Statistical Learning and Data Mining
(Notes)

Data: Collection of data objects and their attributes.

Attribute:

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, characteristic, or feature

Object:

A collection of attributes describe an object

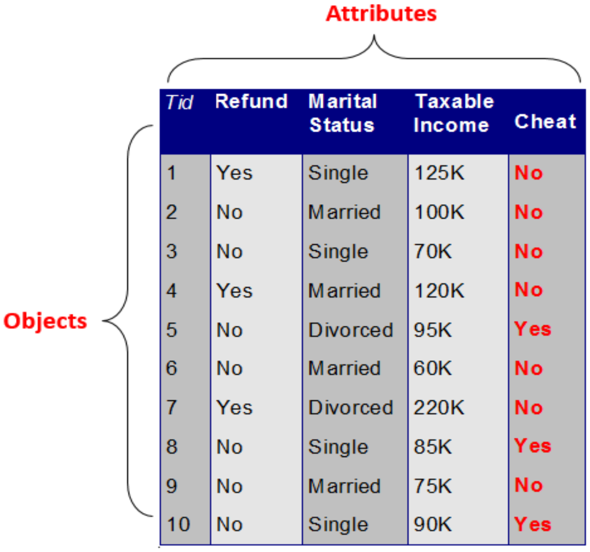
- Object is also known as record, point, case, sample, entity, or instance

Types of Attributes

There are different types of attributes

- Nominal
  - ❖ Examples: ID numbers, eye color, zip codes
- Ordinal
  - ❖ Examples: rankings, grades,
- Interval
  - ❖ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio
  - ❖ Examples: temperature in Kelvin, length, time, counts

Record Data: Data that consists of a collection of records, each of which consists of a fixed set of attributes


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

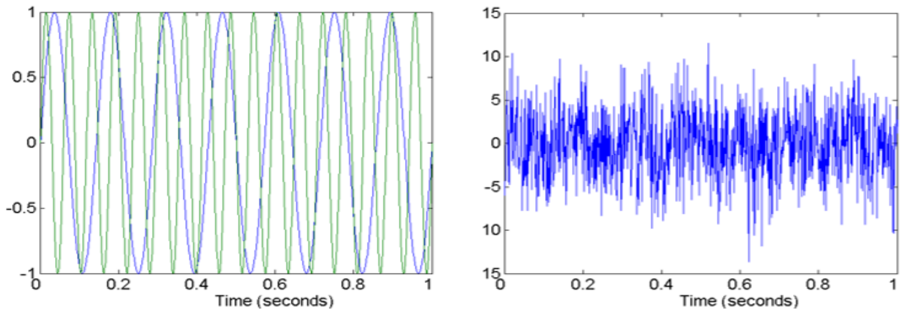
Data Quality

Examples of data quality problems:

- Noise and outliers
- missing values
- duplicate data

Noise

Noise refers to modification of original values



Two Sine Waves

Two Sine Waves + Noise

Outliers

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

Missing Values

Reasons for missing values

- Information is not collected
  (e.g., people decline to give their age and weight)

- Attributes may not be applicable to all cases
  (e.g., annual income is not applicable to children)

Handling missing values

- Eliminate Data Objects
- Estimate Missing Values
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

Duplicate Data

Data set may include data objects that are duplicates, or almost duplicates of one another

- Major issue when merging data from heterogeneous sources

Examples:

- Same person with multiple email addresses

Data cleaning:

- Process of dealing with duplicate data issues

Data Preprocessing

1. Aggregation
2. Sampling
3. Dimensionality Reduction
4. Feature subset selection
5. Feature creation
6. Discretization and Binarization
7. Attribute Transformation

1. Aggreation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - ❖ Data reduction (Reduce the number of attributes or objects)
  - ❖ Change of scale (Cities aggregated into regions, states, countries, etc)
  - ❖ More “stable” data (Aggregated data tends to have less variability)

2. Sampling

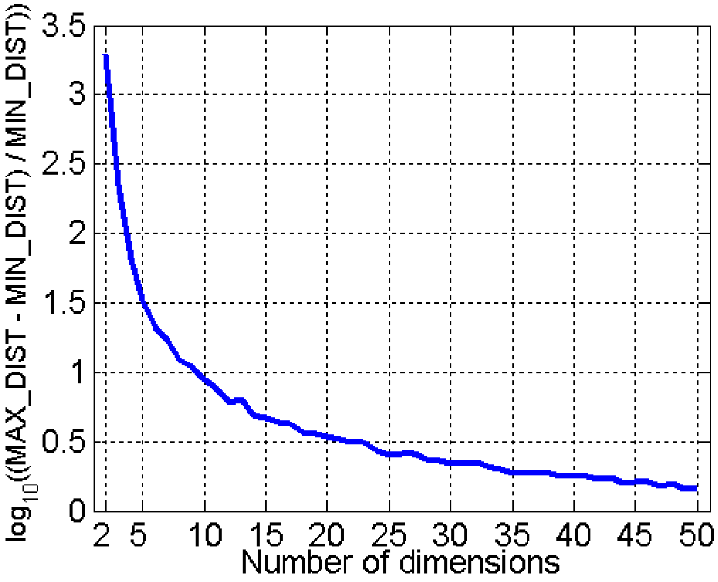
Sampling is the main technique employed for data selection.

- It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because **obtaining** the entire set of data of interest is too expensive or time consuming.

Curse of Dimensionality

When dimensionality increases, data becomes increasingly sparse in the space that it occupies

Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



3. Dimensionality Reduction

Purpose

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

Techniques

- Principle Component Analysis
- Singular Value Decomposition
- Others: supervised and non-linear techniques

4. Future Subset Selection

Another way to reduce dimensionality of data

Redundant features

- duplicate much or all of the information contained in one or more other attributes
- Example: purchase price of a product and the amount of sales tax paid

Irrelevant features

- contain no information that is useful for the data mining task at hand
- Example: students' ID is often irrelevant to the task of predicting students' GPA

Techniques:

- Brute-force approch:
  - ❖ Try all possible feature subsets as input to data mining algorithm
- Embedded approaches:
  - ❖ Feature selection occurs naturally as part of the data mining algorithm
- Filter approaches:
  - ❖ Features are selected before data mining algorithm is run

5. Feature Creation

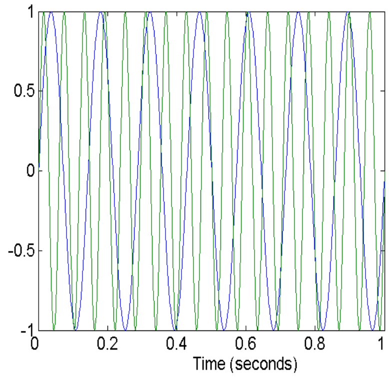
Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

Three general methodologies:

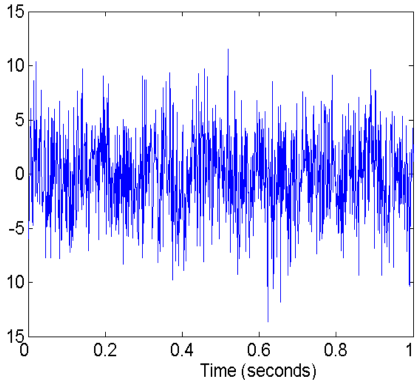
- Feature Extraction (domain-specific)
- Mapping Data to New Space
- Feature Construction (combining features)

Mapping Data to a New Space

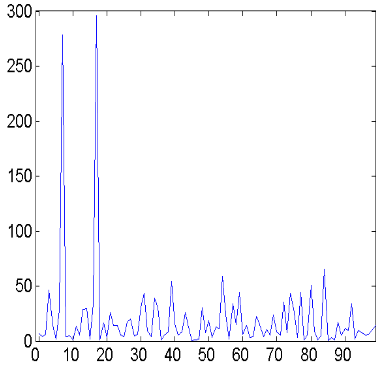
- Fourier transform
- Wavelet transform



Two Sine Waves



Two Sine Waves + Noise



Frequency

Similarity and Dissimilarity

Similarity

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range [0, 1]

Dissimilarity

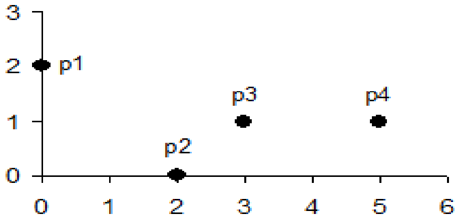
- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
-

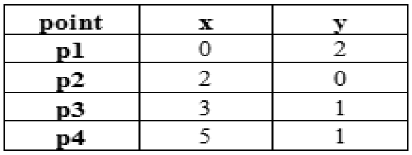
Proximity refers to a similarity or dissimilarity

Euclidean Distance

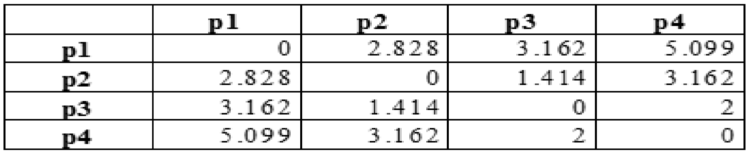
$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q. Standardization is necessary, if scales differ.





point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1



	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\sum^{-1}(p-q)^T$$

Algorithm: self study

Common Properties of a Similarity

Similarities, also have some well known properties.

1. s(p, q) = 1 (or maximum similarity) only if p = q.
2. s(p, q) = s(q, p) for all p and q. (Symmetry)

Where s(p, q) is the similarity between points (data objects), p and q.

Similarity between Binary Vectors

- Common situation is that objects, p and q, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

Cosine Similarity

If d_1 and d_2 are two document vectors, then

$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,

Where $\bullet$ indicates vector dot product and || d || is the length of vector d.

<u>Example:</u>

$d_1 =$ **3 2 0 5 0 0 0 2 0 0**

$d_2 =$ **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = 0.3150$

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product
-

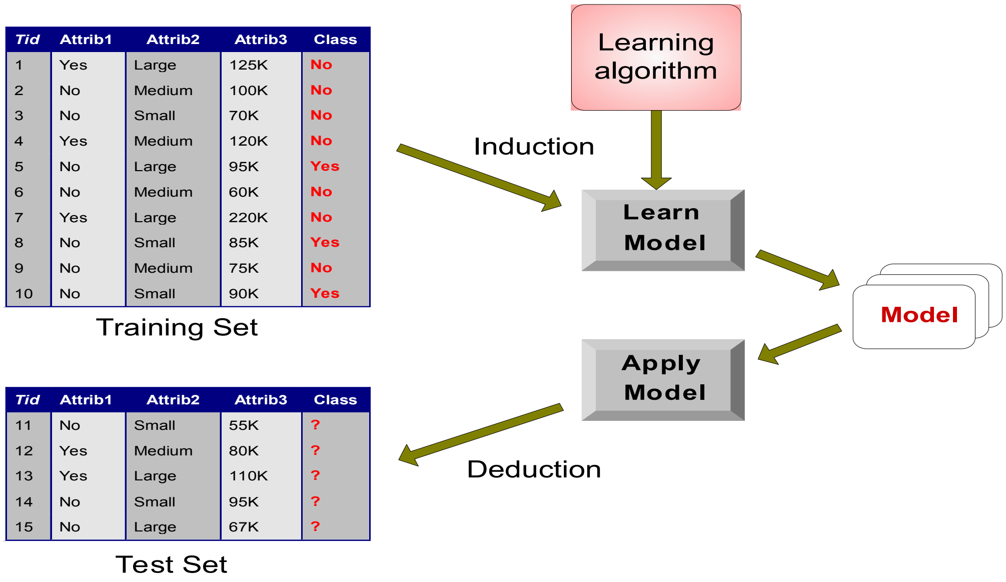
$$p'_k = (p_k - mean(p))/std(p)$$
$$q'_k = (q_k - mean(q))/std(q)$$
$$correlation(p,q) = p' \cdot q'$$

Density

Density-based clustering require a notion of density

- Euclidean density
- Probability density
- Graph-based density

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

Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Categorizing news stories as finance, weather, entertainment, sports, etc

Classification Techniques

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines
-

Evaluation Metrics

1. RMSE (Root Mean Square Error)

It represents the sample standard deviation of the differences between predicted values and observed values (called residuals). Mathematically, it is calculated using this formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

2. MAE

MAE is the average of the absolute difference between the predicted values and observed value. The MAE is a linear score which means that all the individual differences are weighted equally in the average. For example, the difference between 10 and 0 will be twice the difference between 5 and 0. However, same is not true for RMSE which we will discuss more in details further. Mathematically, it is calculated using this formula:

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

3. R Squared (R^2) and Adjusted R Squared

R Squared & Adjusted R Squared are often used for explanatory purposes and explains how well your selected independent variable(s) explain the variability in your dependent variable(s). Both these metrics are quite misunderstood and therefore I would like to clarify them first before going through their pros and cons.

Mathematically, R_Squared is given by:

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

The numerator is MSE (average of the squares of the residuals) and the denominator is the variance in Y values. Higher the MSE, smaller the R_squared and poorer is the model.

4. Adjusted R^2

Just like R^2, adjusted R^2 also shows how well terms fit a curve or line but adjusts for the number of terms in a model. It is given by below formula:

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

where n is the total number of observations and k is the number of predictors. Adjusted R^2 will always be less than or equal to R^2

Underfitting:

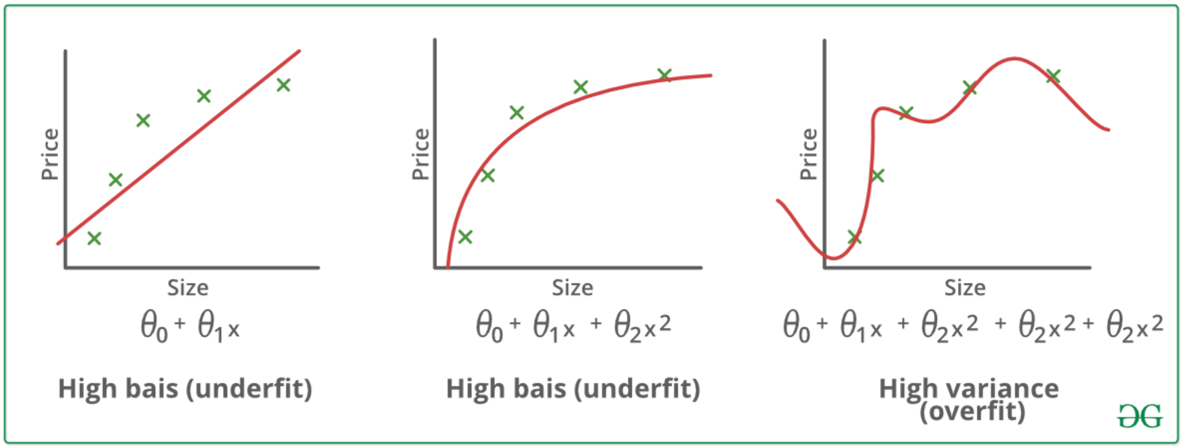
A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. *(It's just like trying to fit undersized pants!)* Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.

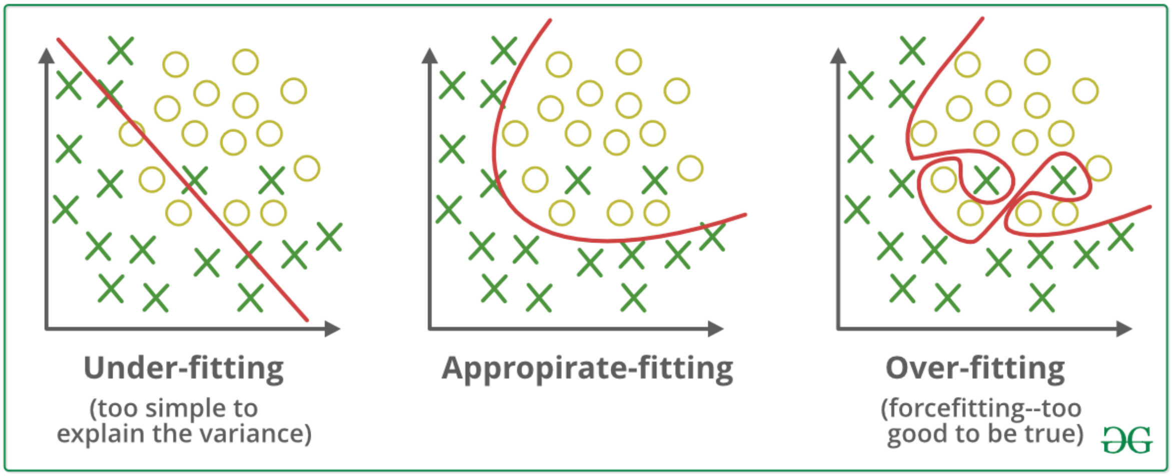
It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data. In such cases the rules of the machine learning model are too easy and flexible to be applied on such a minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

Overfitting:

A statistical model is said to be overfitted, when we train it with a lot of data. When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too much of details and noise.

The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.




Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

How to avoid Overfitting:

The commonly used methodologies are:

- **Cross- Validation:** A standard way to find out-of-sample prediction error is to use 5-fold cross validation.
- **Early Stopping:** Its rules provide us the guidance as to how many iterations can be run before learner begins to over-fit.
- **Pruning:** Pruning is extensively used while building related models. It simply removes the nodes which add little predictive power for the problem in hand.
- **Regularization:** It introduces a cost term for bringing in more features with the objective function. Hence it tries to push the coefficients for many variables to zero and hence reduce cost term.