Data, Features, Statistics, Exploratory Analysis, Classification

by Ahmet Sacan

Data

- Data=?
- Measurement, Observation
  - Numeric, words, image
- Incomplete, Error
- measurements are just a set of samples: formed as a result of random selection of some representatives of the set

Terminology

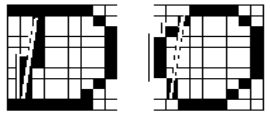
- Sample
  - Input
  - Vector
  - Input vector
- Attribute
  - Feature
  - Variable
  - Property

id	clump thick	uniformity of	uniformity	marginal ad	single epith	bare nuclei	bare chrom	normal nuc	mitoses	class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2

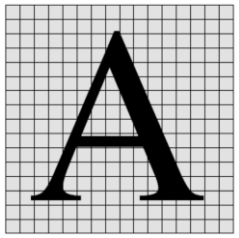
Data from Wisconsin Breast Cancer Dataset: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

Features

- PatternRecognition is mostly based on Feature Spaces
- Feature extraction
  - Transform raw data
- What makes good features?
  - Pixel intensities?
  - Type/number of curves
  - Art/science?



Data Representation



- Raw Data: 16x16 binary/gray image
- Features: moments, strokes, endpoints, holes, number of horizontal/vertical/slanted lines, etc.
  - Features represent the object
  - Summarize/capture relevant aspects
  - In most cases, cannot recreate original data from features

Data Representation

- Most methods work with vectors
- Other data types
  - Time Series
  - Graphs
  - Text: Bag of words
  - Sequences
  - 3D structures

Feature Space

- Map objects to vector space
- Categorical: symbolic or discrete
  - nominal (unordered): sweet, salty, sour
  - ordinal (can be ordered): colors, small < medium < large
- Continuous: numerical

Feature vector

- Usually numeric and continuous, but doesn't have to be
- Binary values: {0,1} or {-1,+1}
- Scalar types: integer, real
- Complex numbers: x+iy
- Vector valued variables: color values in RGB or other multi-channel data

Statistics

- Population
  - people in a country
  - crystal grains in a rock
  - goods manufactured by a particular factory during a given period
- Time series: Data collected about population at various times
- Study subset of population, called a "sample"
  - For practical reasons
- Collect data, analyze
  - Descriptive
  - Inferential

Statistics

- Statistics is a mathematical science pertaining to the
  - collection,
  - analysis,
  - interpretation or explanation, and
  - presentation of data.
- Descriptive statistics
  - summarize or describe a collection of data.
- Inferential statistics
  - Patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, to draw inferences about the process or population being studied.

Descriptive Statistics

- summarize the data, either numerically or graphically, to describe the sample
- E.g., mean and standard deviation.

Inferential Statistics

- Inferential statistics is used to model patterns in the data, accounting for randomness and drawing inferences about the larger population.
- Hypothesis testing: answers to yes/no questions
- Estimation: estimates of numerical characteristics
- Forecasting of future observations,
- Correlation: descriptions of association
- Regression: modeling of relationships
- Other modeling techniques: ANOVA, time series, and data mining.

Statistics

- Univariate
  - Mean
  - Variance
- Multivariate
  - Summaries of each variable separately
  - Relationship between (pairs of) variables
    - Variance
    - Covariance
    - Correlation

Data Analysis

- Exploratory Analysis
  - Exploration: attempts to recognize any non-random pattern or structure
  - Mining: generates possible interesting hypotheses for further study
- Confirmatory Analysis
  - After well-defined hypothesis in mind
  - Some type of (well-known) significance test