Data Mining: Data And Preprocessing

- **Data** [Sec. 2.1]
  - Transaction or market basket data
  - Attributes and different types of attributes
- **Exploring the Data** [Sec. 3]
  - Five number summary
  - Box plots
  - Skewness, mean, median
  - Measures of spread: variance, interquartile range (IQR)
- **Data Quality** [Sec. 2.2]
  - Errors and noise
  - Outliers
  - Missing values

- **Data Preprocessing** [Sec. 2.3]
  - Aggregation
  - Sampling
    - Sampling with(out) replacement
    - Stratified sampling
  - Discretization
    - Unsupervised
    - Supervised
  - Feature creation
  - Feature transformation
  - Feature reduction

Step 1: To describe the dataset

- What do your records represent?
- What does each attribute mean?
- What type of attributes?
  - Categorical
  - Numerical
    - Discrete
    - Continuous
  - Binary
  - Asymmetric
- …

What is Data?

- Collection of data objects and their attributes
- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc
  - **Attribute** is also known as **variable**, **field**, **characteristi**c, or **feature**
- A collection of attributes describe an object
  - **Object** is also known as **record**, **point**, **case**, **entity**, or **instance**

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Class

Transaction Data

- A special type of record data, where
  - each transaction (record) involves a set of items
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Transaction Data

- Transaction data can be represented as sparse data matrix: **market basket representation**
  - Each record (line) represents a transaction
  - Attributes are binary and asymmetric

Tid	*Bread*	*Coke*	*Milk*	*Beer*	*Diaper*
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - **Distinctness**: $= \neq$
  - **Order**: $< >$
  - **Addition**: + -
  - **Multiplication**: * /

  - **Nominal attribute**: distinctness
  - **Ordinal attribute**: distinctness and order
  - **Interval attribute**: distinctness, order and addition
  - **Ratio attribute**: all 4 properties

Types of Attributes

Categorical
- Nominal
  - **Ex**: ID numbers, eye color, zip codes
- Ordinal
  - **Ex**: rankings (e.g. taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

Numeric
- Interval
  - **Ex**: calendar dates, temperature in Celsius or Fahrenheit
- Ratio
  - **Ex**: length, time, counts, monetary quantities

Discrete, Continuous, & Asymmetric Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values

    Ex: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables
  - **Nominal, ordinal, binary attributes**
- **Continuous Attribute**
  - Has real numbers as attribute values
  - **Interval** and **ration** attributes

    Ex: temperature, height, or weight
- **Asymmetric Attribute**
  - Only presence is regarded as important

    Ex: If students are compared on the basis of the courses they do not take, then most students would seem very similar

Step 2: To explore the dataset

- Preliminary investigation of the data to better understand its specific characteristics
  - It can help to answer some of the data mining questions
  - To help in selecting pre-processing tools
  - To help in selecting appropriate data mining algorithms
- Things to look at
  - Class balance
  - Dispersion of data attribute values
  - Skewness, outliers, missing values
  - Attributes that vary together
  - …
- Visualization tools are important [Sec. 3.3]
  - Histograms, box plots, scatter plots
  - …

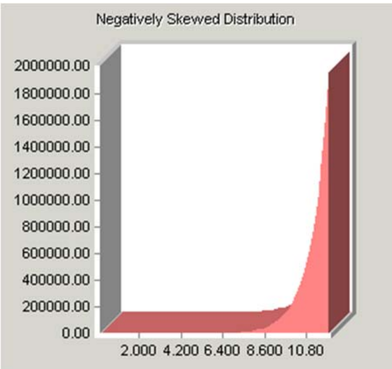
Class Balance

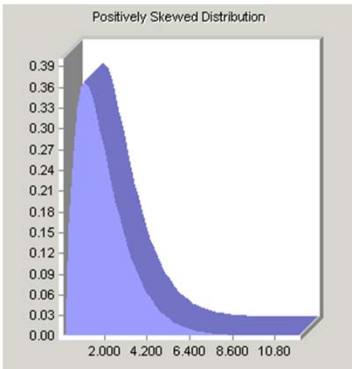
- Many datasets have a discrete (binary) attribute class
  - What is the frequence of each class?
  - Is there a considerable less frequent class?
- Data mining algorithms may give poor results due to class imbalance problem
  - Identify the problem in an initial phase

Useful statistics

- Discrete attributes
  - Frequency of each value
  - **Mode** = value with highest frequency
- Continuous attributes
  - Range of values, i.e. **min** and **max**
  - **Mean** (average)
    - Sensitive to outliers
  - **Median**
    - Better indication of the ”middle” of a set of values in a skewed distribution
  - **Skewed distribution**
    - mean and median are *quite* different

Skewed Distributions of Attribute Values





Five-number summary

- For numerical attribute values

  (minimum, Q_1, Q_2, Q_3, maximum)

Attribute values: **6 47 49 15 42 41 7 39 43 40 36**

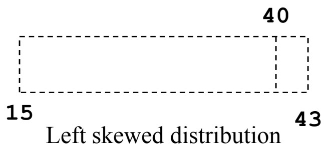
Sorted: **6 7 15 36 39 40 41 42 43 47 49**

$Q_1 = 15$ lower quartile

Q_2 = median = 40 (*mean = 33.18*)

$Q_3 = 43$ upper quartile

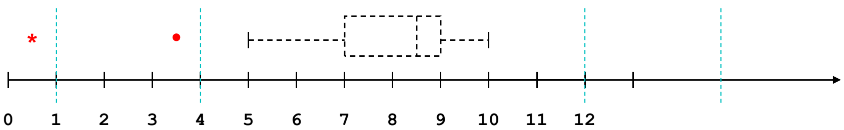
$Q_3 - Q_1 = 28$ interquartile range



Left skewed distribution

Box Plots

http://www.shodor.org/interactivate/activities/BoxPlot/



- $Q_1 = 7$ Q_1 = median = 8.5 $Q_3 = 9$
- Interquartile range IQR = $Q_3 - Q_1 = 2$
- Largest non-outlier = 10 (right whisker)
- Smallest non-outlier = 5 (left whisker)
- Mild outlier (**mo**) = 3.5

  Q_1-3×1.5 IQR ≤ **mo** ≤ Q_1-1.5 IQR

  Q_3+1.5 IQR < **mo** ≤ Q_3+3×1.5 IQR

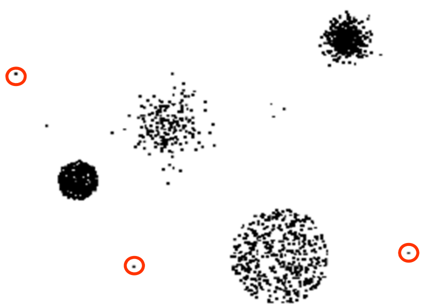
Available in **WEKA**
- **`Filters`**
  `InterquartileRange`

- Extreme outlier(**eo**) = 0.5

  eo < Q_1-3×1.5 IQR or **eo** > Q3+3×1.5 IQR

Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set



• Outliers can be legitimate objects or values

• Outliers may be of interest

Network intrusion detection

Fraud detection

• Some algorithms may produce poor results in the presence of outliers

Identify and remove them

Box Plots

- A box plot can provide information useful information about an attribute
  - sample's range
  - median
  - normality of the distribution
  - skew (asymmetry) of the distribution
  - plot extreme cases within the sample

Dispersion of Data

- How do the values of an attribute spread?
  - Variance

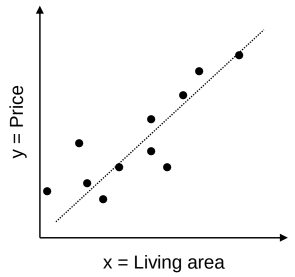
$$\text{variance}(A) = S_A^2 = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \overline{A})^2$$

    - Variance is sensitive to outliers
  - Interquartile range (**IQR**)
  - What if the distribution of values is multimodal, i.e. data has several *bumps*?
    - Vizualization tools are useful

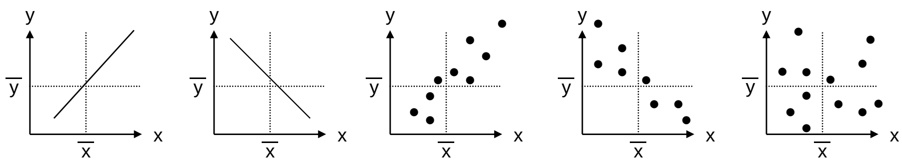
Attributes that Vary Together

- **Correlation** is a measure that describes how two attributes vary together [Sec. 3.2.5]

There is a linear correlation between x and y.



$$\text{corr}(x, y) = \frac{S_{xy}}{S_x S_y}$$



corr(x,y) = 1 corr(x,y) = -1 0 < corr(x,y) < 1 -1 < corr(x,y) < 0 corr(x,y) = 0

Step 3: Data Preprocessing

- Data is often collected for unspecified applications
  - Data may have quality problems that need to be addressed before applying a data mining technique
    - Noise and outliers
    - Missing values
    - Duplicate data

- **Preprocessing** may be needed to make data more suitable for data mining

"If you want to find gold dust, move the rocks out of the way first!"

Data Preprocessing

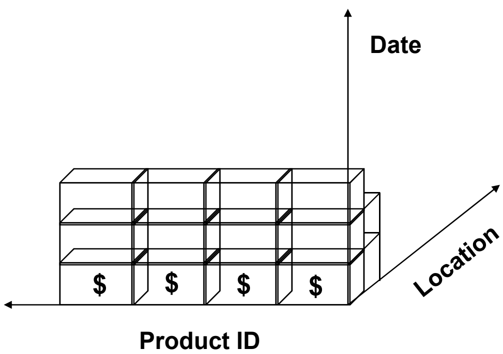
- Data transformation might be need
  - Aggregation
  - Sampling [[sec. 2.3.2]
  - Feature creation
  - Feature transformation
    - Normalization (back to it when clustering is discussed)
  - Discretization [sec. 2.3.6]
  - Feature reduction [sec. 2.3.4]

Missing Values

- Handling missing values
  - Eliminate objects with missing values
    - Not more than 5% of the records
  - Estimate missing values
    - Replace by most frequent or average
  - Use non-missing data to predict the missing values
    - Linear regression
    - Maintain the between-attribute relationships
    - Different replacements can be generated for the same attribute
  - Use expert knowledge
  - Apply a data mining technique that can cope with missing values (e.g. decision trees)

Aggregation

- Combining two or more objects into a single object.



- Reduce the possible values of date from 365 days to 12 months.
- Aggregating the data per store location gives a view per product monthly.

Attribute "Location" is eliminated

Online Analytical Processing

(**OLAP**)

[Sec. 3.4.2]

Aggregation

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - High-level view of the data
    - Easier to discover patterns
  - More “stable” data
    - Aggregated data tends to have less variability

Sampling [sec. 2.3.2]

- **Sampling** is a technique employed for selecting a subset of the data
- Why is it used in data mining?
  - It may be too expensive or too time consuming to process all data
  - To measure a classifier's performance the data may be divided in a training set and a test set
  - To obtain a better balance between class distributions

Sampling

- Sampling techniques should create representative samples
  - A sample is representative, if it has approximately the same property (of interest) as the original set of data
    - **Ex**: Each of the classes in the full dataset should be represented in about the right proportion in the training and test sets
  - Using a representative sample will work almost as well as using the full datasets
- There are several sampling techniques
  - Which one to choose?

Sampling Techniques

- Simple Random Sampling
  - Every sample of size n has the same chance of being selected
  - Perfect random sampling is difficult to achieve in practice
  - Use random numbers

Random sampling

- Sampling without replacement

  A selected item cannot be selected again - removed from the full dataset once selected

- Sampling with replacement

  Items can be picked up more than once for the sample – not removed from the full dataset once selected

  Useful for small data sets

- **Drawback**: by bad luck, all examples of a less frequent (rare) class may be missed out in the sample

Sampling Techniques

- Stratified sampling
  - Split the data into several partitions (strata); then draw random samples from each partition
  - Each strata may correspond to each of the possible classes in the data
  - The number of items selected from each strata is proportional to the strata size
  - However, stratification provides only a primitive safeguard against uneven representation of classes in a sample

Filters in Weka

- Filters – algorithms that transform the input dataset in some way

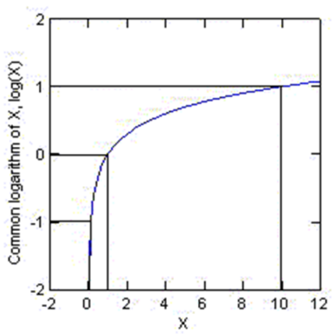
Filters		
Unsupervised	Attribute filter	**`ReplaceMissingValues`** **`NumericTransform`**
	Instance filter	**`Resample`**
Supervised		
	Attribute filter	**`AttributeSelection`** **`Discretize`**
	Instance filter	**`Resample`** **`SpreadSubsample`**

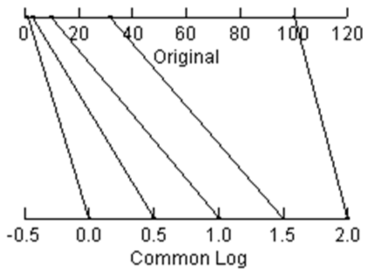
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature Extraction
    - domain-specific
    - From an image as a set of pixels one might extract features such as whether certain types of edges are present
  - Feature Construction
    - combining features **Ex**: *density = mass / volume*

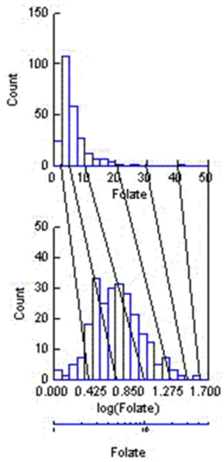
Feature Transformation

- A function is applied to each value of an attribute
  - Use $\log_{10} x$ to transform data that does not have a normal distribution into data that does
    - See http://www.jerrydallal.com/LHSP/logs.htm for more details





The logarithmic transformation



Discretization [Sec. 2.3.6]

- To transform a continuous attribute into a categorical attribute
  - Some data mining algorithms only work with discrete attributes
    - E.g. Apriori for ARM
  - Better results may be obtained with discretized attributes

Discretization

- **Unsupervised discretization**
  - Equal-interval binning
  - Equal-frequency binning

  • **Class labels are ignored**

  • The best number of bins k is determined experimentally

- **Supervised discretization**
  - Entropy-based discretization
  - It tries to maximize the "purity" of the intervals (i.e. to contain as less as possible mixture of class labels)

Unsupervised Discretization

- **Equal-interval binning**
  - Divide the attribute values x into k equally sized bins
  - If $\mathbf{x_{min}} \leq \mathbf{x} \leq \mathbf{x_{max}}$ then the bin width $\boldsymbol{\delta}$ is given by

$$\delta = \frac{x_{max} - x_{min}}{k}$$

  - Construct bin boundaries at $\mathbf{x_{min}} + \mathbf{i\delta}$, i = 1,…, k-1

- **Disadvantage**: Outliers can cause problems

Unsupervised Discretization

– **Equal-frequency binning**

- An equal number of values are placed in each of the k bins.

- **Disadvantage**: Many occurrences of the same continuous value could cause the values to be assigned into different bins

Supervised Discretization

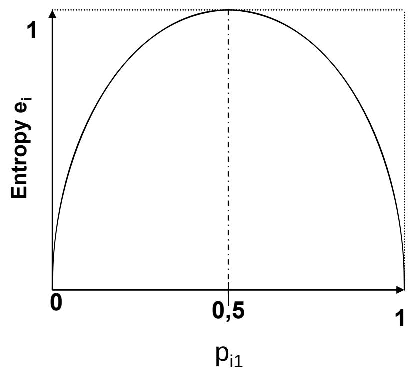
– **Entropy-based discretization**

- The main idea is to split the attribute's value in a way that generates bins as "pure" as possible
- We need a measure of "**impurity of a bin**" such that
  – A bin with uniform class distribution has the highest impurity
  – A bin with all items belonging to the same class has zero impurity
  – The more skewed is the class distribution in the bin the smaller is the impurity

- **Entropy** can be such measure of impurity

Entropy of a Bin i

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$



Two class problem
K = 2

Entropy

- n   number of bins
- m   total number of values
- k   number of class labels
- m_i number of values in the ith bin
- m_{ij} number of values of class j in the ith bin
- $p_{ij} = m_{ij} / m_i$
- $w_i = m_i / m$

$$\text{entropy} = \sum_{i+1}^{n} w_i e_i$$

Splitting Algorithm

Splitting Algorithm

```
1.  Sort the values of attribute X (to be
    discretized) into a sorted sequence S;
2.  Discretize(S);
```

Discretize(S)

```
while ( StoppingCriterion(S) == False ) {
    % minimize the impurity of left and right bins
    % if S has n values then n-1 split points need to be considered
    (leftBin, rightBin) = GetBestSplitPoint(S);
    Discretize(leftBin);
    Discretize(rightBin);
}
```


Discretization in Weka

Attribute Filter		Options
Unsupervised	`Discretize`	`bins`
		`useEqualFrequency`
Supervised	`Discretize`	

Feature Reduction [sec. 2.3.4]

- Purpose:
  - Many data mining algorithms work better if the number of attributes is lower
    - More easily interpretable representation of concepts
    - Focus on the more relevant attributes
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to reduce noise
- Techniques
  - **Single attribute evaluators**
  - **Attribute subset evaluators**
    - A search strategy is required

Feature Reduction

- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - **Ex**: students' ID is often irrelevant to the task of predicting students grades
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - **Ex**: price of a product and the amount of sales tax paid
  - Select a subset of attributes whose pairwise correlation is low

Single Attribute Evaluators

1. Measure how well each attribute individually helps to discriminate between each class
   - Which measure to use?
     - Information gain Weka: **`InfoGainAttributeEval`**
     - Chi-square statistic Weka: **`ChiSquareAttributeEval`**
2. Rank all attributes
3. The user can then discard all attributes that do not meet a specified criterion
   - e.g. retain the best 10 attributes

Single Attribute Evaluator: Information Gain

- How much information is gained about the **classification** of a record by knowing the value of **A**?
  - Assume **A** has three possible values $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$
  - Using attribute **A**, it is possible to divide the data **S** into 3 subsets
    - $\mathbf{S}_1$ is the set of records with $\mathbf{A = v}_1$
    - $\mathbf{S}_2$ is the set of records with $\mathbf{A = v}_2$
    - $\mathbf{S}_3$ is the set of records with $\mathbf{A = v}_3$

InfoGain(A) = Entropy(S) –

$$[w_1 \times \text{Entropy}(S_1) + w_2 \times \text{Entropy}(S_2) + w_3 \times \text{Entropy}(S_3)]$$

Attribute Subset Evaluators

- Use a search algorithm to search through the space of possible attributes to find a "*suitable*" sub-set
  - How to measure the predictive ability of a sub-set of attributes?
  - What search strategy to use?
- Principal component analysis

How to Measure the Predictive Ability of a Set of Attributes?

- Measure how well each attribute correlates with the class but has little correlation with the other attributes
  - Weka: **`CfsSubsetEval`**
- Use a classifier to evaluate the attribute set
  - Choose a classifier algorithm
  - The **accuracy** of the classifier is used as the predictive measure
    - **accuracy** = number of correctly classified records in training or a test dataset
  - Weka: **`ClassifierSubsetEval`**

    accuracy estimated on a training or separate test data set

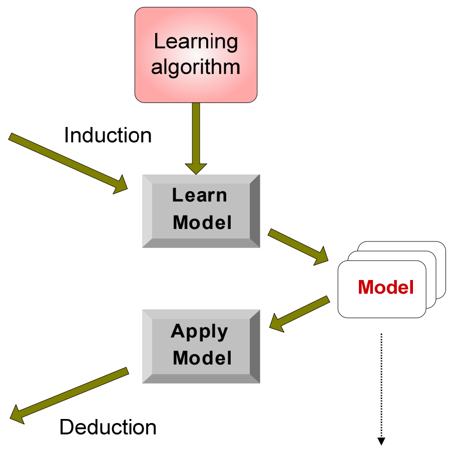
    `WrapperSubsetEval`

    accuracy estimated by cross-validation

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

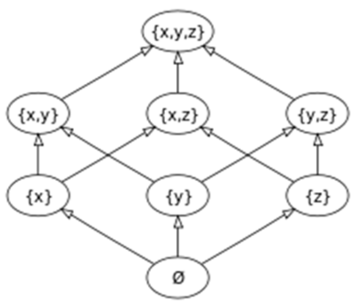
Test Set

(Attrib1 = yes) → Class = No

(Attrib1 = No) ∧ (Attrib3 < 95K) → Class = Yes

What Search Strategy to Use?

- Exhaustive search
  - Visits all 2^n sub-sets, for n attributes
- Greedy search
- Best first search
- Random search
- …



Weka: Select Attributes

Single Attribute Evaluator	Ranking Method
InfoGainAttributeEval	Ranker
ChiSquareAttributeEval	

- Ranker options:
  - startSet
  - threshold

Weka: Select Attributes

Attribute Subset Evaluator	Classifier Algorithm	Search Method
CfsSubsetEval	-	ExhaustiveSearch
ClassifierSubsetEval	Rules	BestFirst
WrapperSubsetEval	Trees	GreedyStepwise
		RandomSearch