Data-driven Intelligent Systems

Lecture 4
Preprocessing Methods

KNOWLEDGE
TECHNOLOGY

http://www.informatik.uni-hamburg.de/WTM/

Data Preprocessing

- Data Cleaning
- Data Integration
- Values Reduction
  - Chi-Square Statistical Test (nominal data)
  - ChiMerge Discretization
  - Binning

Major Tasks in Data Preprocessing

- ***Data cleaning***
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- ***Data integration***
  - Integration of multiple databases, data cubes, or files
- ***Data reduction***
  - Data compression, e.g. values reduction, discretization
  - Dimensionality reduction, i.e. features reduction
- ***Data transformation***
  - Normalization
  - PCA

Why We should Clean Dirty Data

Why preprocess the Data?

Measures for data quality: A multidimensional view

- Completeness: is the data fully available? What to do if not?
- Consistency: differences in data units or name conventions?
- Timeliness: measurements from different epochs? Old measure devices?
- Believability: is the data source reliable?
- Interpretability: how easily can the data be understood?

Noisy Data

- Data in the real world is “noisy” or incorrect
- “Noisy” attribute values may be due to:
  - technology limitation
  - faulty data collection instruments
  - data entry problems; human error
  - data transmission problems
  - inconsistency in naming convention
  - duplicate records
- Noise: random error or variance
  in a measured variable

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks
https://arxiv.org/abs/2103.14749

Data Cleaning

Types of error

- ***Incomplete***: lacking attribute values, lacking certain attributes of interest, or only aggregate data available
  - e.g., *Occupation*=“ ” (missing data)
- ***Noisy***: containing noise, errors, or outliers
  - e.g., *Salary*=“−10” (an error)
- ***Inconsistent***: containing discrepancies in codes or names, e.g.,
  - *Age*=“42”, *Birthday*=“03/07/2012”
  - Was rating “1, 2, 3”, now rating “A, B, C”
  - discrepancy between duplicate records
- ***Intentionally imprecise***
  - Jan. 1 as everyone’s birthday?
  - → *disguised missing* data

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data not considered important at time of entry
  - did not register history or changes of the data
- Missing data may need to be inferred

How to handle missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification) — not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill it in automatically with
  - a global constant: e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value using inference such as Bayesian formula or decision tree based on other attributes

Missing Data

- One possible interpretation of missing values – ***"don't care"*** values:

  `X = {1,?,3}`

  → for the second feature the domain is `[0,1,2,3,4]`, then extend to:

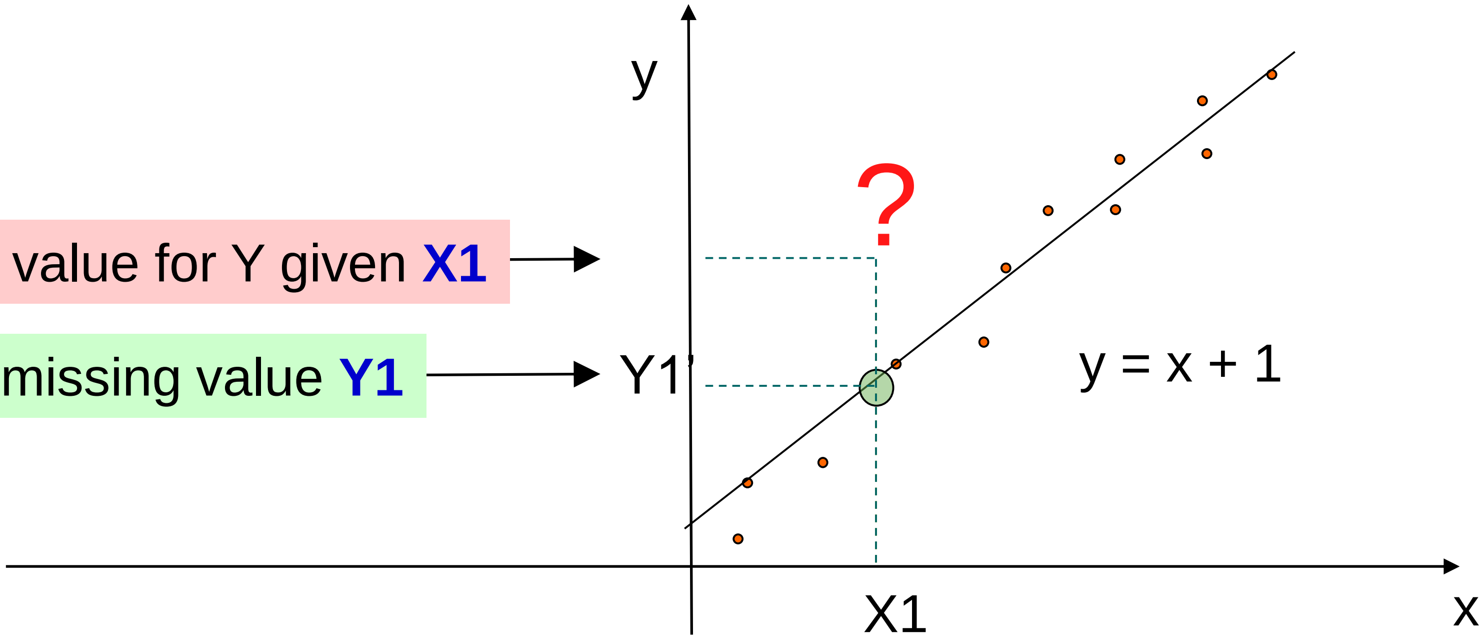
  `X1 = {1,0,3}, X2 = {1,1,3}, X3 = {1,2,3}, X4 = {1,3,3}, X5 = {1,4,3}`

- *Data miner* can generate model of ***correlation between features***.
  - Different techniques possible: regression, Bayesian formalism, clustering, or decision tree induction.

Missing Data Replacement with Regression Analysis

Missing value for Y given **X1**

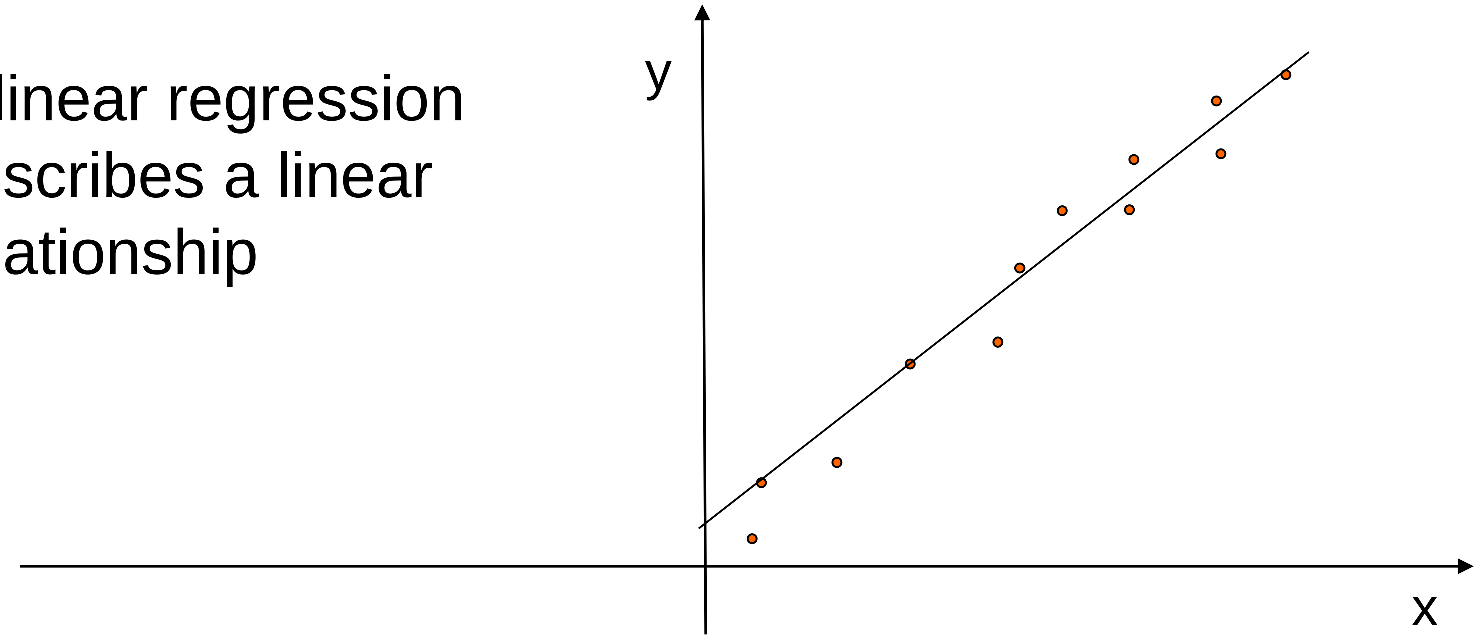
Predicted missing value **Y1**



- In general, replacement of missing values using a simple, artificial schema of data preparation is *speculative and often misleading*.
- It is best to generate multiple solutions of data mining algorithms ***with and without features*** that have missing values. Then compare, analyse, interpret.

Refresher on Regression Analysis

- A linear regression describes a linear relationship



dependent variable

independent variable

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

residual error

y-intercept

slope coefficient

Regression: Least Squares Criterion

- A good fit minimizes the residual error *E*

data

estimated value

$$E = \sum_{data} (y - (\beta_0 + \beta_1 \cdot x))^2$$

- Obtain β_0 and β_1 by minimizing *E*

www.feinberg.northwestern.edu/sites/bcc/docs/2011-lectures/lecture3linearlogisticregression2011.pdf

Regression: Least Squares Equation

- Minimizing *E* leads to the following values:

$$\beta_1 = \frac{\sum_{data} (x - \bar{x}) \cdot (y - \bar{y})}{\sum_{data} (x - \bar{x})^2}$$

slope: estimated change of *y* as a result of a one-unit change of *x*

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

intercept: estimated average value of *y* when *x* is zero

$\bar{x}$ and $\bar{y}$ are the mean values of *x* and *y*

How to handle noisy Data?

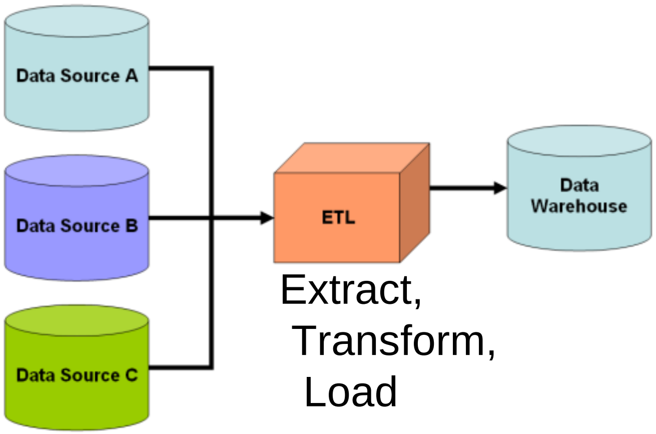
- Data visualization
  - Boxplot ✓ to detect outliers
- Regression ✓
  - smooth by fitting the data into regression model
  - non-linear regression: careful to not overfit the model
  - note: data does not get more rich in information
- Data Discretization
  - Binning (→ later today)
  - Clustering (→ later Lecture)
    - detect and remove outliers when forming data clusters

Data Preprocessing

- Data Cleaning
- Data Integration
- Values Reduction
  - Chi-Square Statistical Test (nominal data)
  - ChiMerge Discretization
  - Binning

Data Quality: Why Preprocess the Data?

Data Integration



Data integration combines data from multiple sources into a coherent store

- Integrate metadata from different sources
  - Schema integration: e.g., A.cust-id ≡ B.cust-#
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., BER = TXL = Berlin-Tegel
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- ***Redundant data*** occur often when integrating multiple databases
  - ***Object identification***: The same attribute or object may have different names in different databases
  - ***Derivable data:*** One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Careful integration can
  - reduce/avoid redundancies and inconsistencies and
  - improve mining speed and quality
- Redundant attributes may be possible to detect by ***correlation analysis*** (→ next Lecture)

Data Preprocessing

- Data Cleaning
- Data Integration
- Values Reduction
  - ▶ Chi-Square Statistical Test (nominal data)
  - ChiMerge Discretization
  - Binning

Correlation Analysis for Nominal Data

- Nominal Data: labels for variables, e.g. hair color
- Labels can be counted, e.g. how often we observe blond, brown, ginger hair
- **Task: Compare the counts to expected counts**
- ***X^2 (Chi-square) test***

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

- Tells whether two data distributions are statistically different
- Used also to identify correlations between variables → Example

Chi-Square Calculation: an Example

- Questionnaire among N=1500 participants:

	Play chess	Not play chess	Sum (row)
Like science fiction	250	250	500
Not like science fiction	50	950	1000
Sum (column)	300	1200	N=1500

- Expected results e_{ij} from the null hypothesis stating that "preferred reading" and "game favour" are uncorrelated:

	Play chess	Not play chess	Sum (row)
Like science fiction			500
Not like science fiction			1000
Sum (column)	300	1200	N=1500

$$e_{ij} = sum(col\ \ i) \cdot sum(row\ \ j) / N$$

Chi-Square Calculation

- Plug in the values into the equation:

$$\chi^2 = \frac{(250-100)^2}{100} + \frac{(50-200)^2}{200} + \frac{(250-400)^2}{400} + \frac{(950-800)^2}{800} = \underline{421.9}$$

- Ok, we now have a number… but what does it tell us?
  - We need the Chi-Square Distribution →

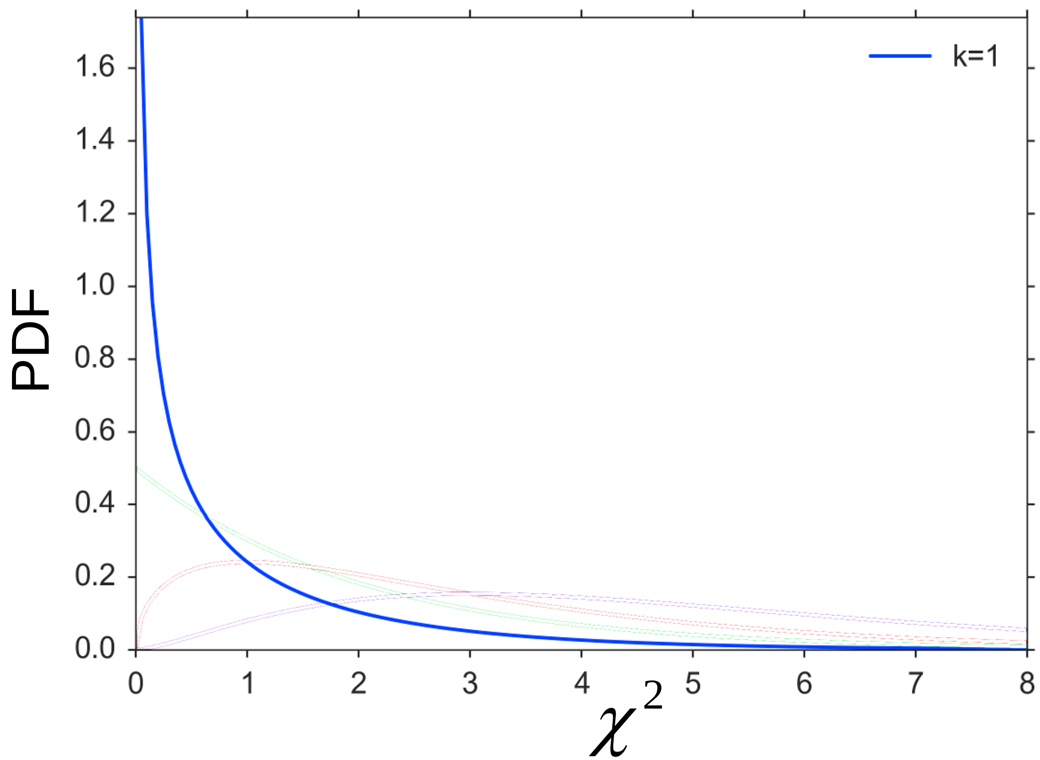
Chi-Square Distribution

- Small deviations are more expected than large deviations:

probability

density

function

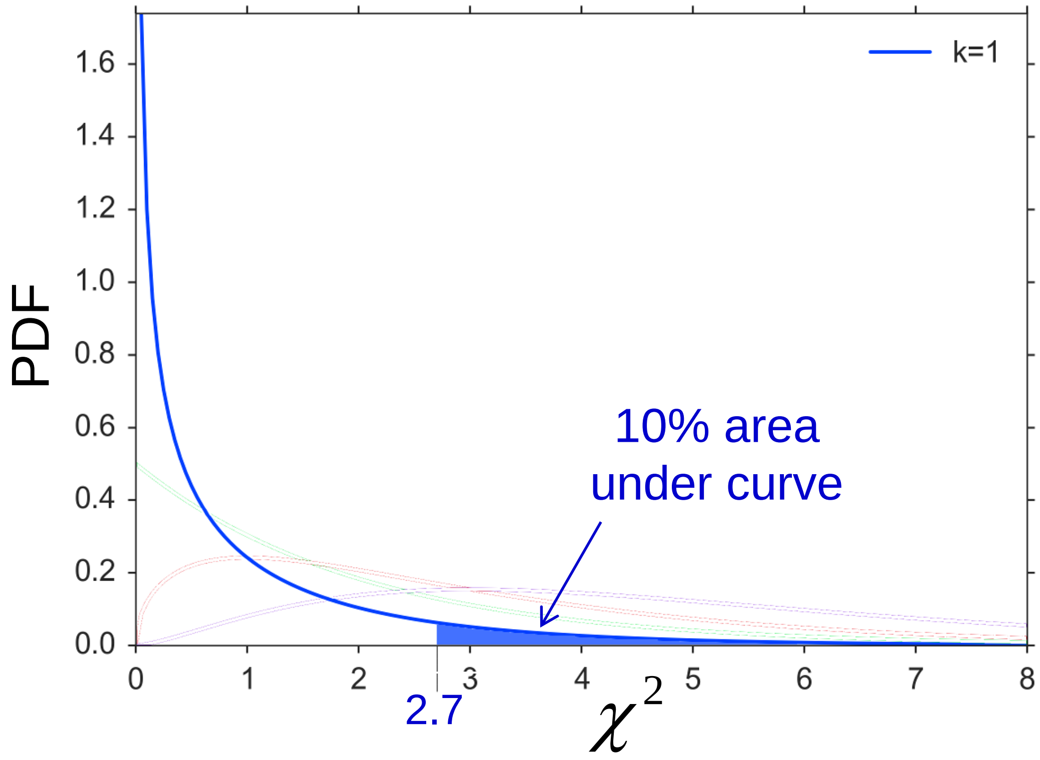


- If a random variable Y has a normal distribution (Gaussian) $\rightarrow$ Y^2 has a Chi-square distribution (with k=1 degree of freedom)

http://work.thaslwanter.at/Stats/html/statsDistributions.html

Chi-Square Test

- Some percentage of expected deviations is over a threshold



- E.g.: 10% of all χ^2 values are larger than critical value 2.7
- Values can be looked up in a chi-square distribution table

Chi-Square Distribution Table

probability level

α	0.5	**0.1**	**0.05**	0.02	0.01	0.001
	0.455	**2.706**	**3.841**	5.412	6.635	10.827

- α: significance level
- α=0.1: 90% of values are below critical value 2.706
- α=0.05: 95% of values are below critical value 3.841
- Earlier, we have found a value of 421.9

$\longrightarrow$ Our data are extremely unlikely given the null hypothesis

This means, that our result speaks in favor of rejecting the null hypothesis H_o under the selected α-level and accepting the alternative hypothesis H_1

Chi-Square Calculation: an Example

- X^2 (chi-square) calculation:

$$\chi^2 = \frac{(250-100)^2}{100} + \frac{(50-200)^2}{200} + \frac{(250-400)^2}{400} + \frac{(950-800)^2}{800} = \underline{421.9}$$

- It shows that "preferred reading" and "likes chess" are correlated in the group (since X^2 larger than 3.481, from X^2 table – a statistical measure for significance of 2x2 table)
- What if all numbers were 10x smaller (N=150 participants)?

$$\rightarrow \quad \chi^2 = \ldots = 42.19$$

- What if all numbers were 50x smaller (N=30 participants)?

$$\chi^2 = \frac{(5-2)^2}{2} + \frac{(1-4)^2}{4} + \frac{(5-8)^2}{8} + \frac{(19-16)^2}{16} = 8.4375$$

Chi-Square Calculation: Degrees of Freedom

		Category 1 Levels				Sum (row)
		L1	L2	…	LJ	
Category 2 Levels	L1					
	…					
	LI					
Sum (col.)						N

- If the two categories have several levels (*J* levels for category 1 and *I* levels for category 2), then there are more degrees of freedom in which the entries can differ
- *I x J* contingency table
- Number of degrees of freedom: (*I*-1) *x* (*J*-1)

Degrees of Freedom

“The number of degrees of freedom in a problem, distribution, etc., is the number of parameters which may be independently varied.”

(from Wolfram MathWorld)

Chi-Square Calculation: an Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250	250	500
Not like science fiction	50	950	1000
Sum (column)	300	1200	N=1500

- If one entry is changed …

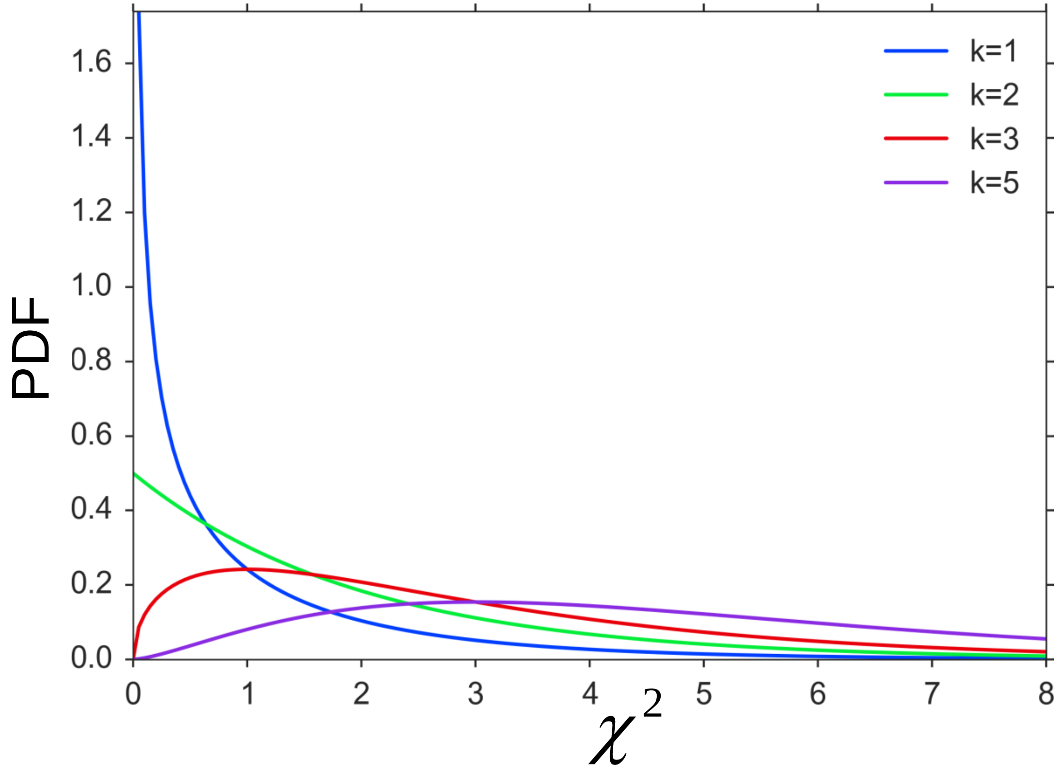
	Play chess	Not play chess	Sum (row)
Like science fiction	**200**	300	500
Not like science fiction	100	900	1000
Sum (column)	300	1200	N=1500

… this fixes all other entries

→ 1 degree of freedom

Chi-Square Calculation: Degrees of Freedom

- More degrees of freedom make larger deviations probable:



- Sum of k independ. random variables Y_i with normal distrib. $\rightarrow \Sigma_i Y_i^2$ has a χ^2 distribution with k degrees of freedom

Chi-square Distribution Table

probability level

DF	0.5	0.1	**0.05**	0.02	0.01	0.001
1	0.455	2.706	**3.841**	5.412	6.635	10.827
2	1.386	4.605	**5.991**	7.824	9.210	13.815
3	2.366	6.251	**7.815**	9.837	11.345	16.268
4	3.357	7.779	**9.488**	11.668	13.277	18.465
5	4.351	9.236	**11.070**	13.388	15.086	20.517

degrees of freedom

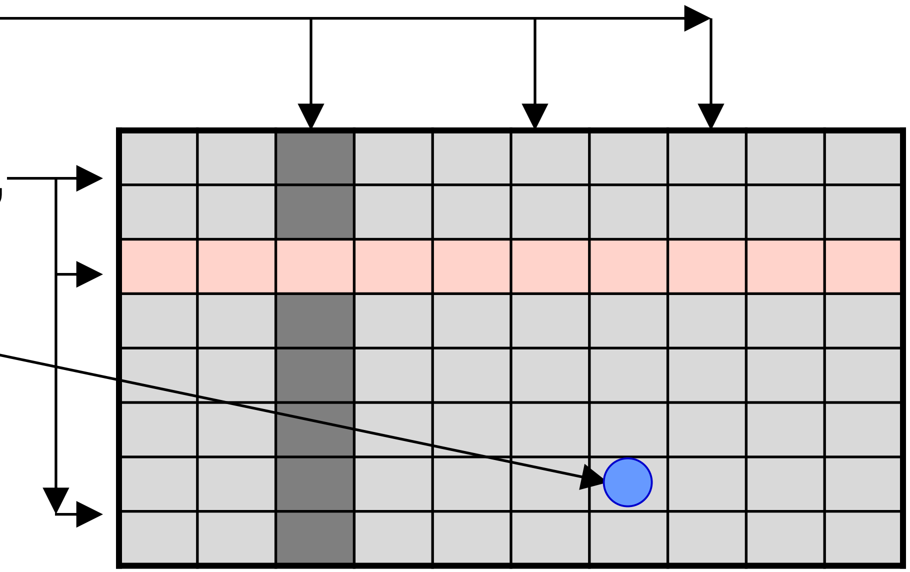
Data Preprocessing

- Data Cleaning
- Data Integration
- Values Reduction
  - Chi-Square Statistical Test (nominal data)
  - ChiMerge Discretization
  - Binning

Dimensions Reduction of Large Data Sets

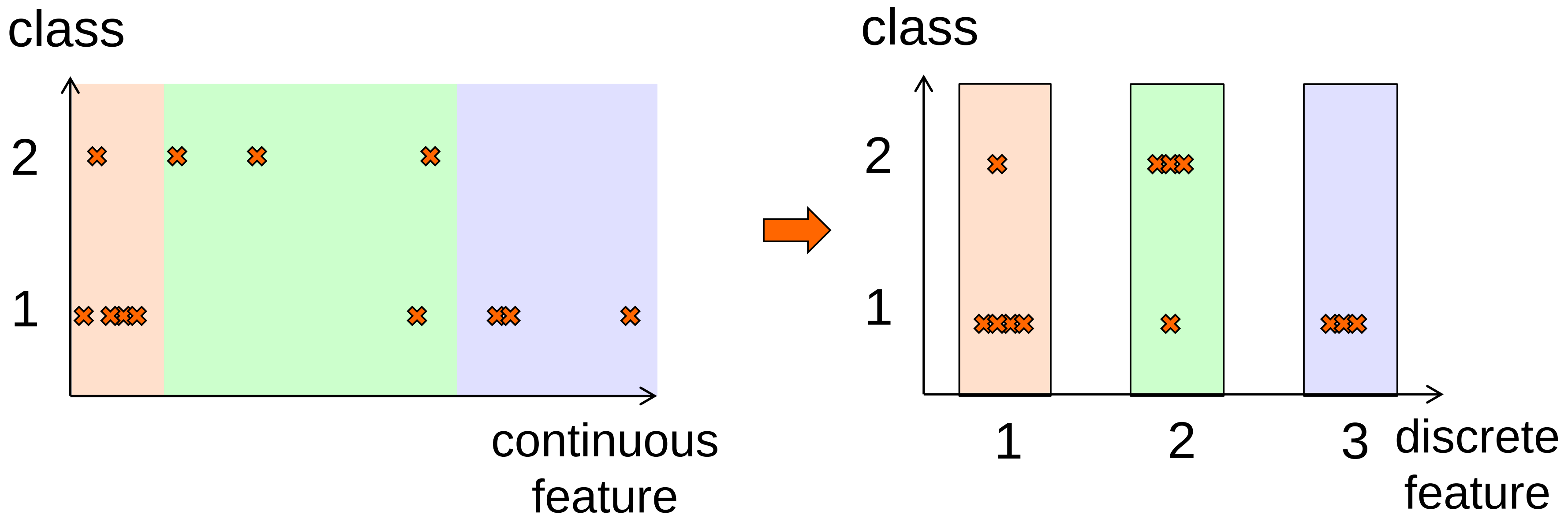
Many dimension reduction techniques rely on data transformations. Main dimensions are:

- ***columns*** (features),
- ***rows*** (cases or samples),
- ***values*** of the features for the given sample



Values Reduction – Our Aim

Data often have continuous features



Some classification algorithms prefer data with discrete attributes
(e.g. Decision Trees, later Lecture)

Values Reduction – ChiMerge Technique

Apply the Chi-square calculations for feature discretization

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***:

   3.1 Compute X^2 tests for each pair of adjacent intervals

   3.2 Merge two adjacent intervals with the lowest X^2 value, if calculated X^2 is less than threshold

   Until no X^2 test of any two adjacent intervals is less than threshold value

Values Reduction – Contingency Table

ChiMerge makes X^2 test for the 2x2 table of categorical data:

	Class 1	Class 2	Σ
Interval-1	A_{11}	A_{12}	R_1
Interval-2	A_{21}	A_{22}	R_2
Σ	C_1	C_2	N

X^2 test is:

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

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 = $\sum A_{ij}$, $j = 1,\ldots k$,
- C_j = number of instances in the j-th class = $\sum A_{ij}$, $i = 1,2$,
- N = total number of instances = $\sum R_i$, $i = 1,2$.

Test whether interval assignment and class label are correlated!

(→ merge intervals if uncorrelated)

Values Reduction – ChiMerge Example

Feature values

Class labels

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

0
2
5
7.5
8.5
.
.
.

Initial interval points

Values Reduction – ChiMerge Example

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

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

	Class 1	Class 2	Σ
Interval [7.5,8.5]	$A_{11}=1$	$A_{12}=0$	$R_1=1$
Interval [8.5,10]	$A_{21}=1$	$A_{22}=0$	$R_2=1$
Σ	$C_1=2$	$C_2=0$	N=2

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

E11 = 1*2/2 = 1,   E12 = 1*0/2 = 0,
E21 = 1*2/2 = 1, & E22 = 1*0/2 = 0

and corresponding X^2 test:

$$X^2 = (1-1)^2/1 + (0-0)^2/0 + (1-1)^2/1 + (0-0)^2/0 = \mathbf{0}$$

For d=1 degree of freedom: $X^2 = 0 < 2.706$ (α=0.1) → **merge !**

Values Reduction – ChiMerge Example

- … one of the following iterations:

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

	Class 1	Class 2	Σ
Interval [0.0,7.5]	$A_{11}=2$	$A_{12}=1$	$R_1=3$
Interval [7.5,10]	$A_{21}=2$	$A_{22}=0$	$R_2=2$
Σ	$C_1=4$	$C_2=1$	N=5

E11 = 3*4/5 = 2.4, E12 = 3*1/5 = 0.6,
E21 = 2*4/5 = 1.6, E22 = 2*1/5 = 0.4

$$X^2 = (2-2.4)^2/2.4 + (1-0.6)^2/0.6 + (2-1.6)^2/1.6 + (0-0.4)^2/0.4 = \mathbf{0.834}$$

We check the statistical table for the X^2 distribution and use d=1 (degree of freedom) and significance level $\alpha=0.1$. We obtain:

X^2 **= 0.834 < 2.706** ⟶ **merge !**

Values Reduction – ChiMerge Example

- … one of the further iterations:

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

	Class 1	Class 2	Σ
Interval [0.0,10]	$A_{11}=4$	$A_{12}=1$	$R_1=5$
Interval [10 ,42]	$A_{21}=1$	$A_{22}=3$	$R_2=4$
Σ	$C_1=5$	$C_2=4$	N=9

E11 = 2.78, E12 = 2.22,
E21 = 2.22, E22 = 1.78

$X^2 =$ 2.72 > 2.706 → **NO merge !**

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

Interval representatives: 5 (low) 26 (medium) 51 (high)

Values Reduction – ChiMerge Example

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

Original set

Final data set with reduced set of values F:

Sample:	F	K
1	5	1
2	5	2
3	5	1
4	5	1
5	5	1
6	26	2
7	26	2
8	26	1
9	26	2
10	51	1
11	51	1
12	51	1

Data Preprocessing

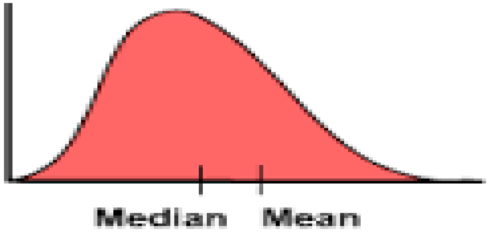
- Data Cleaning
- Data Integration
- Values Reduction
  - Chi-Square Statistical Test (nominal data)
  - ChiMerge Discretization
  - Binning

Data Discretization Methods

- Reduce number of values for given continuous attribute by dividing into intervals
  - ***Correlation analysis*** (e.g., χ^2) (bottom-up merge) ✓
  - ***Clustering analysis*** (unsupervised, top-down split or bottom-up merge) → later Lecture
  - ***Decision-tree analysis*** (supervised, top-down split) → later Lecture
  - ***Binning***: e.g. equal width binning and replacing bin by mean
    - Top-down split, unsupervised, no class information used

Simple Discretization: Binning

- ***Equal-width*** (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - If *A* and *B* are the min and max values of the attribute, the width of intervals will be: $w = (B - A)/N$.
  - Boundaries: *min+w, min+2w, ... , min+(N-1)w*
  - Simple method, but outliers may dominate partitioning
  - Skewed data is not handled well



Example: Data: 0, 4, 12, 16, 16, 18, 24, 26, 28

$N=3 \quad \rightarrow \quad w = (28-0)/3$

Bin 1: 0, 4

Bin 2: 12, 16, 16, 18

Bin 3: 24, 26, 28

Simple Discretization: Binning

- ***Equal-depth*** (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Example with the same data, and N = 3:

    0, 4, 12, 16, 16, 18, 24, 26, 28

    Bin 1: 0, 4, 12

    Bin 2: 16, 16, 18

    Bin 3: 24, 26, 28

Binning Methods for Data Smoothing

Sorted data for price (in $): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into equal-frequency (*equal-depth*) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

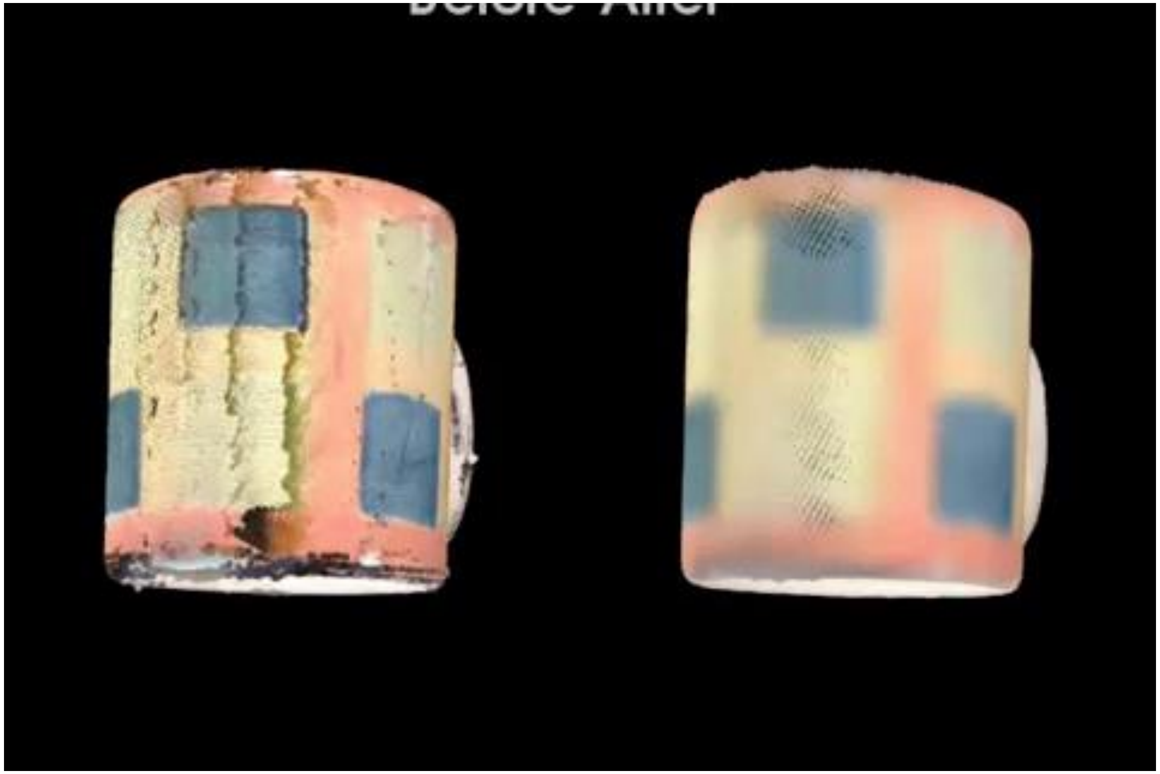
Smoothing by ***bin means***:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

Smoothing by ***bin boundaries***:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 21, 25
- Bin 3: 26, 26, 26, 34

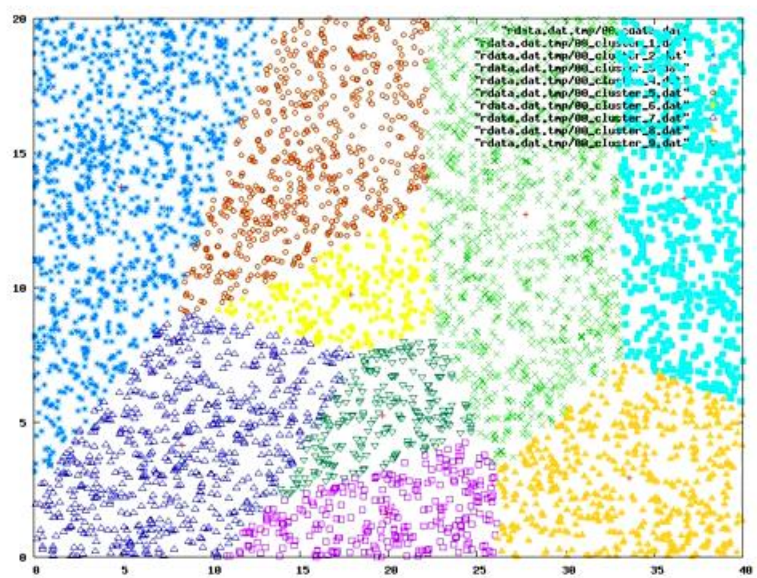
Example: Data Resampling and Smoothing in Point Cloud Application



Clustering

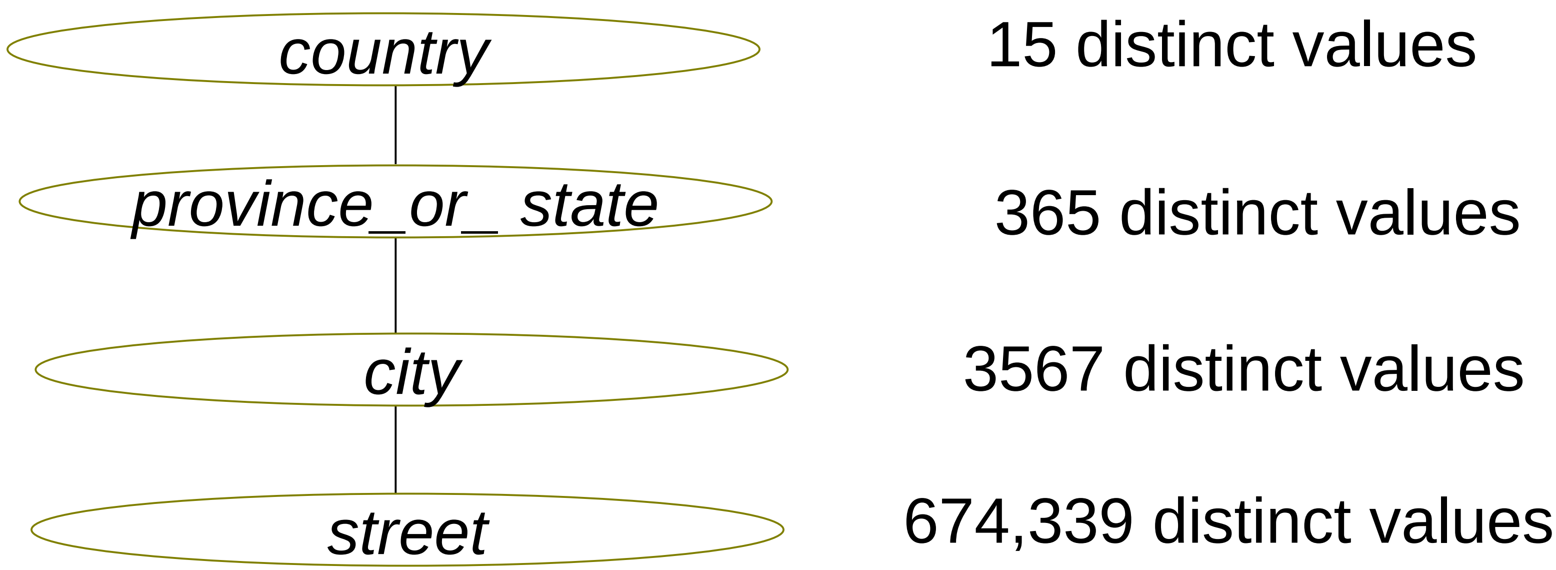
- Partition data set into clusters based on similarity (metrics), and store cluster representation (e.g., centroid and diameter) only
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering algorithms

  → later Lecture



Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy



- Exceptions, e.g., weekday, month, quarter, year

Summary

- ***Data quality***: accuracy, completeness, consistency, timeliness, believability, interpretability
- ***Data cleaning***: e.g. missing/noisy values, outliers
  - Regression to fill in missing values
- ***Data integration*** from multiple sources:
  - Entity identification problem; Remove redundancies; Detect inconsistencies
  - Chi-square for correlation analysis
- ***Data discretization***
  - ChiMerge, Binning, Clustering, Concept hierarchy generation