Data Preprocessing Tasks

1. Data Cleaning

2. Data Transformation
   Next, let’s look at this task.

3. Data Reduction

4. Discretization
Data Transformation

- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scale data to fall within a small, specified range
- Feature construction: new features constructed from the given ones
Aggregation

• Sometimes “less is more.”
• Aggregation is the combining of two or more objects into a single object.
Normalization or Standardization

The goal of standardization or normalization is to make an entire set of values have a particular property.
Data Transformation: Normalization

- min-max normalization
- z-score normalization
- Normalization by decimal scaling
Min-Max Normalization

Transform the data from measured units to a new interval from $\text{new}_{\text{min}}_F$ to $\text{new}_{\text{max}}_F$ for feature $F$:

$$v' = \frac{v - \text{min}_F}{\text{max}_F - \text{min}_F} (\text{new}_{\text{max}}_F - \text{new}_{\text{min}}_F) + \text{new}_{\text{min}}_F$$

where $v$ is the current value of feature $F$. 
Min-Max Normalization: Example

Suppose that the minimum and maximum values for the feature income are $120,000 and $98,000, respectively. We would like to map income to the range $[0.0, 1.0]$. By min-max normalization, a value of $73,600$ for income is transformed to:

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0.0) + 0 = 0.716$$
z-score (zero-mean) Normalization

Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, \( v \), of \( A \) is normalized to \( v' \) by computing:

\[
v' = \frac{v - \bar{F}}{\sigma_F}
\]

where \( \bar{F} \) and \( \sigma_F \) are the mean and standard deviation of feature \( F \), respectively.
z-score (zero-mean) Normalization: Example

Suppose that the mean and standard deviation of the values for the feature income are $54,000 and $16,000, respectively. With z-score normalization, a value of $73,6000 for income is transformed to

$$\frac{73,600 - 54,000}{16,000} = 1.225$$
Decimal Scaling Normalization

Transform the data by moving the decimal points of values of feature $F$. The number of decimal points moved depends on the maximum absolute value of $F$. A value $v$ of $F$ is normalized to $v'$ by computing:

$$v' = \frac{v}{10^j},$$

where $j$ is the smallest integer such that $Max(|v'|) < 1.$
Decimal Scaling Normalization

Suppose that the recorded values of $F$ range from $-986$ to $917$. The maximum absolute value of $F$ is $986$. To normalize by decimal scaling, we therefore divide each value by $1,000$ (i.e., $j = 3$) so that $-986$ normalizes to $-0.986$ and $917$ normalizes to $0.917$. 
Data Preprocessing Tasks

1. Data Cleaning
2. Data Transformation
3. Data Reduction
   Next, let’s look at this task.
4. Discretization
Data Reduction

• Do we need all the data?
• Data mining/analysis can take a very long time
• Computational complexity of algorithms
Data Reduction

• Sampling: selecting a subset of the data objects to be analyzed.
• Feature selection: selecting a subset of the features to be analyzed.
• Dimensionality reduction: creating new features that are a combination of the old features.
Data Sampling

- Sampling is commonly used approach for selecting a subset of the data to be analyzed.
- Typically used because it is too expensive or time consuming to process all the data.
- Key idea:

  *Obtain a representative sample of the data.*
Sampling With or Without Replacement

SRSWOR (simple random sample without replacement)

SRSWR
Systematic Sampling

Select instances from an ordered sampling window.

**Equal-probability method:** First, select an element from the list at random. Then, every \( k \)th element in the window is selected, where \( k \), the sampling interval, is:

\[
k = \frac{N}{n}
\]

*Risk of interaction with unsuspected regularity in the data.*
Simple Random Sampling

Shuffle the data and then select examples.

Avoids regularities. But what if the dataset is imbalanced?
Stratified Random Sampling

• Assume original known class distribution in data is $\Phi$.
• Sample from the data such that $\Phi$ is maintained.
• New sample of data reflects original distribution.

Even works for imbalanced data. But often inefficient.
Cluster Sampling

- Group or segment the data based on similarities.
- Randomly select from each group.

Efficient, but won’t necessarily optimize performance.
Stratified or Cluster Sampling

Simple Random Sample

Stratified/Cluster Sample
Imbalanced Data

• Sometimes, classes have very unequal frequencies.
  – Attrition prediction: 97% stay, 3% attrite (in a month)
  – Medical diagnosis: 90% healthy, 10% diseased
  – eCommerce: 99% don’t buy, 1% buy
  – Security: > 99.99% of Americans are not terrorists

• Similar situation with multiple classes.
• Predictions can be 97% correct, but useless.
Handling Imbalanced Data

• With two classes: let positive targets be a minority.
• Separate raw held-aside set (e.g., 30% of data) and raw training.
• Select remaining positive targets (e.g., 70% of all targets) from raw training.
• Join with equal number of negative targets from raw training, and sort it.
• Separate randomized balanced set into balanced training and balanced testing.
Feature Selection

• Select a minimal set of features such that the probability distribution of the class is close to the one obtained by all the features.

• A good feature vector is defined by its capacity to discriminate between examples from different classes.
  • Maximize the inter-class separation and minimize the intra-class separation.
Feature Selection Properties

- Linear separability
- Non-linear separability
- Highly correlated features
- Multi-modal
Feature Properties

- Linear Separability
- Non-Linear Separability
- Multi-Model
- Highly Correlated Features
Feature Selection

• Given a feature set \( X = \{x_i| i = 1 \ldots N\} \), find a subset \( XM = \{x_{i1}, x_{i2}, \ldots, x_{iM}\} \), with \( M < N \), that optimizes an objective function \( \Phi(X) \), ideally error minimization.

• But remember, finding optimal number of features is an approximation.
Feature Selection: The Key Idea

“There is no problem-independent or privileged or ‘best’ set of features or feature attributes.”

—Ugly Duckling Theorem
Feature Selection Properties

- Problem with four features, where 0 shows feature exclusion and 1 feature inclusion.
- Potentially, any node in the lattice may be reached from any start point, by repeated actions.
- The lattice for an $n$ dimensional feature set can have $2^n$ nodes.
Curse of Dimensionality: Key Idea

“A function defined in high dimensional space is likely to be much more complex than a function defined in a lower-dimensional space, and those complications are harder to discern.”

—Milton Friedman (Famous Dude)
Feature Selection Schemes

• Filters
  – Use a feature selection technique or heuristic to rank-order relevant features based on certain value $v$.
  – Select features with $v$ exceeding certain threshold.

• Wrapper
  – Stepwise forward selection.
  – Stepwise backward elimination.
  – Combine forward selection and backward elimination.
Filters

• Examples: information gain, correlation measures, RELIEF-F, odds ratio, PCA
Filters

• Advantages
  – Typically faster execution
  – Non iterative process
  – Generality
  – Without an inductive bias of an underlying classifier as in the wrapper

• Disadvantage
  – Require an arbitrary cut-off strategy
  – Does not capture inter-feature interactions
Wrappers

• Examples: sequential stepwise search techniques, bidirectional search.
Wrappers

• Advantages
  – Tuned for the underlying inductive bias

• Disadvantage
  – Slow execution: reduce (almost) exhaustive search
  – Lack of generality
    • Tuned for a particular classifier
Wrappers
Let’s see some data transformation!