Programming Machine Learning Applications

Lecture Two: Understanding Characteristics of Data

Dr. Aleksandar Velkoski

Course Overview

Machine Learning

Fundamental Concepts

Getting to Know your Data

Approach to Preparation

Review of Lecture One

Types of Datasets

Instances & Features

Interactive Workshop

Lecture Two

About Datasets

Types of Datasets

Record

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vector
- Transaction data

Graph and network

- World Wide Web
- Social or information networks
- Molecular Structures

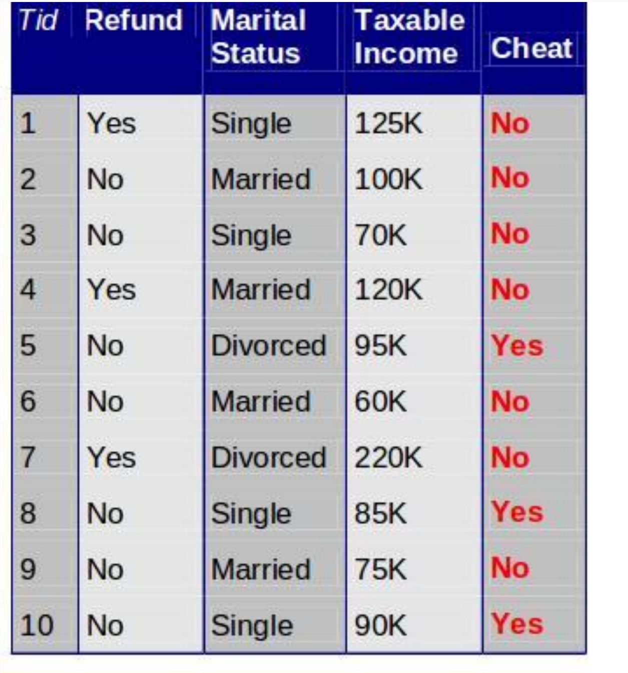
Ordered

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

Spatial and Multimedia

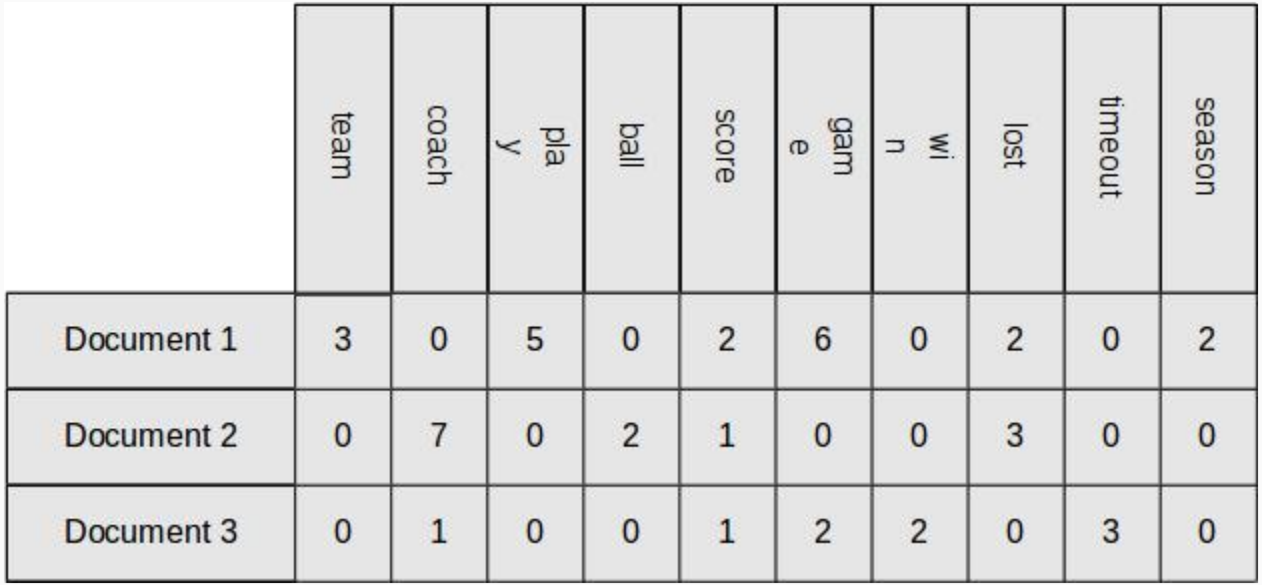
- Spatial data: maps
- Image data
- Video data

Tabular Data



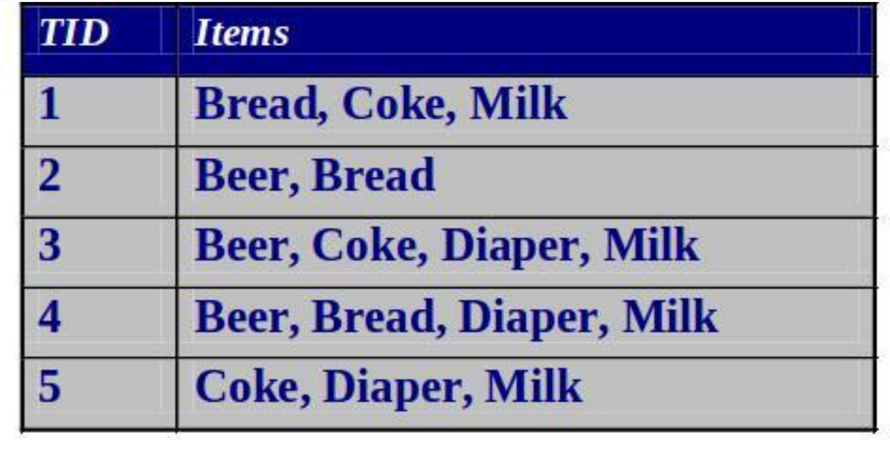
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Document Data



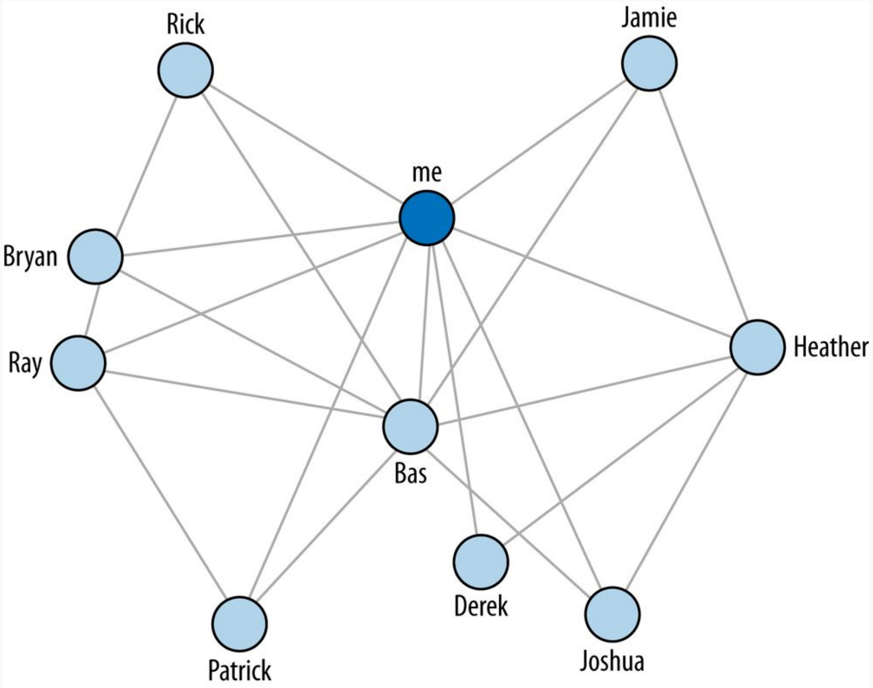
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

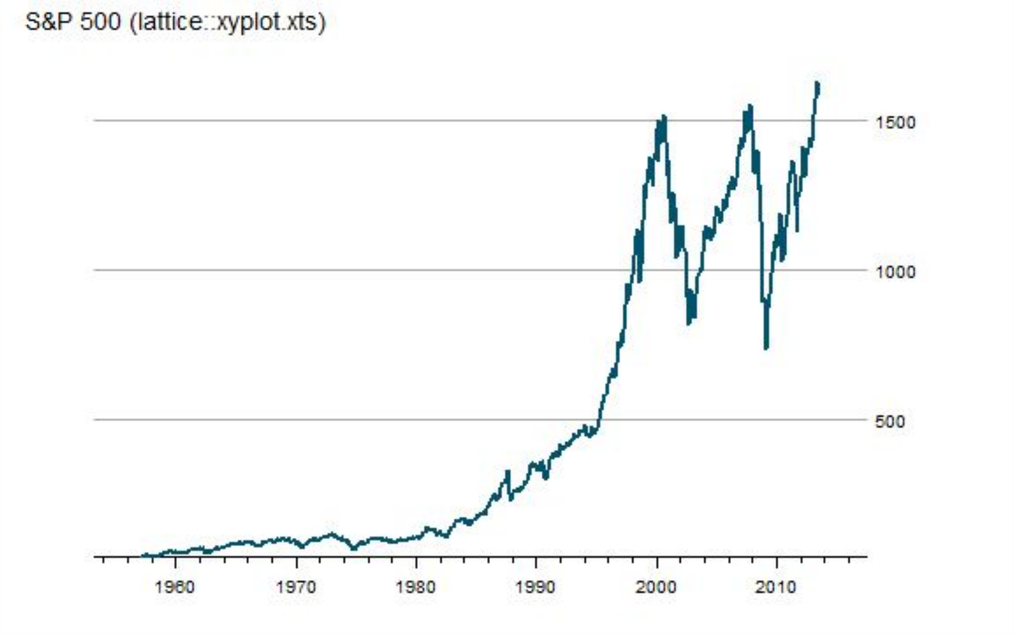


TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data



Ordered Data



Instances & Features

Data Instances and Features

Data sets are made up of data instances.

A data instance represents a subject:

- sales database: customers, products
- medical database: patients, treatments
- university database: students, professors, courses

Also called objects (book usage), rows, samples, examples, data points, tuples.

Instances are described by data features: customer _ID, name, address, age

Data Instances and Features

Nominal / Categorical: categories, states, or “names”

Binary: Symmetric and Asymmetric

Ordinal: Values have a meaningful order (ranking) but magnitude unknown

Numeric: Interval-scaled, Ratio Scaled Quantity

Data Instances and Features

Discrete Features

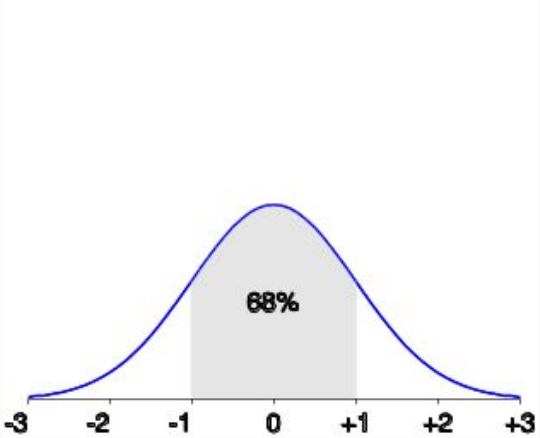
- Finite or countably infinite set of values

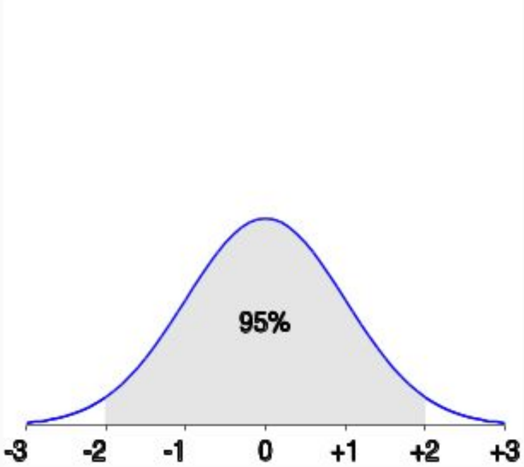
Continuous

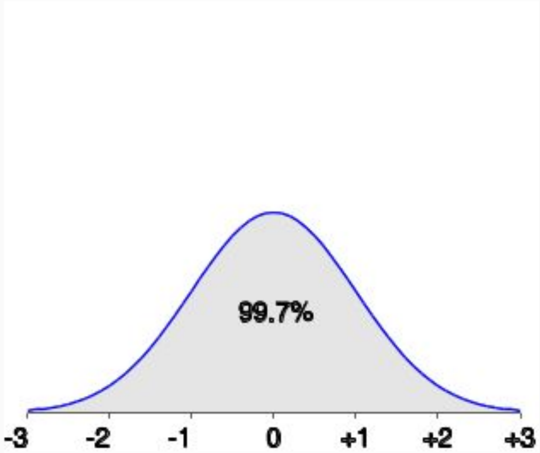
- Real numbers as feature values
- Usually floating points

Properties of Distributions

- From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements (μ: mean, σ: standard deviation)
- From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
- From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it

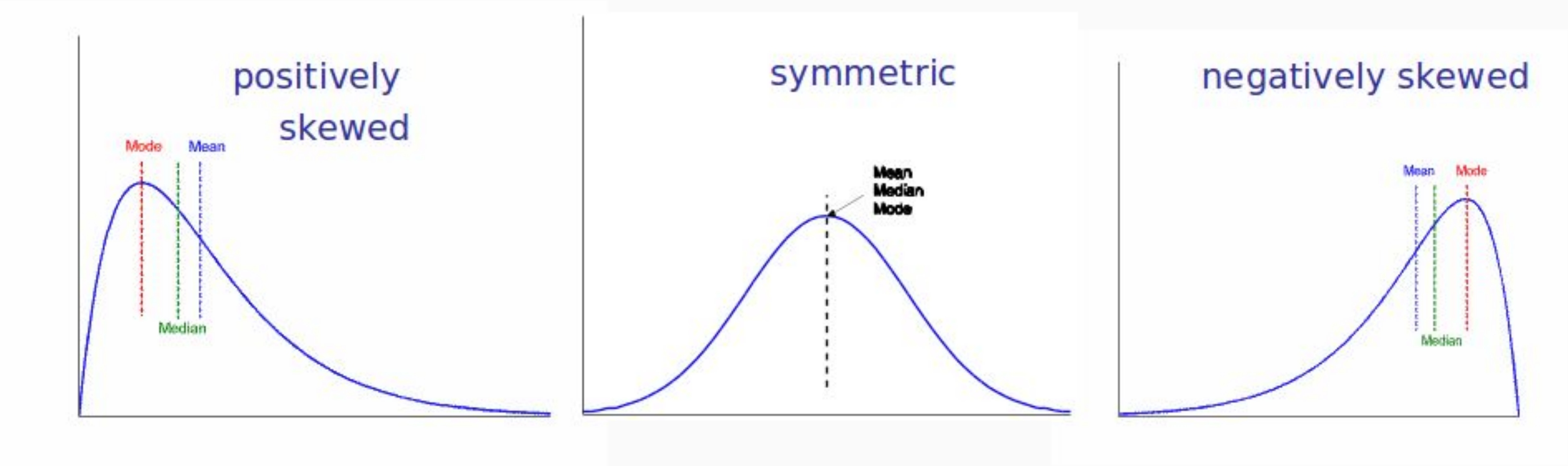






Symmetric vs. Skewed

Median, mean and mode of symmetric, positively and negatively skewed data



Graphic Display of Statistical Descriptions

Boxplot: graphic display of five-number summary

Histogram: x-axis → values; y-axis → frequencies

Quantile plot: plots univariate distribution. each xi is paired with fi indicating that ~ fi *100% of data are xi

Quantile-quantile (q-q) plot: graphs quantiles of one univariate distribution against corresponding quantiles of another.

Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

NumPy & Pandas Tutorial

Wrapping-up the Lecture

Questions

What method can you use to get basic statistical descriptions of features in a DataFrame?

How do you create a correlation matrix in Pandas?

What ways can you use to select data in Pandas?

How do you convert a DataFrame to a NumPy Array?