Data Mining & Machine Learning

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

ILLINOIS TECH
College of Computing

Getting to Know Data

- **Types of the Data**
  - **Qualitative (Categorical/Nominal)**
    - Nominal = Values are strings
    - Special Nominal Variable
      - Binary, such as gender
      - Ordinal, such as letter grade (A, B, C, F)
  - **Quantitative (Numerical)**
    - Discrete, we need to count to get values
      Example: number of students in the class
    - Continuous, we need to measure to get values
      Example: the length of the table

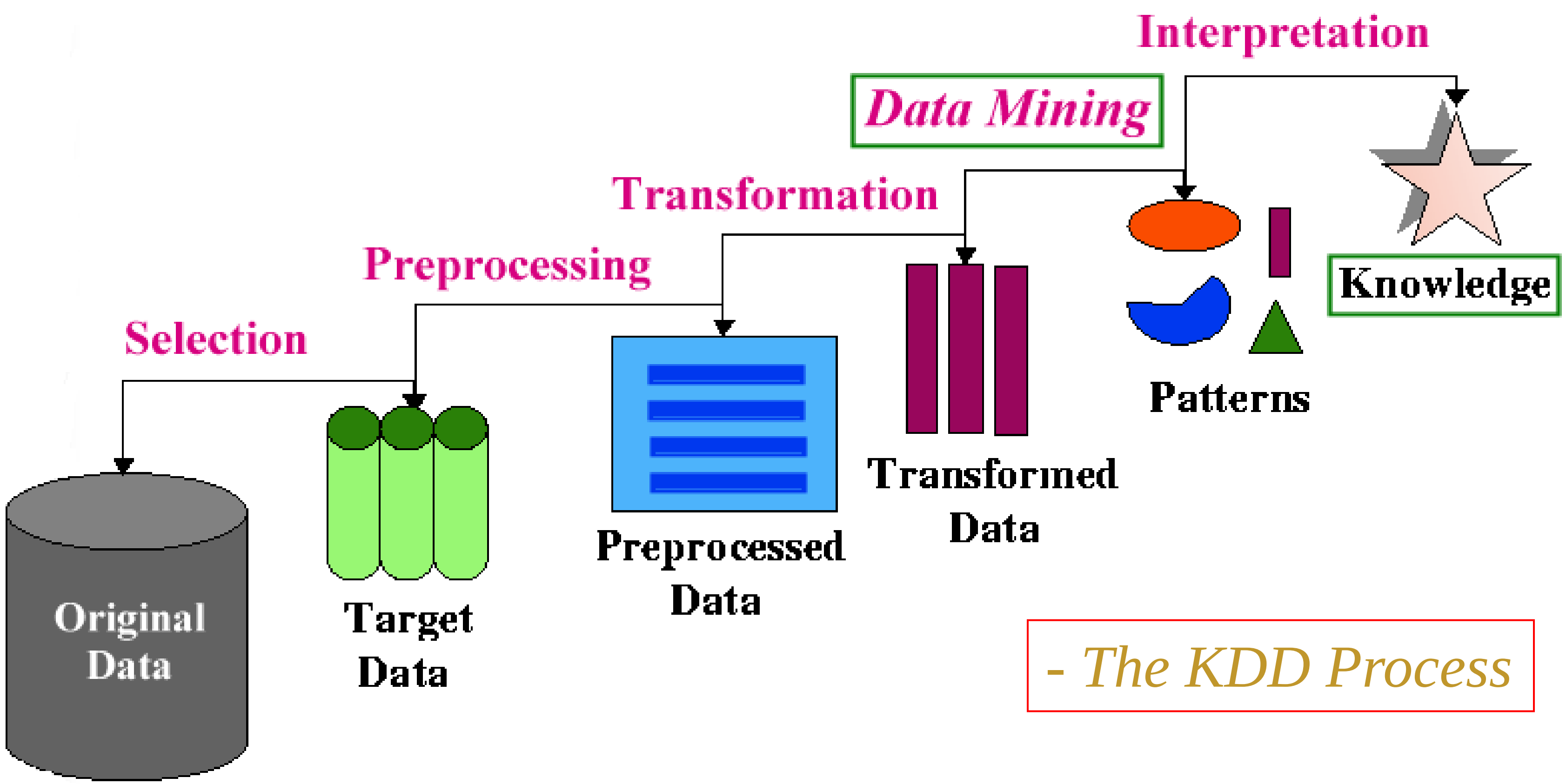
Getting to Know Data

- If you observe that a column of numbers, it is not guaranteed that this variable is a numerical variable
- These numbers may be encoded for some reason, for example, 1 – India, 2 – China, 3- France, 4 – Spain
- You need to be careful about the data types in a data set

Week 2 - Schedule

- KDD Process
- Data Preprocessing
  - Why: Data Quality
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction
  - Summary

(Knowledge Discovery in DB) KDD Process



Week 2 - Schedule

KDD Process: Data PreProcessing

- Why: Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Week 2 - Schedule

KDD Process: Data PreProcessing

- Why: Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

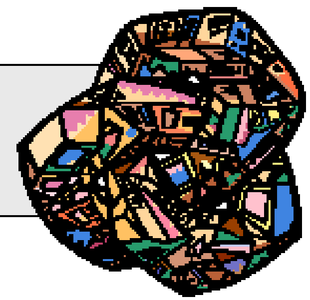
Data Quality

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data PreProcessing

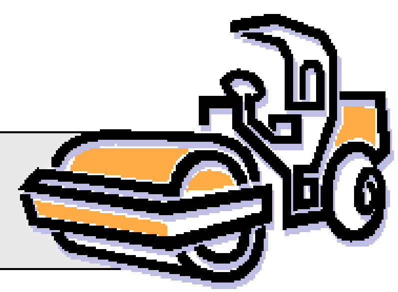
Data Cleaning

Data Integration



Data Transformation

Data Reduction



Week 2 - Schedule

KDD Process: Data PreProcessing

- Why: Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Cleaning

- Real-world application data can be dirty:
  - Incomplete: missing values
  - Noisy: errors, outliers, e.g., salary = -10
  - Inconsistent: 80, 90, A, B, C
- Data cleaning attempts to:
  - Fill in missing values
  - Smooth out noisy data
  - Correct inconsistencies
  - Remove irrelevant data

Data Cleaning: Missing Values

- Data is not always available (missing attribute values in records)
  - equipment malfunction
  - deleted due to inconsistency or misunderstanding
  - not considered important at time of data gathering
- Solving Missing Data if it is numerical variable. Exp: age
  - Ignore the record with missing values;
  - Fill in the missing values manually;
  - Fill in the missing values automatically;
    - Use a global constant to fill in missing values
    - Use the attribute mean value to filling missing values of that attribute;
    - Use the attribute mean for all samples belonging to the same class to fill in the missing values;
    - Build a predictive model (e.g., regression model) to predict missing values

Data Cleaning: Missing Values

- Fill in Missing Data if it is numerical variable, Exp: age
  - Use a global constant to fill in missing values
  - Use the attribute mean value to filling missing values of that attribute;
  - Use the attribute mean for all samples belonging to the same class to fill in the missing values;
  - Build a predictive model (e.g., regression model) to predict missing values
- Fill in Missing Data if it is nominal variable, Exp: gender
  - Use a global constant to fill in missing values, e.g., NULL
  - Use the most frequent value to filling missing values of that attribute;
  - Use the most frequent value belonging to the same class to fill in the missing values;
  - Build a predictive model (e.g., classification model) to predict missing values

Data Cleaning: Noisy Data

Solutions to reduce noisy data when the variance is large

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning: Noisy Data

- Binning (when a numerical variable has large variance/outliers)

Original Data for "price" (after sorting): 4, 8, 15, 21, 21, 24, 25, 28, 34

Binning

Partition into equidepth bins
Bin1: 4, 8, 15
Bin2: 21, 21, 24
Bin3: 25, 28, 34

Each value in a bin is replaced by the mean value of the bin.

means
Bin1: 9, 9, 9
Bin2: 22, 22, 22
Bin3: 29, 29, 29

boundaries
Bin1: 4, 4, 15
Bin2: 21, 21, 24
Bin3: 25, 25, 34

Min and Max values in each bin are identified (boundaries). Each value in a bin is replaced with the closest boundary value.

Data Cleaning: Noisy Data

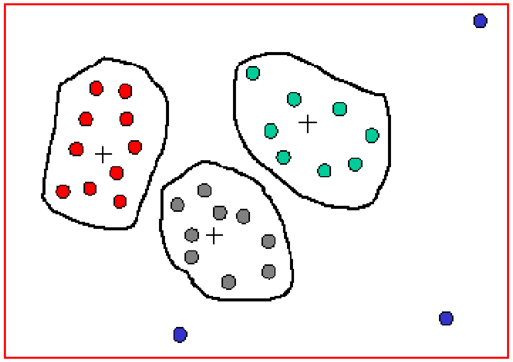
Steps in Binning

- Step 1: Rank the values from smaller to larger
- Step 2: Make a decision how many bins you need, i.e., you need to decide a bin size if you want to create bins with equal length
- Step 3: Create bins equally (Note: the last bin may not have the equal length)
- Step 4: Choose a strategy (by means or boundaries) to transform value in each bin

Data Cleaning: Noisy Data

Other Methods

Clustering



Similar values are organized into groups (clusters). Values falling outside of clusters may be considered “outliers” and may be candidates for elimination.

Week 2 - Schedule

KDD Process: Data PreProcessing

- Why: Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Integration

- Data analysis may require a combination of data from multiple sources into a coherent data store
- Challenges in Data Integration:
  - Schema integration: CID = C_number = Cust-id = cust#
  - Identity identification problem: Bill Clinton = William Clinton
  - Data value conflicts (different representations or scales, e.g., $ and ￥)
  - Redundant attributes (redundant if it can be derived from other attributes) -- may be able to identify redundancies via correlation analysis:

Data Integration: Correlation Analysis

- What is correlation?
  If two variables have strong correlations, it means that they may change together!

Student	Gender	Dept	TimeStudy	TimeGame	Grade
1	M	ITMD	20	1	A
2	F	ITMS	25	2	A
3	M	ITMD	5	20	C
4	F	ITMS	6	18	C

- Can you observe some correlations in this table?

Data Integration: Correlation Analysis

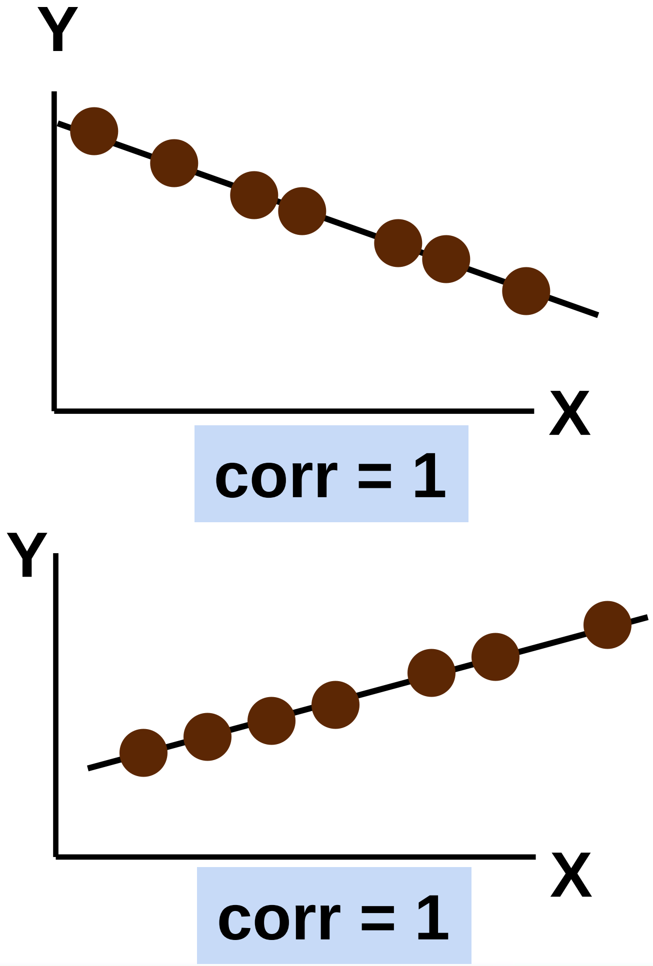
- Two numerical variables: Pearson correlation
- One numerical vs one nominal variable: ANOVA
- Two nominal variables
  - Conditional probabilities
  - Chi-square test

Data Integration: Correlation Analysis

- Two numerical variables: Pearson correlation
- One numerical vs one nominal variable: ANOVA
- Two nominal variables
  - Conditional probabilities
  - Chi-square test

Data Integration: Correlation Analysis

- For Numeric Data Only: Pearson correlation

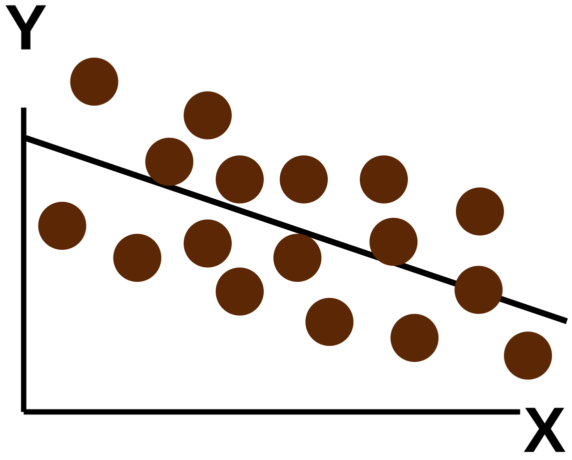


Perfect linear correlation between X and Y:

100% of the variation in Y is explained by variation in X

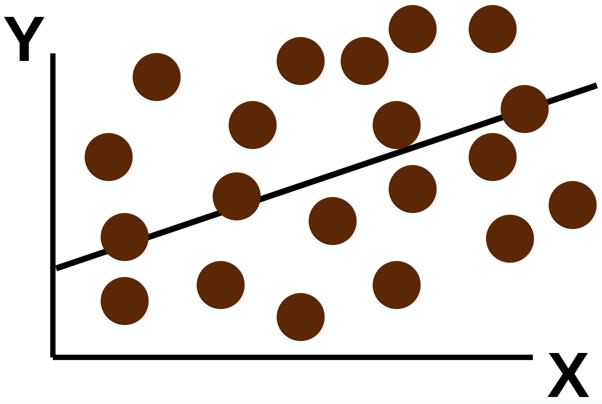
Data Integration: Correlation Analysis

- For Numeric Data Only: Pearson correlation



-1 < corr < 1

Weaker linear correlation between X and Y:



Some but not all of the variation in Y is explained by variation in X

Data Integration: Correlation Analysis

- For Numeric Data Only: Pearson correlation

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

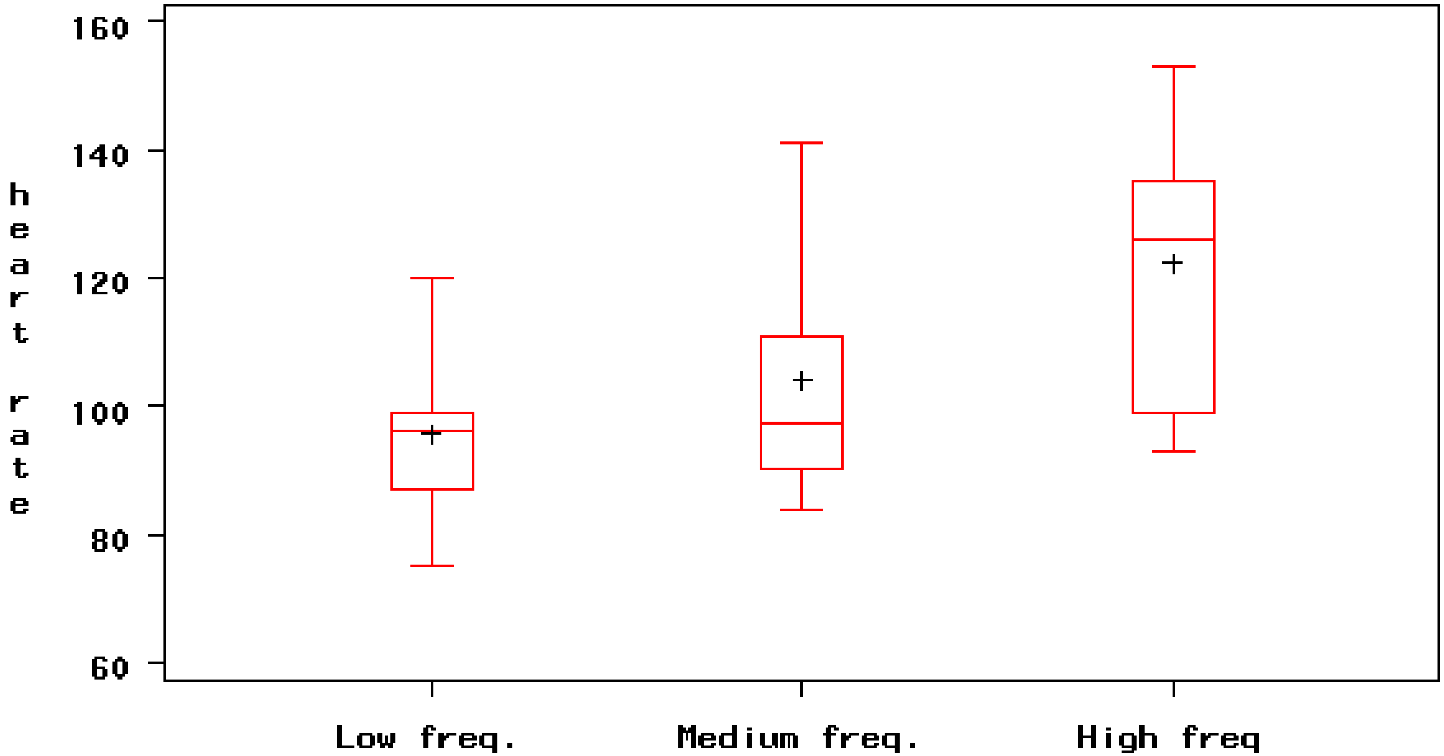
- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Data Integration: Correlation Analysis

- Two numerical variables: Pearson correlation
- One numerical vs one nominal variable: ANOVA
- Two nominal variables
  - Conditional probabilities
  - Chi-square test

Data Integration: Correlation Analysis

- Between Nominal and Numerical Variables



Data Integration: Correlation Analysis

- Between Nominal and Numerical Variables: ANOVA

1. Be sure that the observations arise from independent groups!

2. Draw side-by-side box plots for the groups, to visualize the differences among the groups and the within-group variation

3. Estimate the ANOVA regression model for t=1,…,K

   where the errors e_{it} are normally distributed and with constant standard deviation σ. <u>Use the regression F-test to check the hypothesis that the averages are equal.</u>

4. Examine the residuals to verify that the model assumptions are satisfied.

Data Integration: Correlation Analysis

- Between Nominal and Numerical Variables: ANOVA

Student	Gender	Dept	TimeStudy	TimeGame	Grade
1	M	ITMD	20	1	A
2	F	ITMS	25	2	A
3	M	ITMD	5	20	C
4	F	ITMS	6	18	C

- How about Grade vs TimeStudy?

Data Integration: Correlation Analysis

- Two numerical variables: Pearson correlation
- One numerical vs one nominal variable: ANOVA
- Two nominal variables
  - Conditional probabilities
  - Chi-square test

Data Integration: Correlation Analysis

- Dependency between values: Conditional probabilities

Correlation analysis: Pr(A,B) / (Pr(A).Pr(B))
= 1: independent,
> 1: positive correlation,
< 1: negative correlation.

Correlation between
<u>two nominal values</u>

Student	Gender	Dept	TimeStudy	TimeGame	Grade
1	M	ITMD	20	1	A
2	F	ITMS	25	2	A
3	M	ITMD	5	20	C
4	F	ITMS	6	18	C

- How about Gender = M vs. Dept = ITMD?

Data Integration: Correlation Analysis

- Dependency between variables: Chi-square test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

The larger the X^2 value, the more likely the variables are related
The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count

The test is applied when you have <u>two categorical variables</u> from a single population. It is used to determine whether there is a significant association between the two variables.

Data Integration: Correlation Analysis

- For Nominal Data Only: Chi-square test

Example: "Which holiday do you prefer?"

	Hiking	Cruise
Men	209	280
Women	225	248

Does Gender affect Preferred Holiday?

If Gender (Man or Woman) **does** affect Preferred Holiday we say they are **dependent**.

By doing some special calculations (explained later), we come up with a "p" value:

p value is 0.132

Now, $p < 0.05$ is the usual test for dependence. In this case **p is greater than 0.05**, so we believe the variables are **independent** (ie not linked together).

In other words Men and Women probably do **not** have a different preference for hiking Holidays or Cruises.

Data Integration: Correlation Analysis

- For Nominal Data Only: Chi-square test

	Voting Preferences			Row total
	Republican	**Democrat**	**Independent**	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

$$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

where $O_{r,c}$ is the observed frequency count at level *r* of Variable A and level *c* of Variable B, ar

$E_{r,c}$ is the expected frequency count at level *r* of Variable A and level *c* of Variable B.

$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = (n_r * n_c) / n$$

$$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$$

$$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$$

$$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

$$X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40$$
$$+ (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$$

$$X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$$

$$\boxed{X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2 \quad P(X^2 > 16.2) = 0.0003.}$$

http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP

Data Integration: Correlation Analysis

- For Nominal Data Only: Chi-square test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Null hypothesis: two variables are independent
- Tutorial and Example
  https://online.stat.psu.edu/stat500/lesson/8/8.1
- Coding by R and Python
  - http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence
  - https://thinkingneuron.com/how-to-measure-the-correlation-between-two-categorical-variables-in-python/

Data Integration: Correlation Analysis

- For Nominal Data Only: Chi-square test
  - P-value tells whether we should reject H0
  - P-value also tells the degree of significance
  - The contingency coefficient can tell the degree of dependency/correlation

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

    - Value ranges in [0, 1]
    - Larger value, larger dependency or correlation
    - N = number of observations

Week 2 - Schedule

KDD Process: Data PreProcessing

- Why: Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Transformation

Why we need transformation?

- Attribute values are at different scales
- Difficult for comparison
- Different Data Formats
- Special requirements by specific data mining tasks

Data Transformation

What are the popular transformation tasks

- Smoothing by binning
- Data Normalization
- Data Discretization

Data Transformation: Normalization

Sometimes, we need to use values in the same scale

- Min-max Normalization
- Z-score Normalization
- Decimal Scaling for Normalization

Data Transformation: Normalization

- Min-max Normalization

$$x'_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}(new\max - new\min) + new\min$$

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

ID	Gender	Age	Salary
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

Data Transformation: Normalization

- Z-score Normalization : v' = (v - Mean) / Stdev

Humidity
85
90
78
96
80
70
65
95
70
80
70
90
75
80

Mean = 80.3
Stdev = 9.84

Humidity
0.48
0.99
-0.23
1.60
-0.03
-1.05
-1.55
1.49
-1.05
-0.03
-1.05
0.99
-0.54
-0.03

After transformation, mean = 0, Stdev = 1

Data Transformation: Normalization

- Decimal Scaling for Normalization
  - moves the decimal point of v by *j* positions such that *j* is the minimum number of positions moved so that absolute maximum value falls in [0..1].
  - $v' = v / 10^j$
  - Ex: if v ranges between -56 and 9976, *j*=4 ==> v' ranges between -0.0056 and 0.9976

Data Transformation: Normalization

- Normalization
  - Min-Max → can produce values in any new scale
  - Decimal scaling → can produce values in [-1, 1]
  - Z-score method → no controls on the new scales

Data Transformation: Discretization

Data Conversion between Numeric and Nominal data

- From Numeric to Nominal/Ordinal Data
- From Nominal to Numeric Data

Data Transformation: Discretization

Data Conversion between Numeric and Nominal data

- From Numeric to Nominal/Ordinal Data

Humidity
85
90
78
96
80
70
65
95
70
80
70
90
75
80

Low = 60-69
Normal = 70-79
High = 80+

Humidity
High
High
Normal
High
High
Normal
Low
High
Normal
High
Normal
High
Normal
High

Data Transformation: Discretization

Data Conversion between Numeric and Nominal data

- From Nominal to Numeric Data

ID	Outlook	Temperature	Humidity	Windy
1	sunny	85	85	FALSE
2	sunny	80	90	TRUE
3	overcast	83	78	FALSE
4	rain	70	96	FALSE
5	rain	68	80	FALSE
6	rain	65	70	TRUE
7	overcast	58	65	TRUE
8	sunny	72	95	FALSE
9	sunny	69	70	FALSE
10	rain	71	80	FALSE
11	sunny	75	70	TRUE
12	overcast	73	90	TRUE
13	overcast	81	75	FALSE
14	rain	75	80	TRUE

OutLook	OutLook	OutLook	Temp	Humidity	Windy	Windy
overcast	rain	sunny			TRUE	FALSE
0	0	1	85	85	0	1
0	0	1	80	90	1	0
1	0	0	83	78	0	1
0	1	0	70	96	0	1
0	1	0	68	80	0	1
0	1	0	65	70	1	0
1	0	0	64	65	1	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Data Transformation: Discretization

Data Conversion between Numeric and Nominal data

- From Nominal to Numeric Data

ID	Outlook	Temperature	Humidity	Windy
1	sunny	85	85	FALSE
2	sunny	80	90	TRUE
3	overcast	83	78	FALSE
4	rain	70	96	FALSE
5	rain	68	80	FALSE
6	rain	65	70	TRUE
7	overcast	58	65	TRUE
8	sunny	72	95	FALSE
9	sunny	69	70	FALSE
10	rain	71	80	FALSE
11	sunny	75	70	TRUE
12	overcast	73	90	TRUE
13	overcast	81	75	FALSE
14	rain	75	80	TRUE

Assume there are N values in a variable, you just need to create N-1 new columns

OutLook overcast	OutLook rain	OutLook sunny	Temp	Humidity	Windy TRUE	Windy FALSE
0	0	1	85	85	0	1
0	0	1	80	90	1	0
1	0	0	83	78	0	1
0	1	0	70	96	0	1
0	1	0	68	80	0	1
0	1	0	65	70	1	0
1	0	0	64	65	1	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Two columns are enough

Not necessary

Data Transformation: Discretization

Data Conversion between Numeric and Nominal data

- From Nominal to Numeric Data
- Special case: when a nominal variable is ordinal variable
  - In this case, you can encode them by numbers directly

Grade	Grade1	Grade2	Grade3
A	0	0	4
B	1	1	3
C	2	3	2
F	3	5	1

Week 2 - Schedule

KDD Process: Data PreProcessing

- Why: Data Quality
- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Summary

Data Reduction

- Data is often too large; reducing data can improve performance
- Data reduction consists of reducing the representation of the data set while producing the same (or almost the same) results
- Data reduction includes:
  - Data cube aggregation
  - Dimensionality reduction
  - Discretization
  - Numerosity reduction
    - Regression
    - Histograms
    - Clustering
    - Sampling

Data Reduction Techniques

- Data reduction is necessary in most of the data mining tasks
- Not all of the data are useful
- Irrelevant data may leave negative impact on DM
- We will have a special session "Feature Selection and Reduction" in the later class
- We briefly introduce it in this class

Summary

- Data Cleaning
  Missing values, smoothing by binning
- Data Integration
  Correlation analysis
- Data Transformation
  Normalization, Discretization
- Data Reduction
  We will introduce more later in the lecture
  "Feature Selection and Reduction"