(v)}{sd(v)}$$

$$\bar{v} = mean(v) = \frac{\sum v}{n_v}$$

$$\sigma_v = sd(v) = \sqrt{\frac{\sum (v - \bar{v})^2}{n_v - 1}}$$

?

- E.g., the initial set of values of the attribute $v = \{1, 2, 3\}$ has

$$mean(v) = 1,\ sd(v) = 1 \text{ and new set of } v' = \{-1, 0, 1\}$$

Normalizations (cont.)

- An identical normalization should be applied both on the observed (training) and future (new) data
  - The normalization parameters must be saved along with a solution

Data Smoothing

- Minor differences between the values of a feature (attribute) are not significant and may degrade the performance of data mining
  - They may be caused by noises

- Reduce the number of distinct values for a feature
  - E.g., round the values to the given precision

$$F = \{0.93, 1.01, 1.001, 3.02, 2.99, 5.03, 5.01, 4.98\}$$
$$\Rightarrow F_{smoothed} = \{1.0, 1.0, 1.0, 3.0, 3.0, 5.0, 5.0, 5.0\}$$

  - The dimensionality of the data space (number of distinct examples) is also reduced at the same time

Differences and Ratios

- Can be viewed as a kind of attribute/feature construction
  - New attributes are constructed from the given attributes
  - Can discover the missing information about the relationships between data attributes
  - Can be applied to the *input* and *output* features for data mining

- E.g.,
  1. Difference
     - E.g., "$s(t+1)$ - $s(t)$", relative moves for control setting
  2. Ratio
     - E.g.,"$s(t+1)$ / $s(t)$", levels of increase or decrease
     - E.g., Body-Mass Index (BMI) $Weight(Kg) \big/ Height(m^2)$

Missing Data

- In real-world application, the subset of samples or future cases with complete data may be relatively small
  - Some data mining methods accept missing values
  - Others require all values be available
    - Try to drop the samples or fill in the missing attribute values in during data preparation

Missing Data (cont.)

- Two major ways to deal with missing data (values)
  1. Reduce the data set and eliminate all samples with missing values
     – If large data set available and only a small portion of data with missing values

  2. Find values for missing data
     a. Domain experts examine and enter reasonable, probable, and expected values for the missing data

     b. Automatically replace missing values with some constants

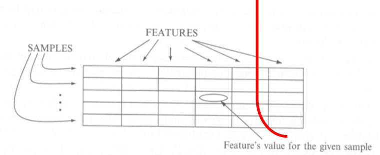
        b.1 Replace a missing value with a single global constant

        b.2 Replace a missing value with its feature mean

        b.3 Replace a missing value with its feature mean for the given class (if class labeling information available)

        b.4 Replace a missing value with the most probable value (e.g., according to the values of other attributes of the present data)

will bias the data



Missing Data (cont.)

- The replaced value(s) (especially for b.1~b.3) will homogenize the cases / samples with missing values into an artificial class

- Other solutions

  1. "Don't Care"

     - Interpret missing values as "don't care" values

     $$\vec{x} = \langle 1,\ ?,\ 3 \rangle, \text{with feature values in domain } [0,1,2,3,4]$$

     $$\Rightarrow \vec{x}_1 = \langle 1,\ 0,\ 3 \rangle, \vec{x}_2 = \langle 1,\ 1,\ 3 \rangle, \vec{x}_3 = \langle 1,\ 2,\ 3 \rangle, \vec{x}_4 = \langle 1,\ 3,\ 3 \rangle, \vec{x}_5 = \langle 1,\ 4,\ 3 \rangle$$

     - A explosion of artificial samples being generated !

  2. Generate multiple solutions of data-mining with and without missing-value features and then analyze and interpret them !

$$\begin{array}{ll} A_1, B_1, C_1 & \\ A_2, B_2, C_2 & \\ \ldots & \Rightarrow (A, B, ?), (A, ?, C), (?, B, C) \\ A_N, B_N, C_N & \end{array}$$

Time-Dependent Data

- Time-dependent relationships may exist in specific features of data samples
  - E.g., “temperature reading” and speech are a univariate time series, and video is a multivariate time series

$$X = \{t(0), t(1), t(2), t(3), t(4), t(5), t(6), t(7), t(8), t(9), t(10)\}$$

- Forecast or predict $t(n+1)$ from previous values of the feature

TABLE 2.1 Transformation of Time Series to standard tabular form (window = 5)

Sample	WINDOW					Next Value
	M1	M2	M3	M4	M5	
1	t(0)	t(1)	t(2)	t(3)	t(4)	t(5)
2	t(1)	t(2)	t(3)	t(4)	t(5)	t(6)
3	t(2)	t(3)	t(4)	t(5)	t(6)	t(7)
4	t(3)	t(4)	t(5)	t(6)	t(7)	t(8)
5	t(4)	t(5)	t(6)	t(7)	t(8)	t(9)
6	t(5)	t(6)	t(7)	t(8)	t(9)	t(10)

Time-Dependent Data (cont.)

- Forecast or predict *t*(*n*+*j*) from previous values of the feature

TABLE 2.2 **Time-series samples in standard tabular form (window = 5) with postponed predictions (j = 3)**

Sample	WINDOW					Next Value
	M1	M2	M3	M4	M5	
1	t(0)	t(1)	t(2)	t(3)	t(4)	t(7)
2	t(1)	t(2)	t(3)	t(4)	t(5)	t(8)
3	t(2)	t(3)	t(4)	t(5)	t(6)	t(9)
4	t(3)	t(4)	t(5)	t(6)	t(7)	t(10)

- As mentioned earlier, forecast or predict the differences or ratios of attribute values
  - $t(n+1) - t(n)$
  - $t(n+1) / t(n)$

Time-Dependent Data (cont.)

- "Moving Averages" (MA)– a single average summarizes the most *m* feature values for each case at each time moment *i*
  - Reduce the random variation and noise components

$$MA(i, M) = \frac{1}{M} \cdot \sum_{j=i-M+1}^{i} t(j),$$

$$t(j): \text{noisy data}, \hat{t}(j): \text{clean data}$$

$$t(j) = \hat{t}(j) + error, \quad \text{error is assumed to be a constant}$$

$$\Rightarrow MA(i, M) = \frac{1}{M} \cdot \sum_{j=i-M+1}^{i} t(j) = mean(j) + error$$

$$, where \quad mean(j) = \sum_{j=i-M+1}^{i} \hat{t}(j)$$

$$\Rightarrow t(j) - MA(i, M) = \hat{t}(j) - mean(j)$$

Time-Dependent Data (cont.)

- "Exponential Moving Averages" (EMA) – give more weight to the most recent time periods

$$EMA(i, M) = p \cdot t(i) + (1 - p) \cdot EMA(i - 1, M - 1)$$
$$EMA(i, 1) = t(i)$$

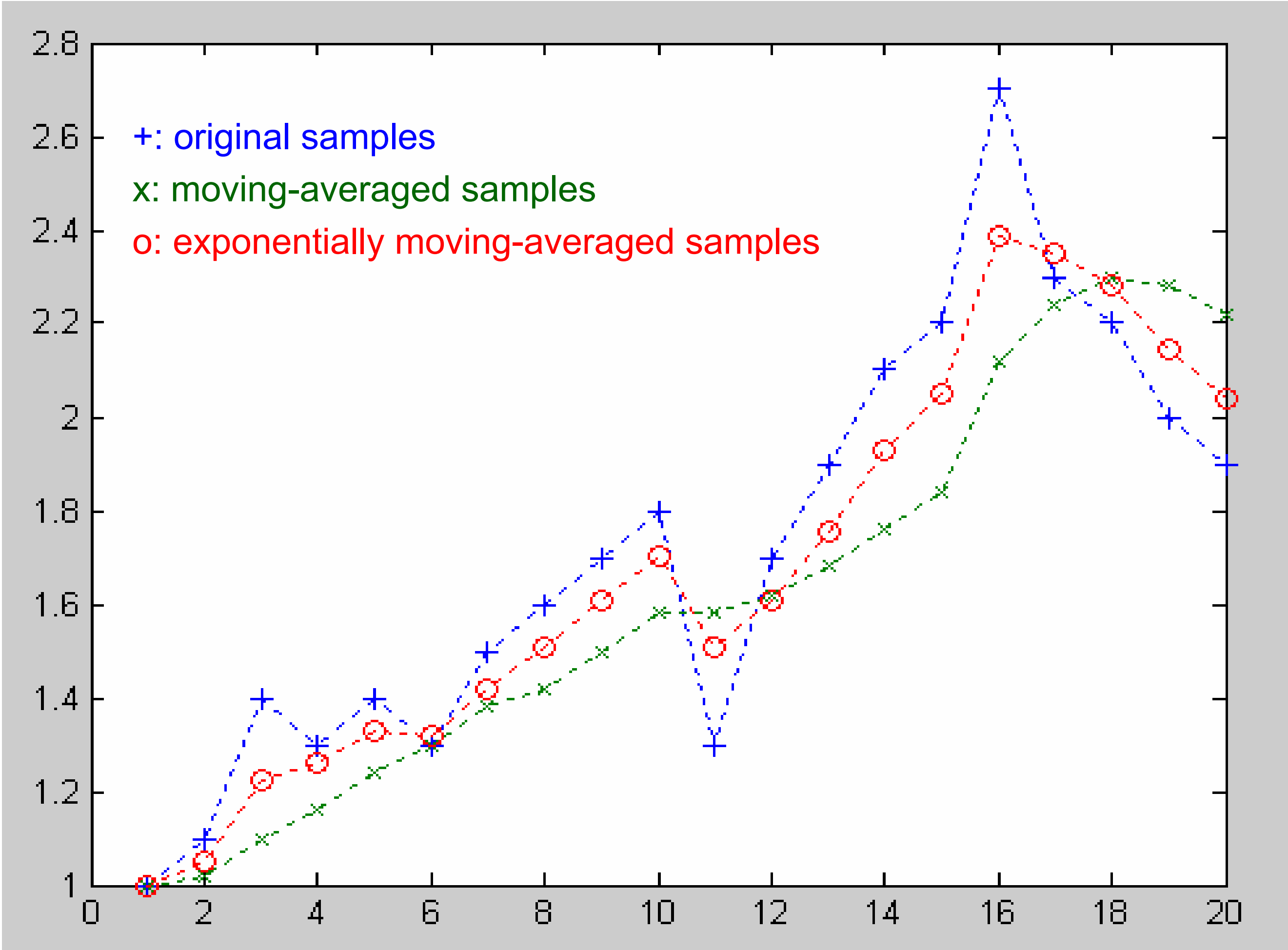
$$\text{if } p = 0.5$$
$$EMA(i, 2) = 0.5 \cdot t(i) + 0.5 \cdot EMA(i - 1, 1)$$
$$\begin{aligned} EMA(i, 3) &= 0.5 \cdot t(i) + 0.5 \cdot EMA(i - 1, 2) \\ &= 0.5 \cdot t(i) + 0.5 \cdot [0.5 \cdot t(i - 1) + 0.5 \cdot EMA(i - 2, 1)] \\ &= 0.5 \cdot t(i) + 0.5 \cdot [0.5 \cdot t(i - 1) + 0.5 \cdot t(i - 2)] \end{aligned}$$

$$t(i) \longrightarrow \boxed{\text{System}} \longrightarrow \hat{t}(i)$$

Causal or Noncausal Filter

Time-Dependent Data (cont.)

X=[1.0 1.1 1.4 1.3 1.4 1.3 1.5 1.6 1.7 1.8 1.3 1.7 1.9 2.1 2.2 2.7 2.3 2.2 2.0 1.9];



Time-Dependent Data (cont.)

- Appendix: *MATLab Codes* for Moving Averages (MA)

```
W=1:20;
X=[1.0 1.1 1.4 1.3 1.4 1.3 1.5 1.6 1.7 1.8 1.3 1.7 1.9 2.1 2.2 2.7 2.3 2.2 2.0 1.9];
U=zeros(5,20);

for M=0:10
 for i=1:20
  sum=0.0;
  for m=0:M
   if i-m>0
    sum=sum+X(i-m);
   else
    sum=sum+X(1);
   end
  end
  U(M+1,i)=sum/(M+1);
 end
end
plot(W,U(1,:),':+',W,U(5,:),':x');
```

Homework-1: Data Preparation

- Exponential Moving Averages (EMA)

  X=[1.0 1.1 1.4 1.3 1.4 1.3 1.5 1.6 1.7 1.8 1.3 1.7 1.9 2.1 2.2 2.7 2.3 2.2 2.0 1.9];

$$EMA(i,m) = p \cdot t(i) + (1-p) \cdot EMA(i-1, m-1)$$
$$EMA(i,1) = t(i)$$

  - Try out different settings of *m* and *p*
  - Discuss on the results you observed
  - Discuss the applications in which you would prefer to use exponential moving averages (EMA) instead of moving averages (MA)

- Due date: 2004/3/17

Time-Dependent Data (cont.)

- Example: multivariate time series

spatial information

Temporal information

Time	a	b
1	5	117
2	8	113
3	4	116
4	9	118
5	10	119
6	12	120

a) Initial time-dependent data

Sample	a(n-2)	a(n-1)	a(n)	b(n-2)	b(n-1)	b(n)
1	5	8	4	117	113	116
2	8	4	9	113	116	118
3	4	9	8	116	118	119
4	9	10	12	118	119	120

b) Samples prepared for data mining with time window = 3

date reduction

cases ?

feature ?

values ?

FIGURE 2.3 Tabulation of time-dependent features a and b

High dimensions of data generated during the transformation of time-dependent can be reduced through "data reduction"

Outlier Analysis

- Outliers
  - Data samples that do not comply with the general behavior of the data model and are significantly different or inconsistent with the remaining set of data
  - E.g., a person's age is "-999", the number of children for one person is "25", …. (typographical errors/typos)

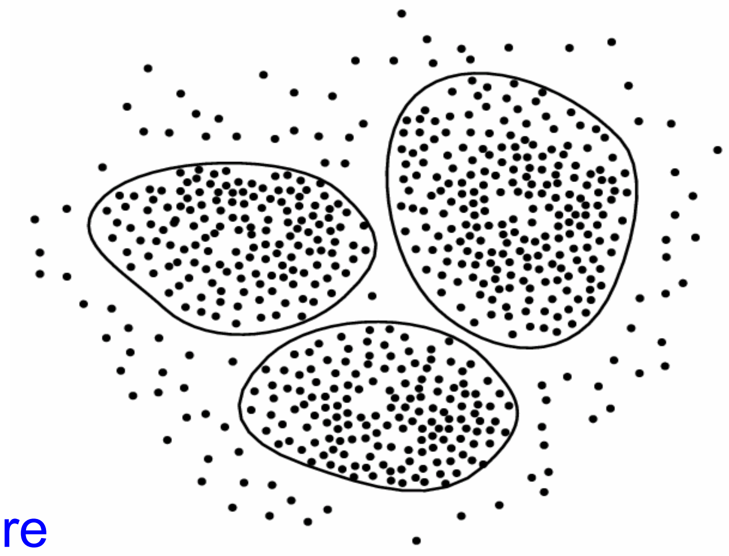
- Many data-mining algorithms try to minimize the influence of outliers or eliminate them all together
  - However, it could result in the loss of important hidden information
  - "one person's noise could be another person's signal", e.g., outliers may indicate abnormal activity

Outlier Analysis (cont.)

- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis

Outlier Analysis (cont.)

- Outlier detection/mining
  - Given a set of *n* samples, and *k*, the expected number of outliers, find the top *k* samples that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data
  - Can be viewed as two subproblems
    - Define what can be considered as inconsistent in a given data set
      - Nontrivial
    - Find an efficient method to mine the outliers so defined
      - Three methods introduced here



Visual detection of outlier ?

Outlier Analysis (cont.)

1. Statistical-based Outlier Detection
   - Assume a distribution or probability model for the given data set and then identifies outliers with respect to the model using a *discordance* test
     - Data distribution is given/assumed (e.g., normal distribution)
     - Distribution parameters: mean, variance
       - Threshold value as a function of variance

$$Age = \{3, 56, 23, 39, 156, 52, 41, 22, 9, 28, 139, 31,$$
$$55, 20, -67, 37, 11, 55, 45, 37\}$$

$$Mean = 39.9$$

$$Standard\ deviation = 45.65$$

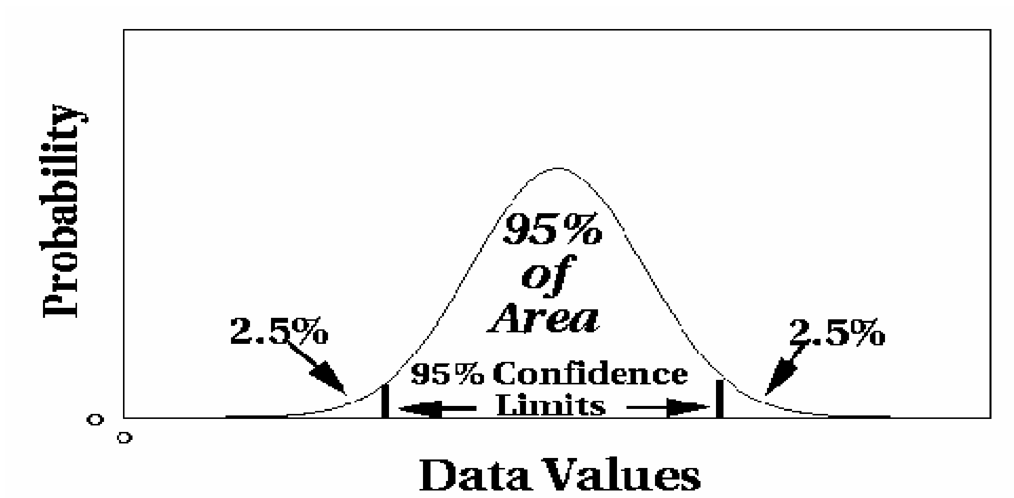
$$Threshold = Mean \pm 2 \times Standard\ deviation$$

$[-54., \ 131.2] \Rightarrow [0, \ 131.2]$ Age is always greater than zero !

$$\Rightarrow outliers : 156, 139, -67$$

Outlier Analysis (cont.)

1. Statistical-based Outlier Detection (cont.)
   – Drawbacks
     - Most tests are for single attribute
     - In many cases, data distribution may not be known



Outlier Analysis (cont.)

2. Distance-based Outlier Detection

– A sample s_i in a data S is an outlier if at least a fraction p of the objects in S lies at a distance greater than d, denoted as $DB<p, d>$

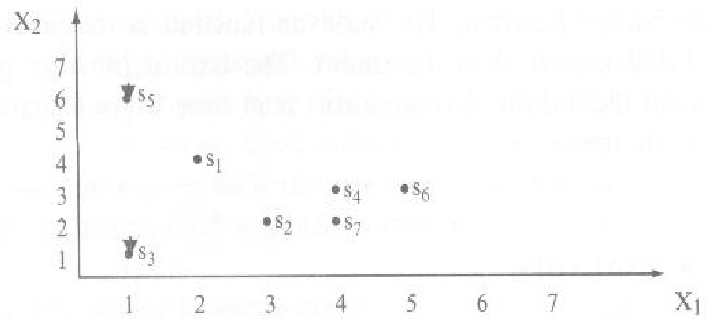


FIGURE 2.4 Visualization of two-dimensional data set for outlier detection

TABLE 2.3 Table of distances for data set S

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1		2.236	3.162	2.236	2.236	3.162	2.828
s_2			2.236	1.414	4.472	2.236	1.000
s_3				3.605	5.000	4.472	3.162
s_4					4.242	1.000	1.000
s_5						5.000	5.000
s_6							1.414

- If $DB<p, d>=DB<4, 3>$

$$d = \left[(x_1 - x_1)^2 + (y_1 - y_1)^2 \right]^{1/2}$$

– Outliers: s_3, s_5

the distance greater then d for each given point in S

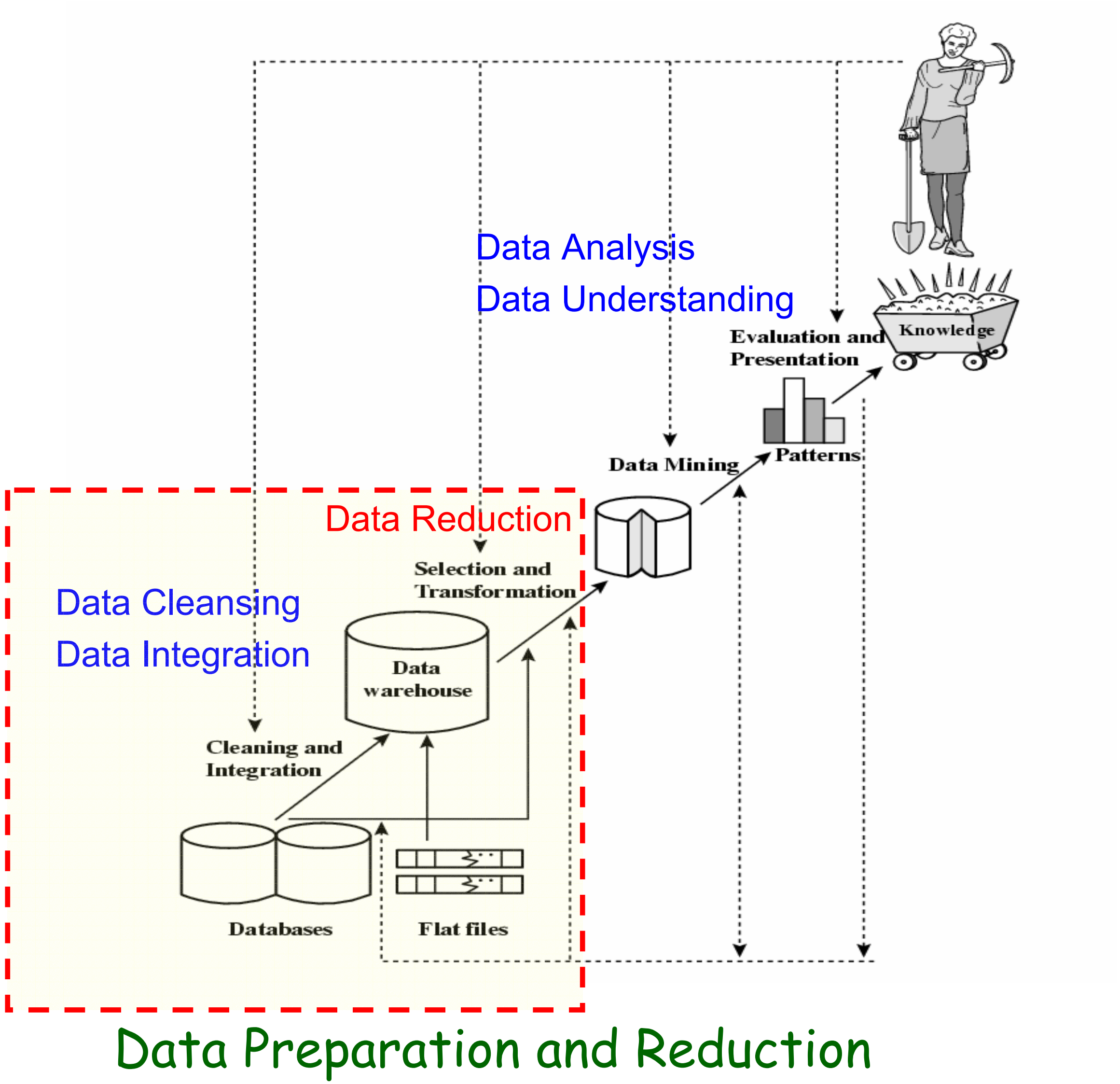
Sample	p
s_1	2
s_2	1
s_3	5
s_4	2
s_5	5
s_6	3

Outlier Analysis (cont.)

3. Deviation-based Outlier Detection
   - Define the basic characteristics of the sample set, and all samples that deviate from these characteristics are outliers

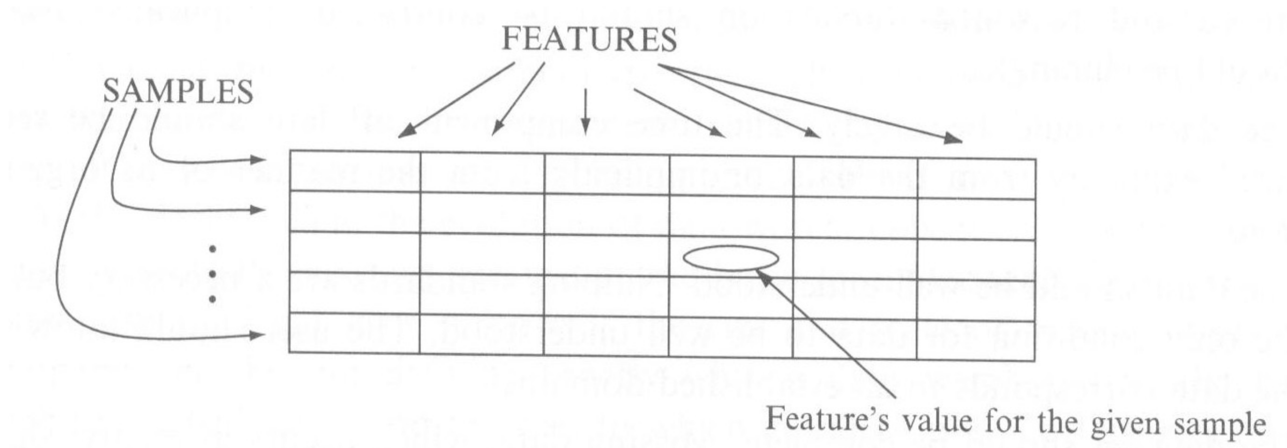
   - The "sequence exception technique"
     - Based on a dissimilarity function, e.g., variance $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$
     - Find the smallest subset of samples whose removal results in the greatest reduction of the dissimilarity function for the residual set (a NP-hard problem)

Where Are We Now ?



Introduction to Data Reduction

- Three dimensions of data sets
  - Rows (cases, samples, examples)
  - Columns (features)
  - *Values* of the features
- We hope that the final reduction doesn't reduce the quality of results, instead the results of data mining can be even improved



Introduction to Data Reduction (cont.)

- Three basic operations in data reduction
  - Delete a column
  - Delete a row
  - Reduce the number of values in a column

Preserve the characteristic of original data
Delete the nonessential data

- Gains or losses with data reduction
  - Computing time
    - Tradeoff existed for preprocessing and data-mining phases
  - Predictive/descriptive accuracy
    - Faster and more accurate model estimation
  - Representation of the data-mining model
    - Simplicity of model representation (model can be better understood)
      - Tradeoff between simplicity and accuracy

Introduction to Data Reduction (cont.)

- Recommended characteristics of data-reduction algorithms
  - Measure quality
    - Quality of approximated results using a reduced data set can be determined precisely
  - Recognizable quality
    - Quality of approximated results can be determined at preprocessing phrase
  - Monotonicity
    - Iterative, and monotonically decreasing in time and quality
  - Consistency
    - Quality of approximated results is correlated with computation time and input data quality
  - Diminishing returns (Convergence)
    - Significant improvement in early iterations and which diminished over time
  - Interruptability
    - Can be stopped at any time and provide some answers
  - Preemptability
    - Can be suspended and resumed with minimal overhead

Feature Reduction

- Also called “column reduction”
  - Also have the side effect of case reduction

- Two standard tasks for producing a reduced feature set
  1. Feature selection
     - Objective: find a subset of features with performances comparable to the full set of features
  2. Feature composition (do not discuss it here!)
     - New features/attributes are constructed from the given/old features/attributes and those given ones are discarded later on !
     - For example $\frac{Weight(Kg)}{Height(m^2)}$
       - » Body-Mass Index (BMI)
       - » New features/dimensions retained after principal component analysis (PCA)
     - Interdisciplinary approaches and domain knowledge

Feature selection (cont.)

- Select a subset of the features based domain knowledge and data-mining goals

- Can be viewed as a search problem
  - Manual or automated

Feature selection as searching

$\{A_1, A_2, A_3\}$

$\Rightarrow$ {0,0,0}, {1,0,0}, {0,1,0},…, {1,1,1}

1: with the feature

0: without the feature

  - Find optimal or near-optimal solutions (subsets of features) ?

Feature selection (cont.)

- Methods can be classified as
  a. Feature ranking algorithms
  b. Minimum subset algorithms

  Need a feature-evaluation scheme

  - Button-up: starts with an empty set and fill it in by choosing the most relevant features from the initial set of features
  - Top-down: begin with a full set of original features and remove one-by-one those that are irrelevant

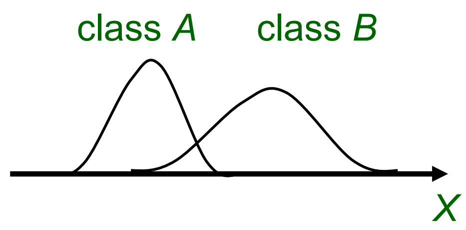
- Methods also can be classified as
  a. Supervised : Use class label information
  b. Unsupervised: Do not use class label information

Supervised Feature Selection (cont.)

- Method I: Simply based on comparison of means and variances
  - Assume the distribution of the feature forms a normal curve
  - Feature means of different categories/classes are normalized and then compared
    - If means are far apart $\rightarrow$ interest in a feature increases
    - If means are indistinguishable $\rightarrow$ interest wanes in that feature

$$SE\left(X_A - X_B\right) = \sqrt{\frac{var\left(X_A\right)}{n_{X,A}} + \frac{var\left(X_B\right)}{n_{X,B}}}$$

$$TEST: \frac{\left|mean\left(X_A\right) - mean\left(X_B\right)\right|}{SE\left(X_A - X_B\right)} > \text{threshold - value}$$



  - Simple but effective
  - Without taking into consideration relationship to other features
    - Assume features are independent of each other

Supervised Feature Selection (cont.)

- Example: threshold - value = 0.5

$$\bar{x} = mean\,(x) = \frac{\sum x}{n_x}$$

$$\text{var}\,(x) = \frac{\sum (x - \bar{x})^2}{n_x - 1}$$

TABLE 3.1 Dataset with three features

X	Y	C
0.3	0.7	A
0.2	0.9	B
0.6	0.6	A
0.5	0.5	A
0.7	0.7	B
0.4	0.9	B

$$X_A = \{0.3, 0.6, 0.5\},\ n_{X,A} = 3$$

$$X_B = \{0.2, 0.7, 0.4\},\ n_{X,B} = 3$$

$$Y_A = \{0.7, 0.6, 0.5\},\ n_{Y,A} = 3$$

$$Y_B = \{0.9, 0.7, 0.9\},\ n_{Y,B} = 3$$

$$SE(X_A - X_B) = \sqrt{\frac{var(X_A)}{n_{X,A}} + \frac{var(X_B)}{n_{X,B}}} = \sqrt{\frac{0.0233}{3} + \frac{0.6333}{3}} = 0.4678$$

$$\frac{|mean(X_A) - mean(X_B)|}{SE(X_A - X_B)} = \frac{|0.4667 - 0.4333|}{0.4678} = 0.0735 < 0.5$$

$$SE(Y_A - Y_B) = \sqrt{\frac{var(Y_A)}{n_{Y,A}} + \frac{var(Y_B)}{n_{Y,B}}} = \sqrt{\frac{0.010}{3} + \frac{0.0133}{3}} = 0.0875$$

$$\frac{|mean(Y_A) - mean(Y_B)|}{SE(Y_A - Y_B)} = \frac{|0.600 - 0.8333|}{0.0875} = 2.6667 > 0.5$$

Supervised Feature Selection (cont.)

- Example: (cont.)
  - X is a candidate for feature reduction
  - Y is significantly above the threshold value $\rightarrow$ Y has the potential to be a distinguishing feature between two classes
  - How to extend such a method to K-class problems
    - $k(k\text{-}1)/2$ pairwise comparisons are needed ?

Supervised Feature Selection (cont.)

- Method II: Features examined collectively instead of independently, additional information can be obtained

$C : m \times m$ covariance matrix, each entry $C_{i,j}$ stands for the correlation between two features i, j ⇦ *m* features are selected

$$C_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\big(v(k,i)-m(i)\big)\cdot\big(v(k,j)-m(j)\big)$$

number of samples

$v(k,i)$: the value of feature i of sample k

$m(i)$: mean of feature i

$$DM = (M_1 - M_2)(C_1 + C_2)^{-1}(M_1 - M_2)^T$$ ⇦ distance measure for multivariate variables

- M1, M2, C1, C2, are respectively mean vectors and covariance matrices for class 1 and class 2
- A subset set of features are selected for this measure (maximizing *DM*)
    - All subsets should be evaluated ! (how to do ? a combinatorial problem)

Review: Entropy

- Three interpretations for quantity of information
  1. The amount of **uncertainty** before seeing an event
  2. The amount of **surprise** when seeing an event
  3. The amount of **information** after seeing an event

- The definition of information:

$$\textit{define} \quad 0\log_2 0 = 0$$

$$I(x_i) = \log_2 \frac{1}{P(x_i)} = -\log_2 P(x_i)$$

  – $P(x_i)$ the probability of an event x_i

- Entropy: the average amount of information

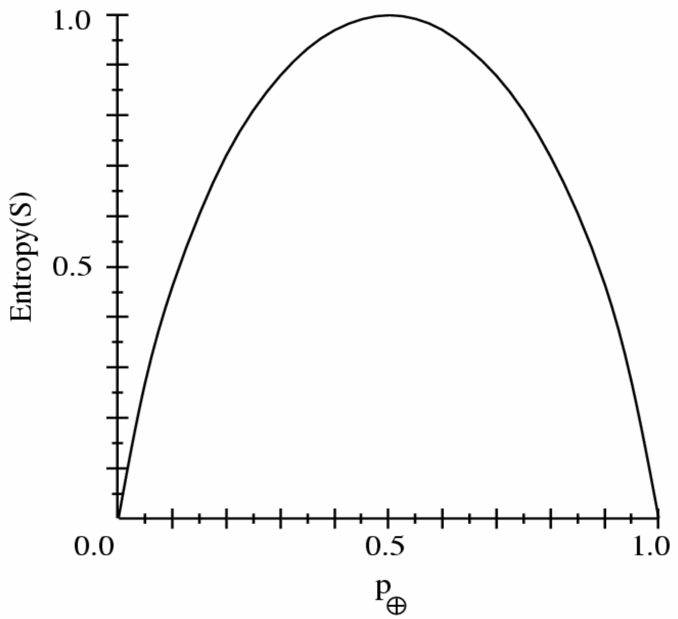
$$H(X) = E[I(X)]_X = E[-\log_2 P(x_i)]_X = \sum_{x_i} -P(x_i) \cdot \log_2 P(x_i)$$

where $X = \{x_1, x_2, \ldots, x_i, \ldots\}$

  – Have maximum value when the probability (mass) function is a uniform distribution

Review: Entropy (cont.)

- For Boolean classification (0 or 1)



$$Entropy\,(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

-相同機率分佈下(如Uniform)，event個數越多，entropy越大
(½, ½) → 1 , (¼, ¼, ¼, ¼) → 2
-event個數固定情況下，機率分佈越平均(如Uniform) ， entropy越大

- Entropy can be expressed as the minimum number of bits of information needed to encode the classification of an arbitrary number of examples
  - If c classes are generated, the maximum of Entropy can be

$$Entropy\,(X) = \log_2 c$$

Review: Entropy (cont.)

- Illustrative Example
  - Discriminate speech portions form non-speech portions for Voice Activity Detection (VAD)
    - Speech has clear formants and entropies of such spectra will be slow
    - Non-speech has flatter spectra and the associated entropies should be higher

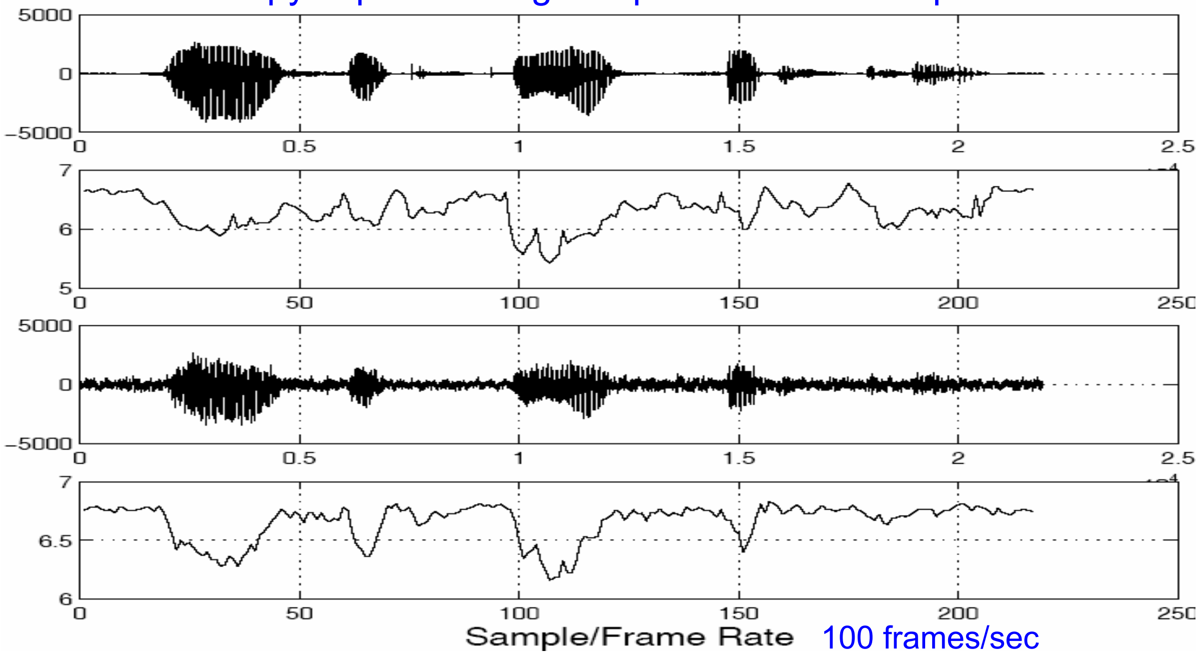
i-th frequency component of spectrum

$$x_i = \frac{X_i}{\sum_{j=1}^{N} X_j}$$

$$H = -\sum_{i=1}^{N} x_i \cdot \log_2 x_i$$

probability mass

Entropy captures the gross peakiness of the spectrum



waveform

entropy

noisy waveform

entropy

100 frames/sec

Unsupervised Feature Selection

- Method I: Entropy measure for ranking features
  - Assumptions
    - All samples are given as vectors of feature values without any categorical information
    - The removal of an irrelevant (redundant) feature may not change the basic characteristics of the data set
      - basic characteristics $\rightarrow$ the similarity measure between any pair of samples
    - Use entropy to observe the change of global information before and after removal of a specific feature
      - Higher entropy for disordered configurations
      - Less entropy for ordered configurations
  - Rank features by iteratively (gradually) removing the least important feature in maintaining the configuration order

Unsupervised Feature Selection (cont.)

- Method I: Entropy measure for ranking features (cont.)
  - Distance measure between two samples x_i and x_j

$$D_{ij} = \left[\sum_{k=1}^{n} \left(\left(x_{ik} - x_{jk} \right) / \left(\max_k - \min_k \right) \right)^2 \right]^{1/2}$$

number of features

  - Change the distance measure to likelihood of proximity/similarity using exponential operator (function)

$$S_{ij} = \exp\left(-\alpha D_{ij}\right)$$

α is simply set to 0.5 or is set as $-(\ln 0.5)/D_{average}$

ranging between 0 ~1

  - $S_{ij} \approx 1$: x_i and x_j is very similar
  - $S_{ij} \approx 0$: x_i and x_j is very dissimilar

- For Categorical (nominal/nonmetric) features
  - Hamming distance

ranging between 0 ~1

$$S_{ij} = \left(\sum_{k=1}^{n} \left| x_{ik} = x_{jk} \right| \right) / n,$$

$$\left| x_{ik} = x_{jk} \right| = 1 \text{ if } x_{ik} = x_{jk} \text{ and } 0 \text{ otherwise}$$

Unsupervised Feature Selection (cont.)

– Use entropy to monitor the changes in proximity between any sample pair in the data set (data set with size *N*)

$$E = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} H_{i,j}$$

$$- \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(S_{ij} \log S_{ij} + \left(1 - S_{ij}\right) \log \left(1 - S_{ij}\right)\right)$$

Likelihood of being similar

Likelihood of being dissimilar

– Example: a simple data set with three categorical features

Sample	F_1	F_2	F_3
R_1	A	X	1
R_2	B	Y	2
R_3	C	Y	2
R_4	B	X	1
R_5	C	Z	3

Data set

	R_1	R_2	R_3	R_4	R_5
R_1		0/3	0/3	2/3	0/3
R_2			2/3	1/3	0/3
R_3				0/3	1/3
R_4					0/3

Table of similarity measures

$$\text{e.g.,} H_{1,2} = H_{2,1} = -\left[(0/3)\log(0/3) + (3/3)\log(3/3)\right]$$

$$H_{1,4} = H_{4,1} = -\left[(2/3)\log(2/3) + (1/3)\log(1/3)\right]$$

Unsupervised Feature Selection (cont.)

- Method I: Entropy measure for ranking features (cont.)
  - Algorithm
    1. Start with the initial set of features F
    2. For each feature f in F, remove f from F and obtain a subset F_f . Find the difference between entropy for F and F_f

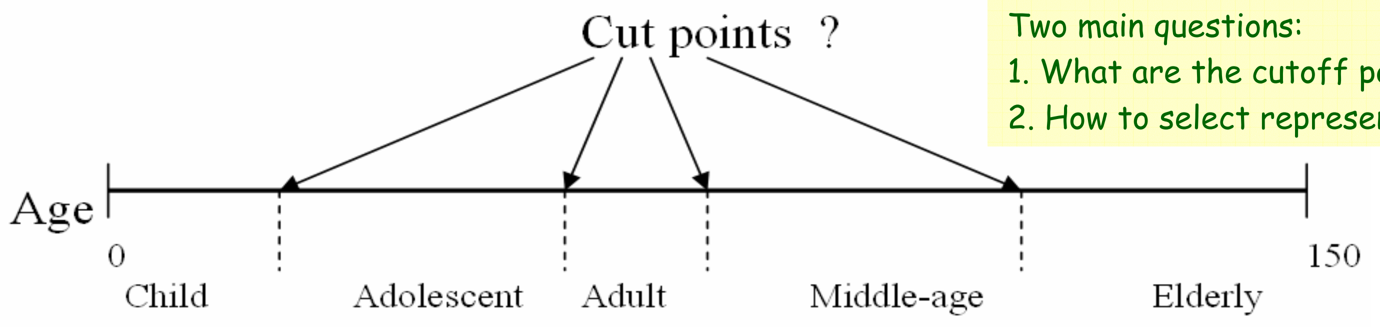
$$\left| E_F - E_{F-f} \right|$$

    3. Find f_k such that its removal makes the entropy difference is minimum, check if the difference is less then the threshold
    4. If so, update the feature set as $F' = F - f_k$ and repeat steps 2~4 until only one feature is retained; otherwise, stop !

Disadvantage: the computational complexity is higher !

Value Reduction

- Also called Feature Discretization
- Goal: discretize the value of continuous features into a small number of intervals, where each interval is mapped to a discrete symbol
  - Simplify the tasks of data description and understanding
  - E.g., a person's age can be ranged from 0 ~ 150
    - Classified into categorical segments: "child, adolescent, adult, middle age, elderly"



Two main questions:
1. What are the cutoff points?
2. How to select representatives of intervals

Discretization of the *age* feature

Unsupervised Value Reduction

- Method I: Simple data reduction (value smoothing)
  - Also called number approximation by rounding
  - Reduce the number of distinct values for a feature
  - E.g., round the values to the given precision

$$f = \{0.93, 1.01, 1.001, 3.02, 2.99, 5.03, 5.01, 4.98\}$$
$$\Rightarrow f_{smoothed} = \{1.0, 1.0, 1.0, 3.0, 3.0, 5.0, 5.0, 5.0\}$$

  - Properties
    - Each feature is smoothed independently of other features
    - Performed only once without iterations
    - The number of data samples (cases) may be also reduced at the same time

Unsupervised Value Reduction (cont.)

- Method II: Placing the value in bins
  - Order the numeric values using great-than or less-than operators
  - Partition the ordered value list into groups with close values
    - Also, these bins have close number of elements
  - All values in a bin is merged into a single concept represented by a single value, for example:
    - Mean or median/mode of the bin's value
    - The closest boundaries of each bin

$$f = \{3,2,1,5,4,3,1,7,5,3\}$$

$$\overset{ordering}{\Rightarrow} \{1,1,2,3,3,3,4,5,5,7\}$$

Based on what criterion ?

$$\overset{splitting}{\Rightarrow} \{\underset{\text{BIN 1}}{1,1,2} \quad \underset{\text{BIN 2}}{3,3,3} \quad \underset{\text{BIN 3}}{4,5,5,7}\}$$

Smoothing based on mean values $\Rightarrow \{\underset{\text{BIN 1}}{1.33,1.33,1.33} \quad \underset{\text{BIN2}}{3,3,3} \quad \underset{\text{BIN 3}}{5.25,5.25,5.25,5.25}\}$

Smoothing based on bin modes $\Rightarrow \{\underset{\text{BIN 1}}{1,1,1} \quad \underset{\text{BIN2}}{3,3,3} \quad \underset{\text{BIN 3}}{5,5,5,5}\}$

Smoothing based on boundary values $\Rightarrow \{\underset{\text{BIN 1}}{1,1,2} \quad \underset{\text{BIN2}}{3,3,3} \quad \underset{\text{BIN 3}}{\boxed{4,4,7,7}}\}$

replaced by the closest of the boundary values

Unsupervised Value Reduction (cont.)

- Method II: Placing the value in bins (cont.)
  - How to determine the optimal selection of *k* bins
    - Criterion: minimize the average distance of a value from its bin mean or median
      - Squared distance for a bin mean
      - Absolute distance for a bin median
    - Algorithm
      1. Sort all values for a given feature
      2. Assign approximately equal numbers of sorted adjacent value (v_i) to each bin, the number of bin is given in advance
      3. Move a border element v_i from one bin to the next (or previous) when that will reduce the global distance error (ER)

Unsupervised Value Reduction (cont.)

- Method II: Placing the value in bins (cont.)
  - Example

$$f = \{5,1,8,2,2,9,2,1,8,6\}$$

$$\overset{ordering}{\Rightarrow} \quad \{1,1,2,2,2,5,6,8,8,9\}$$

$$\overset{splitting\,/\,Initializing}{\Rightarrow} \quad \{\underbrace{1,1,2}_{\text{BIN 1}} \quad \underbrace{2,2,5}_{\text{BIN 2}} \quad \underbrace{6,8,8,9}_{\text{BIN 3}}\}$$

Absolute distance to bin modes
$ER = (0+0+1)+(0+0+3)+(2+0+0+1)=7$

....

$$\Rightarrow \quad \{\underbrace{1,1,2,2,2}_{\text{BIN 1}} \quad \underbrace{5,6}_{\text{BIN 2}} \quad \underbrace{8,8,9}_{\text{BIN 3}}\}$$

Absolute distance to bin modes
$ER = (1+1+0+0+0)+(0+1)+(0+0+1)=4$

$$\Rightarrow \text{corresponding modes}\{2,5,8\}$$

In real-world applications, the number of distinct values is controlled to be 50 ~ 100

Review: Chi-Square Test

- A non-parametric test of statistical significance for bivariate tabular analysis, which can provides degree of confidence in accepting or rejecting an hypothesis
  - E.g. (1), collocations in linguistics

dependent variable/Categories

Independent variable

2x2 contingency table

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

A 2-by-2 table showing the dependence of occurrences of *new* and *companies*. There are 8 occurrences of *new companies* in the corpus, 4667 bigrams where the second word is *companies*, but the first word is not *new*, 15,820 bigrams with the first word *new* and a second word different from *companies*, and 14,287,181 bigrams that contain neither word in the appropriate position.

- Are "new" and "company" independent ?
  - Values of the independent variable has effect on the dependent variable?

Review: Chi-Square Test (cont.)

– E.g. (2), behavior analyses in sociology

Male and Female Footwear Preferences

dependent variable/Categorie *j*

Independent variable *i*

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

2x5 contingency table

- Biological sex and footwear preferences are independent ?
  - Values of the independent variable has effect on the dependent variable?

Ref: http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html

Review: Chi-Square Test (cont.)

- Null Hypothesis
  - In e.g. (2), biological sex and footwear preferences are independent

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male	6	17	13	9	5	50
Female	13	5	7	16	9	50
Total	19	22	20	25	14	100

$$P(male, Sandals) \overset{?}{=} P(male)P(Sandals)$$

$$\Rightarrow N_{male,Sandals} \overset{?}{=} N \times P(male)P(Sandals)$$

$$\Rightarrow N_{male,Sandals} \overset{?}{=} N \times \underbrace{\frac{N_{male}}{N}}_{R} \times \underbrace{\frac{N_{Sandals}}{N}}_{C}$$

$$\Rightarrow \underbrace{N_{male,Sandals}}_{O_{i,j}} \overset{?}{=} \underbrace{\frac{N_{male} \times N_{Sandals}}{N}}_{E_{i,j}}$$

empirical frequency/count

expected frequency/count

$O_{i,j}$

$E_{i,j}$

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$\text{with degrees of freedom} = (I-1) \times (J-1)$$

which is more significiant ?

$(1005-1000)^2 > (13-10)^2$

$\frac{(1005-1000)^2}{1000} < \frac{(13-10)^2}{10}$

Review: Chi-Square Test (cont.)

- Chi-Square Distribution

$$F_{\chi^2}(u,n) = \int_0^u \frac{x^{(n-2)/2} e^{-x/2} dx}{2^{n/2}\left[(n-2)/2\right]!}$$

$n\backslash F$	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	$.0^4393$	$.0^3157$	$.0^3982$	$.0^2393$	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

Review: Chi-Square Test (cont.)

- Chi-Square Distribution (cont.)
  - An asymmetric distribution

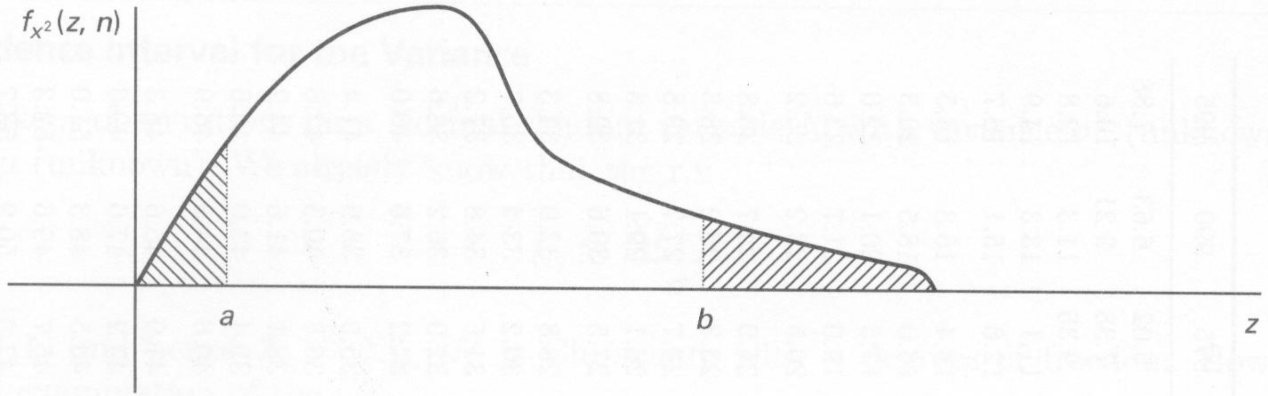


Figure 4.8-2 The points a, b are chosen so that equal amounts of area are removed from the tails. The area of the remainder should equal γ, the confidence level.

- In e.g. (2), for example, we can find $\chi^2 > u$ such that we can have *a* confidence of *P*% (or have error less than 100%-*P*%) to reject the Null Hypothesis

Review: Chi-Square Test (cont.)

- E.g. (2), behavior analyses in sociology (cont.)

Table 1.e. Male and Female Undergraduate Footwear Preferences: Observed and Expected Frequencies

	Sandals	Sneakers	Leather shoes	Boots	Other	Total
Male observed	6	17	13	9	5	50
Male expected	9.5	11	10	12.5	7	
Female observed	13	5	7	16	9	50
Female expected	9.5	11	10	12.5	7	
Total	19	22	20	25	14	100

Male/Sandals:	((19 X 50)/100) = 9.5
Male/Sneakers:	((22 X 50)/100) = 11
Male/Leather Shoes:	((20 X 50)/100) = 10
Male/Boots:	((25 X 50)/100) = 12.5
Male/Other:	((14 X 50)/100) = 7
Female/Sandals:	((19 X 50)/100) = 9.5
Female/Sneakers:	((22 X 50)/100) = 11
Female/Leather Shoes:	((20 X 50)/100) = 10
Female/Boots:	((25 X 50)/100) = 12.5
Female/Other:	((14 X 50)/100) = 7

Male/Sandals:	$((6 - 9.5)^2/9.5) =$	1.289
Male/Sneakers:	$((17 - 11)^2/11) =$	3.273
Male/Leather Shoes:	$((13 - 10)^2/10) =$	0.900
Male/Boots:	$((9 - 12.5)^2/12.5) =$	0.980
Male/Other:	$((5 - 7)^2/7) =$	0.571
Female/Sandals:	$((13 - 9.5)^2/9.5) =$	1.289
Female/Sneakers:	$((5 - 11)^2/11) =$	3.273
Female/Leather Shoes:	$((7 - 10)^2/10) =$	0.900
Female/Boots:	$((16 - 12.5)^2/12.5) =$	0.980
Female/Other:	$((9 - 7)^2/7) =$	0.571

The total chi square value for Table 1 is 14.026.

The degrees of freedom for this Chi-Square distribution is (2-1)X(5-1)=4

Notice that because we originally obtained a balanced male/female sample, our male and female expected scores are the same.

Review: Chi-Square Test (cont.)

- E.g. (2), behavior analyses in sociology (cont.)
  - If we want to reject the Null Hypothesis with confidence larger than 95%, χ^2 must be larger than 9.49 (with degrees of freedom=4)
  - Because 14.2602> 9.49, we can reject the null hypothesis and affirm the claim that males and females differ in their footwear preferences

Supervised Value Reduction

- Method III: ChiMerge technique
  - An automated discretization algorithm that analyzes the quality of multiple intervals for a given feature using χ^2 statistics
  - Determine similarities between distributions of data in two adjacent intervals based on output classification of samples
    - If the χ^2 test indicates that the output class is independent of the feature's intervals, merge them; otherwise, stop merging!

Data Set

Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

initial interval points :

0, 2, 5, 7.5, 8.5, 10, ...,60

Supervised Value Reduction (cont.)

- Method III: ChiMerge technique (cont.)
  - Algorithm
    1. Sort the data for the given feature in ascending order
    2. Define initial intervals so that every value of the feature is in a separate interval
    3. Repeat until no χ^2 of any two adjacent intervals is less then threshold value
       - If no merge is possible, we can increase threshold value in order to increase the possibility of a new merge

Supervised Value Reduction (cont.)

- Method III: ChiMerge technique (cont.)

initial interval points :
0, 2, 5, 7.5, 8.5, 10, ...,60

Data Set

Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} (A_{ij} - E_{ij})^2 / E_{ij}$$

$O_{i,j}$

where:

k = number of classes,
A_{ij} = number of instances in the i-th interval, j-th class,
E_{ij} = expected frequency of A_{ij}, which is computed as $(R_i \cdot C_j) / N$,
R_i = number of instances in the i-th interval = $\sum A_{ij}$, j = 1,…k,
C_j = number of instances in the j-th class = $\sum A_{ij}$, i = 1,2,
N = total number of instances = $\sum R_i$, i = 1,2.

χ^2 was minimum for intervals: [7.5, 8.5] and [8.5, 10]

	K=1	K=2	Σ
Interval [7.5, 8.5]	A_{11}=1	A_{12}=0	R_1=1
Interval [8.5, 9.5]	A_{21}=1	A_{22}=0	R_2=1
Σ	C_1=2	C_2=0	N=2

Based on the table's values, we can calculate expected values:

$E_{11} = 2/2 = 1$,
$E_{12} = 0/2 \approx 0.1$,
$E_{21} = 2/2 = 1$, and
$E_{22} = 0/2 \approx 0.1$

and corresponding $\chi 2$ test:
$\chi^2 = (1-1)^2 / 1 + (0-0.1)^2 / 0.1 + (1-1)^2 / 1 + (0-0.1)^2 / 0.1 = 0.2$

For the degree of freedom d=1, and $\chi^2 = 0.2 < 2.706$
(**MERGE !**)

degrees of freedom $= (I-1)\times(J-1)$

confidence>0.90

Supervised Value Reduction (cont.)

- Method III: ChiMerge technique (cont.)

	K=1	K=2	Σ
Interval [0, 7.5]	A_{11}=2	A_{12}=1	R_1=3
Interval [7.5, 10]	A_{21}=2	A_{22}=0	R_2=2
Σ	C_1=4	C_2=1	N=5

$E_{11} = 12/5 = 2.4$,
$E_{12} = 3/5 = 0.6$,
$E_{21} = 8/5 = 1.6$, and
$E_{22} = 2/5 = 0.4$

$$\chi^2 = (2 - 2.4)^2 / 2.4 + (1 - 0.6)^2 / 0.6 + (2 - 1.6)^2 / 1.6 + (0 - 0.4)^2 / 0.4$$

$\chi^2 = 0.834$
For the degree of freedom d=1, $\chi^2 = 0.834 < 2.706$ (MERGE!)

	K=1	K=2	Σ
Interval [0, 10.0]	A_{11}=4	A_{12}=1	R_1=5
Interval [10.0, 42.0]	A_{21}=1	A_{22}=3	R_2=4
Σ	C_1=5	C_2=4	N=9

$E_{11} = 2.78$, $E_{12} = 2.22$, $E_{21} = 2.22$, $E_{22} = 1.78$, and $\chi^2 = 2.72 > 2.706$
(NO MERGE !)

Final discretization: [0, 10], [10, 42], and [42, 60]

Low Medium High

Case Reduction

- Also called “raw reduction”
- Premise: the largest and the most critical dimension in the initial data set is the number of cases or samples
  - The number of rows in the tabular representation of data
- Simple case reduction can be done in the preprocessing (data-cleansing) phase
  - Elimination of outliers
  - Elimination of samples with missing feature values

There will be many samples remained !

- Or, case reduction achieved by using a sampled subset of samples (called an estimator) to provide some information about the entire data set (using sampling methods)
  - Reduced cost, greater speed, greater scope, even higher accuracy ?
    - Greater scope? can cover equally the rarely and frequently occurred samples

estimator ?
estimate ?
estimation ?

Case Reduction (cont.)

- Method I: Systematic sampling
  - The simplest sampling technique
  - If 50% of a data set should be selected, simply take every other sample in a data set (e.g., 任兩個samples取其一)
  - There will be a problem, if the data set posses some regularities

$$D = \{(\boldsymbol{x}^1, A), (\boldsymbol{x}^2, B), (\boldsymbol{x}^3, A), (\boldsymbol{x}^4, B), (\boldsymbol{x}^5, A), \ldots, (\boldsymbol{x}^N, B)\}$$

$$\overset{\text{Sampling}}{\Rightarrow}$$

$$D' = \{(\boldsymbol{x}'^1, A), (\boldsymbol{x}'^2, A), \ldots, (\boldsymbol{x}'^{N/2}, A)\}$$

Case Reduction (cont.)

- Method II: Random sampling
  - Every sample from the initial data set has the same chance of being selected in the subset
  - Two variants:
    1. Random sampling without replacement
       - Select *n* distinct samples form *N* initial samples without repetition
       - Avoid any bias in a selection
    2. Random sampling with replacement
       - All samples are given really equal chance of being selected, any of samples can be selected more than once

Case Reduction (cont.)

- Method II: Random sampling (cont.)
  - Notice that random sampling is an iterative process which may have two forms

    1. Incremental sampling

       10%, 20%, 33%, 50%, 67%, 100%

       - Perform data mining on increasing larger random subsets to observe the trends in performances
         - The smallest subset should be substantial (e.g., >1000 samples)
       - Stop when no progress is made

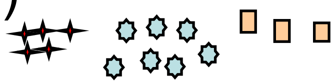
    2. Average sampling
       - Solutions found from many random subsets of samples are averaged or voted
         - Regression problems → averaging
         - Classification problems → voting
       - Drawback: the repetitive process of data mining on smaller sets of samples

$$h_1(x)=A,\ h_2(x)=B,\ h_3(x)=A \ \overset{\text{Voted}}{\Rightarrow}\ h^*(x)=A$$

$$h_1(x)=6,\ h_2(x)=6.5,\ h_3(x)=6.7 \ \overset{\text{Averaged}}{\Rightarrow}\ h^*(x)=6.4$$

Case Reduction (cont.)

- Method III: Stratified(分層的) sampling
  - The entire data set is split into non-overlapping subsets or strata
  - Sampling is performed for each different strata independently of each other
  - Combine all small subsets from different strata to form the final, total subset of samples
  - Better than random sampling if the strata is relatively homogeneous (→smaller variance of sampled data)



- Method IV: Inverse sampling
  - Used when a feature in a data set occurs with rare frequency
    (not enough information can be given to estimate a feature value)
  - Sampling start with the smallest subset and it continues until some conditions about the required number of feature values are satisfied

Data sampling for speech recognition “utterance-陳水扁” >10 times
“utterance-陳水在“ >10 times
….
“utterance-陳萬水” >10 times

HW-2-A: Feature Selection (Due 3/24)

- Unsupervised Feature Selection using Entropy Measure
  - Given four-dimensional samples where features are categorical:

X_1	X_2	X_3	X_4
3	3	1	A
3	6	2	A
5	3	1	B
5	6	2	B
7	3	1	A
5	4	2	B

Apply a method for unsupervised feature selection based on entropy measure to reduce one dimension from the given data set

HW-2-B: Value Reduction (Due 3/24)

- Supervised Value Reduction using ChiMerge
  - Given the data set X with two input features (I_1 and I_2) and one output feature (O) representing the classification of samples:

X:

I_1	I_2	O
2.5	1.6	0
7.2	4.3	1
3.4	5.8	1
5.6	3.6	0
4.8	7.2	1
8.1	4.9	0
6.3	4.8	1

Apply ChiMerge to reduce the number of values (with confidence >0.9)

- Reduce the number of numeric values for feature I_1 and find the final, reduced number of intervals
- Reduce the number of numeric values for feature I_2 and find the final, reduced number of intervals