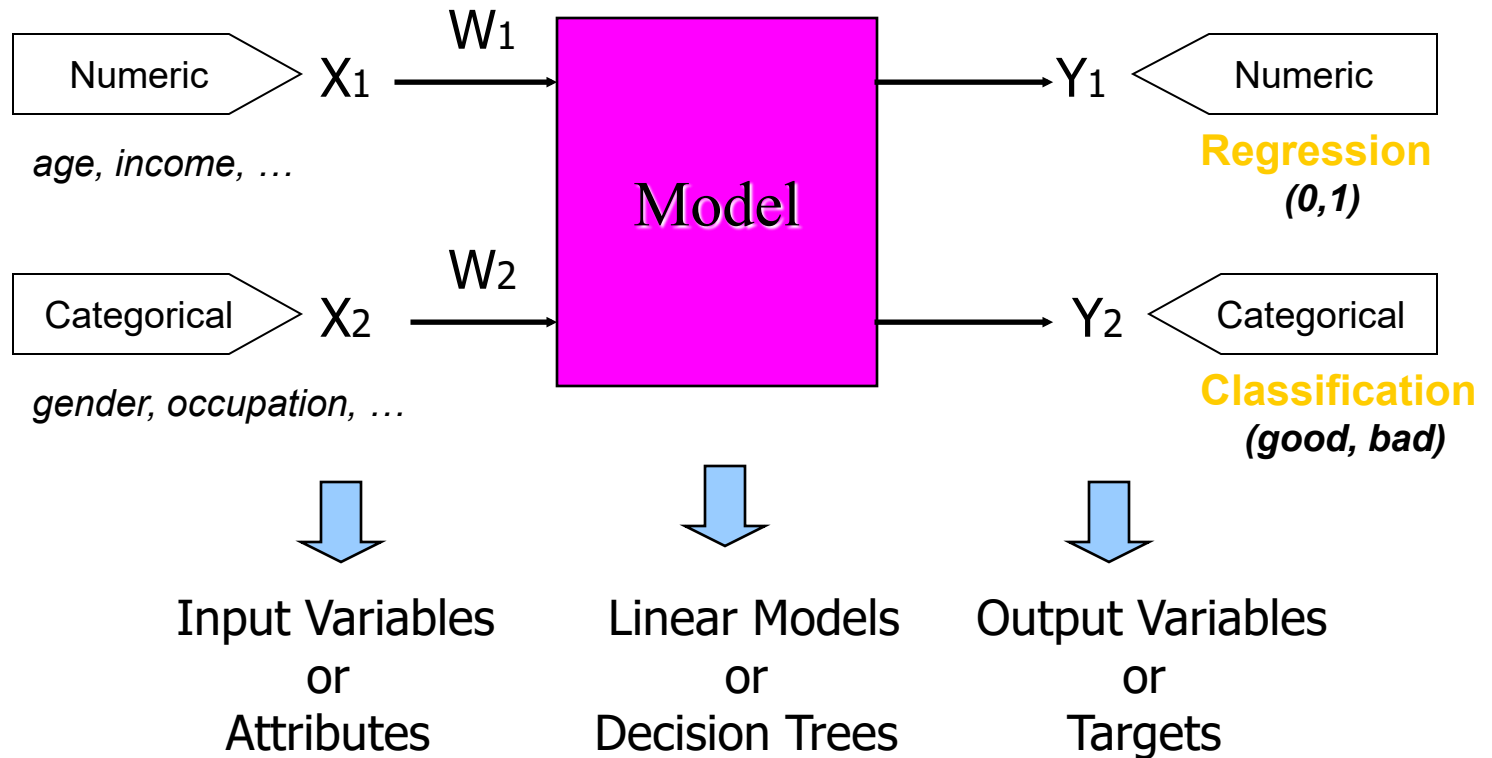

Supervised and Unsupervised Learning Application

Data Analysis



Supervised vs. unsupervised Learning

- **Supervised learning:** classification is seen as supervised learning from examples.
 - **Supervision:** The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a “teacher” gives the classes (**supervision**).
 - Test data are classified into these classes too.
- **Unsupervised learning (clustering)**
 - **Class labels of the data are unknown**
 - Given a set of data, the task is to establish the existence of classes or clusters in the data

Methodology

□ Prediction Methods

- Use some variables to predict unknown or future values of other variables.

□ Description Methods

- Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Methodology

- Classification [Predictive]
- Clustering [Descriptive]
- Regression [Predictive]

Classification: Definition

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification: Application 1

□ Direct Marketing

- Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
- Approach:
 - ◆ Use the data for a similar product introduced before.
 - ◆ We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
 - ◆ Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - ◆ Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

□ Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
 - ◆ Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - ◆ Label past transactions as fraud or fair transactions. This forms the class attribute.
 - ◆ Learn a model for the class of the transactions.
 - ◆ Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - ◆ Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - ◆ Label the customers as loyal or disloyal.
 - ◆ Find a model for loyalty.

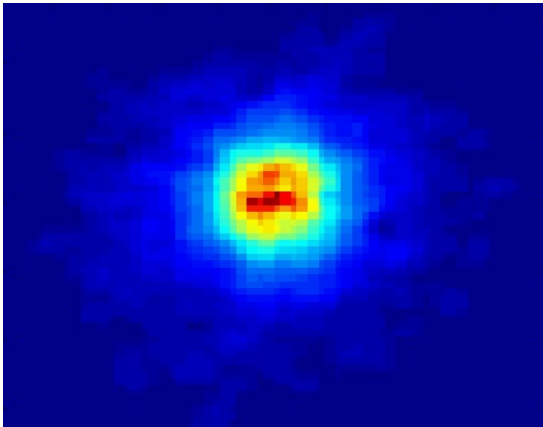
Classification: Application 4

- Sky Survey Cataloging
 - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
 - 3000 images with 23,040 x 23,040 pixels per image.
 - Approach:
 - ◆ Segment the image.
 - ◆ Measure image attributes (features) - 40 of them per object.
 - ◆ Model the class based on these features.
 - ◆ Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

Courtesy: <http://aps.umn.edu>

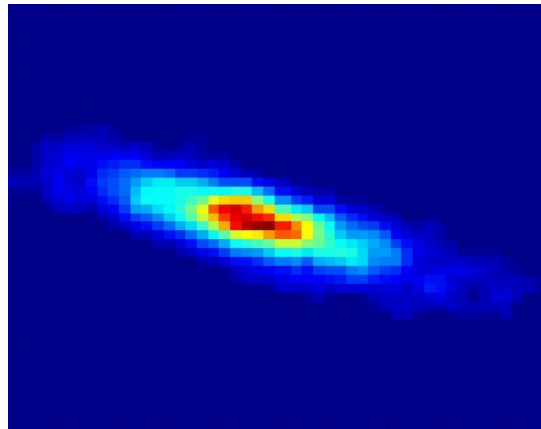
Early



Class:

- Stages of Formation

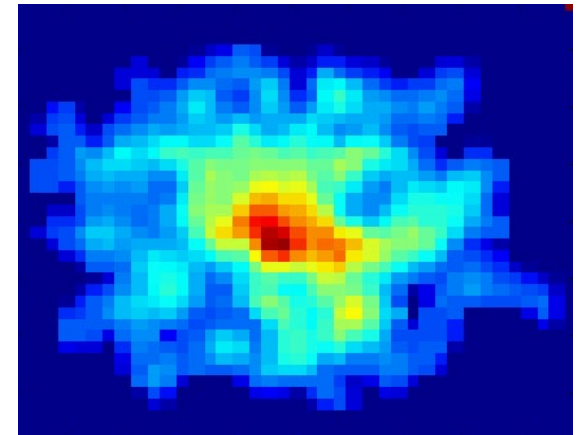
Intermediate



Attributes:

- Image features,
- Characteristics of light waves received, etc.

Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Clustering Definition

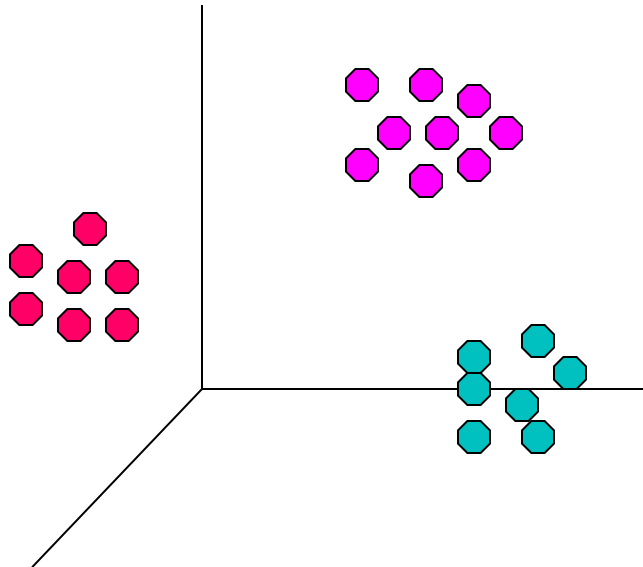
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
 - Data points in one cluster are more similar to one another.
 - Data points in separate clusters are less similar to one another.
- Similarity Measures:
 - Euclidean Distance if attributes are continuous.
 - Other Problem-specific Measures.

Illustrating Clustering

□ Euclidean Distance Based Clustering in 3-D space.

Intracuster distances
are minimized

Intercluster distances
are maximized



Clustering: Application 1

- Market Segmentation:
 - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - ◆ Collect different attributes of customers based on their geographical and lifestyle related information.
 - ◆ Find clusters of similar customers.
 - ◆ Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
 - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
 - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
 - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- ❑ Clustering Points: 3204 Articles of Los Angeles Times.
- ❑ Similarity Measure: How many words are common in these documents (after some word filtering).

<i>Category</i>	<i>Total Articles</i>	<i>Correctly Placed</i>
<i>Financial</i>	555	364
<i>Foreign</i>	341	260
<i>National</i>	273	36
<i>Metro</i>	943	746
<i>Sports</i>	738	573
<i>Entertainment</i>	354	278

Regression

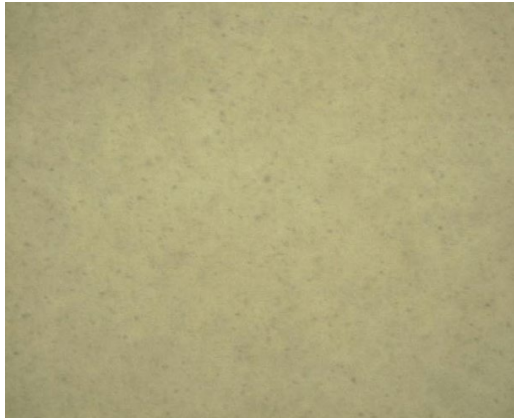
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Example

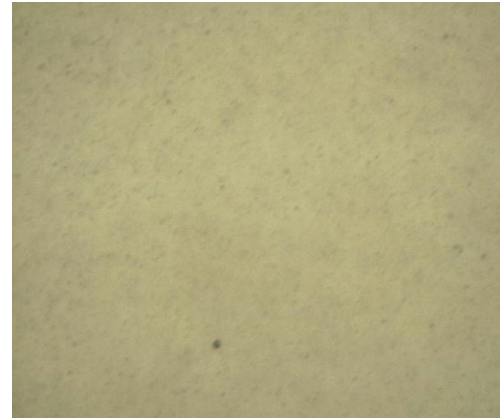
1. Define the problem
2. Build data in database
3. Explore data
4. Prepare data for modeling
5. Build model
6. Evaluate model
7. Deploy model

1. Define the problem

Classify images into those with particles (WP) and those without particles (WO).



WO



WP

2. Build database

- 2000 Images
- 54 Input variables all numeric
- One output variables with two possible values
(With Particle and Without Particle)

4. Prepare data for modeling

- Pre-processed images to remove noise
- Dataset 1 with sharp images: 1350 images including 1257 without particles and 91 with particles
- Dataset 2 with sharp and blurry images: 2000 images including 1909 without particles and blurry particles and 91 with particles
- 54 Input variables, all numeric
- One output variable, with two possible values (WP and WO)

5. Build a model

Classification:

- OneR
- Decision Tree
- 3-Nearest Neighbors
- Naïve Bayesian

6. Evaluate Models

10 -fold cross-validation

Dataset	Attrib.	Class	One-R	C4.5	3.N.N	Bayes
<i>Sharp Images</i>	54	2	99.9	99.8	99.8	95.8
<i>Sharp + Blurry Images</i>	54	2	98.5	97.8	97.8	93.3
<i>Sharp + Blurry Images</i>	54	3	87	87	84	79

If pixel_density_max < 142 then WP

7. Deploy model

- ❖ A Visual Basic program will be developed to implement the model.

