Rensselaer

why not change the world?®

Introduction to Analytic Methods, Types of Data Mining for Analytics

Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/CSCI-4600 BCBP- 4600/ MGM-4600/MGMT-6600

Module 4, January 27th, 2026

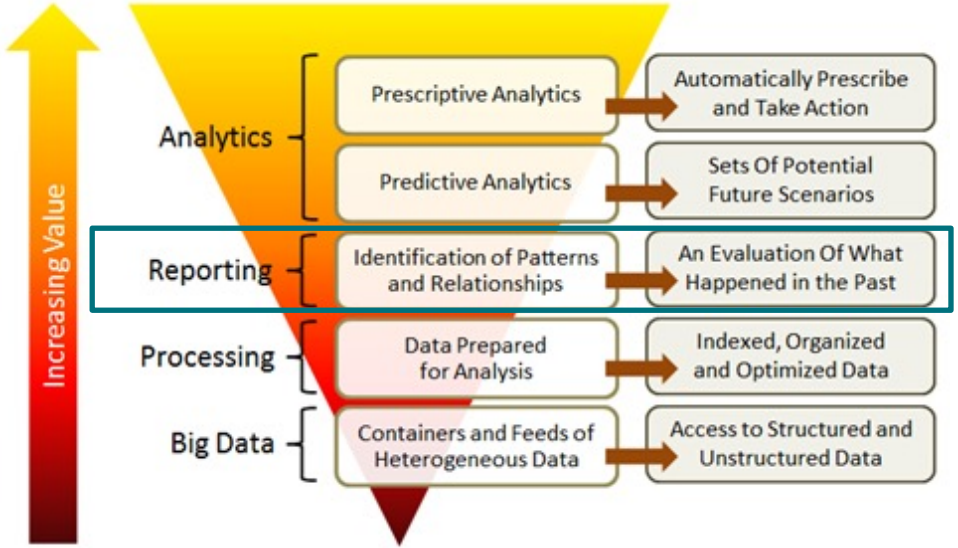
Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Exploratory Data Analysis
  - Variable Distributions
  - Variable Relationships
- Visualizations
  - Boxplots, histograms, scatterplots
- Modeling and Predictive Analysis
  - Patterns & Relationships
  - Objective function
  - Model development & evaluation
- Regression
- Data Mining
  - History
  - Supervised/unsupervised learning
  - kNN, K-Means
  - Rule Mining
  - Model accuracy vs. interpretability
- Prescriptive Analytics

Contents

- Reminder: preliminary/exploratory data analysis, models
- Patterns/ Relations via “Data mining”
- Interpreting results
- Saving the models
- Proceeding with applying the models



Preliminary Data Analysis

- Also called **Exploratory Data Analysis**

– “EDA is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe will be there” (John Tukey)

- Distribution analysis and comparison, visual ‘analysis’, model testing, i.e. pretty much the things you did last lab..

Exploratory Data Analysis

- Initial investigation of data
- Discover patterns and anomalies
- Test hypotheses
- Check assumptions
- Foundation for further analysis

Understanding Your Data

- Central tendency (mean, median, mode)
- Spread (variance, standard deviation)
- Outliers and anomalies
- Frequency distributions

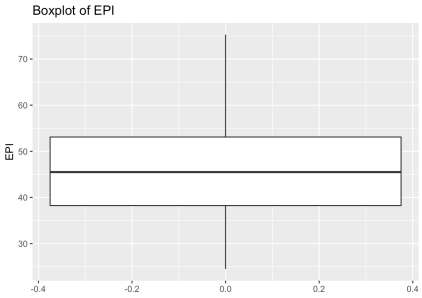
Visualizing Variable Distributions

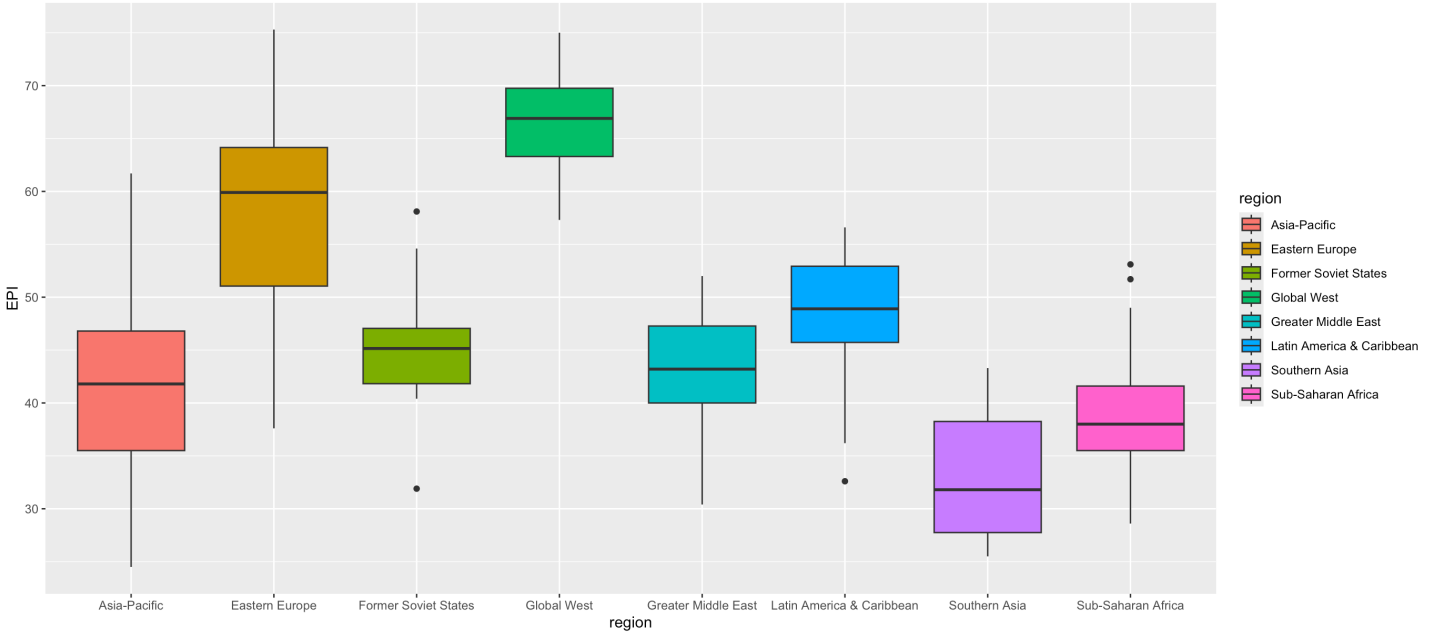
- Boxplot - box-and-whisker of a variable x

- Gives an overview of the range of the variable and where most of the observations are exist along that range
- Make a note of the minimum, maximum, quartiles
- Check for unusual observations (outliers)

Visualizing Variable Distributions

Boxplot - box-and-whisker of a variable x





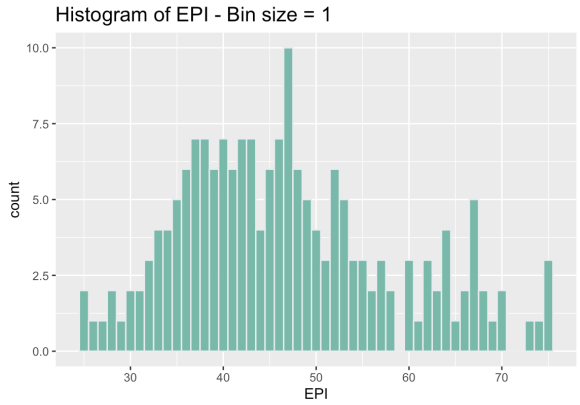
Visualizing Variable Distributions

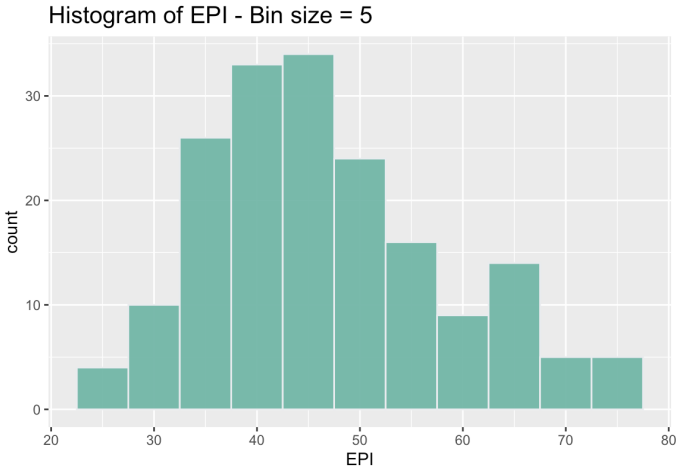
- Histogram – Distribution of variable X

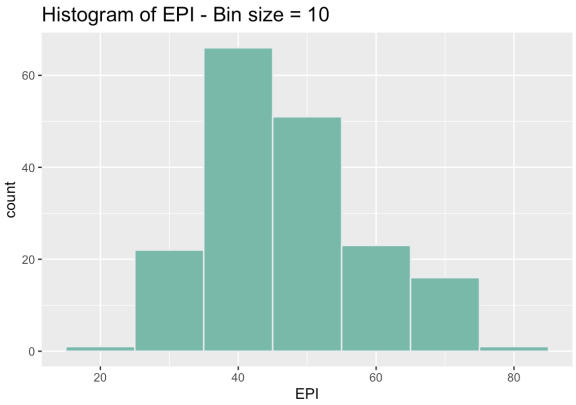
- Describe frequency of occurrence of values or ranges of values (bins) of x
- Tune bin size parameter
- Observe overall shape
- Check for mixed distributions

Visualizing Variable Distributions

Histogram – Distribution of variable X







Relationships between Variables

- Variable Associations
- Covariance
- Correlation analysis

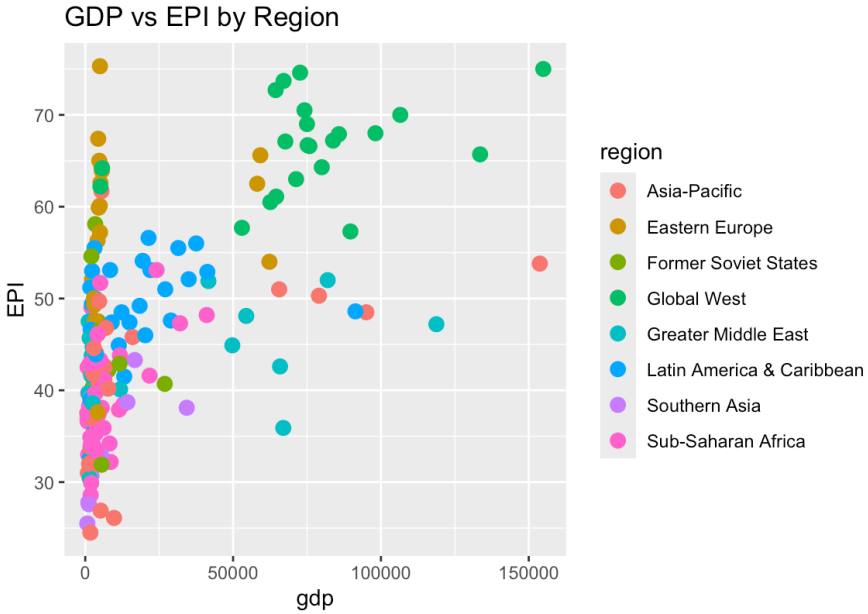
Visualizing Relationships between Variables

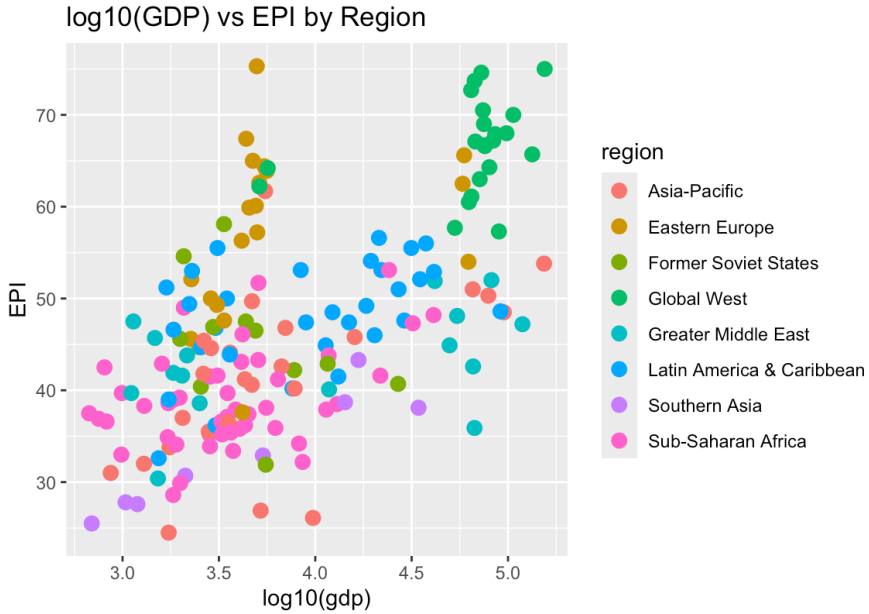
- Scatter Plot – Paired data (x,y)

- Describe the relationship between numerical variables.
- Make a note on the direction of the data points
  - Positive direction
  - Negative Direction
- Check for unusual observations
- See the relationship - Linear or Non-linear

Visualizing Relationships between Variables

Scatter Plot – Paired data (x,y)





Predictive Modeling

Predictive Analysis

- Use historical data to estimate future activity
- Identify patterns
- Develop predictive models
- Quantify uncertainty
- Support decision-making

Finding Patterns in Data

- Linear relationships
  - Ads vs. sales
- Non-linear patterns
  - Hours slept vs. happiness
- Temporal trends
  - Stock price trends
- Spatial patterns
  - mineral diversity in region

Defining Success

- Minimize error
- Maximize accuracy
- Optimize performance metrics
- Balance trade-offs
- This is as much an art as a science and guides model selection

Objective function

Constraint function(s)

Objective Function

- Also called loss function or cost function
- Quantifies how well the model performs
- Guides the learning algorithm, stops after threshold reached
- Lower values = better performance (typically)

e.g. minimizing prediction error

- **Predictions**: Model outputs (ŷ)
- **True Values**: Actual labels (y)
- **Error/Loss**: Difference between predictions and truth

Model Development Process

1. Define problem
2. Collect and prepare data
3. Select algorithm
4. Train model
5. Validate results

Models

- Model of a phenomenon / thing of interest *not* model of the data (structural form)

- Assumptions are often used when developing models, e.g. the *sample* being representative of the *population*

- “All models are wrong but some are useful” (*generally attributed to the statistician ~ George Box*)

Assessing Performance

- Training vs. testing error
- Cross-validation
- Confusion matrices
- Performance metrics (accuracy, precision, recall)

*more on these in upcoming lectures..

Regression

Regression in Statistics

• Regression is a statistical process for *estimating* the relationships among continuous numerical variables.

• In regression analysis the focus is on the relationship between a dependent variable and one or more independent variables.

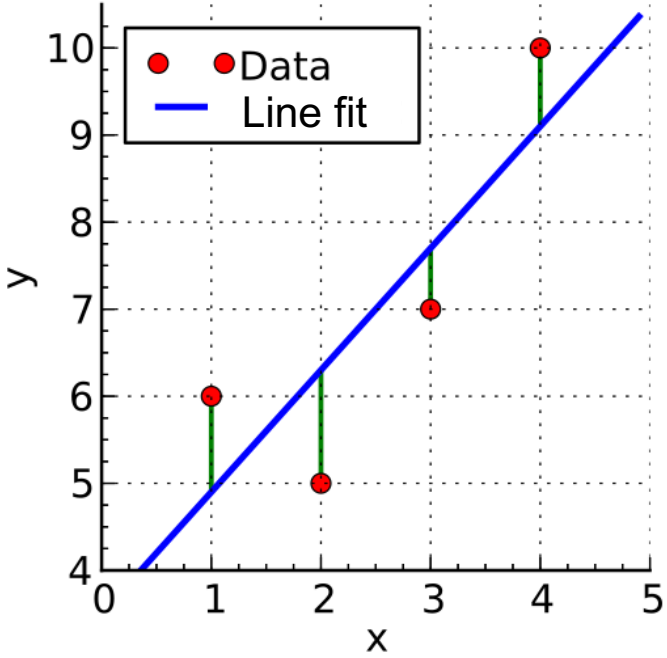
• Independent variables are also called *predictors*, *covariates*.. these are **inputs**.

• Dependent variables are also called *target* or *response* variables, these are **outputs**.

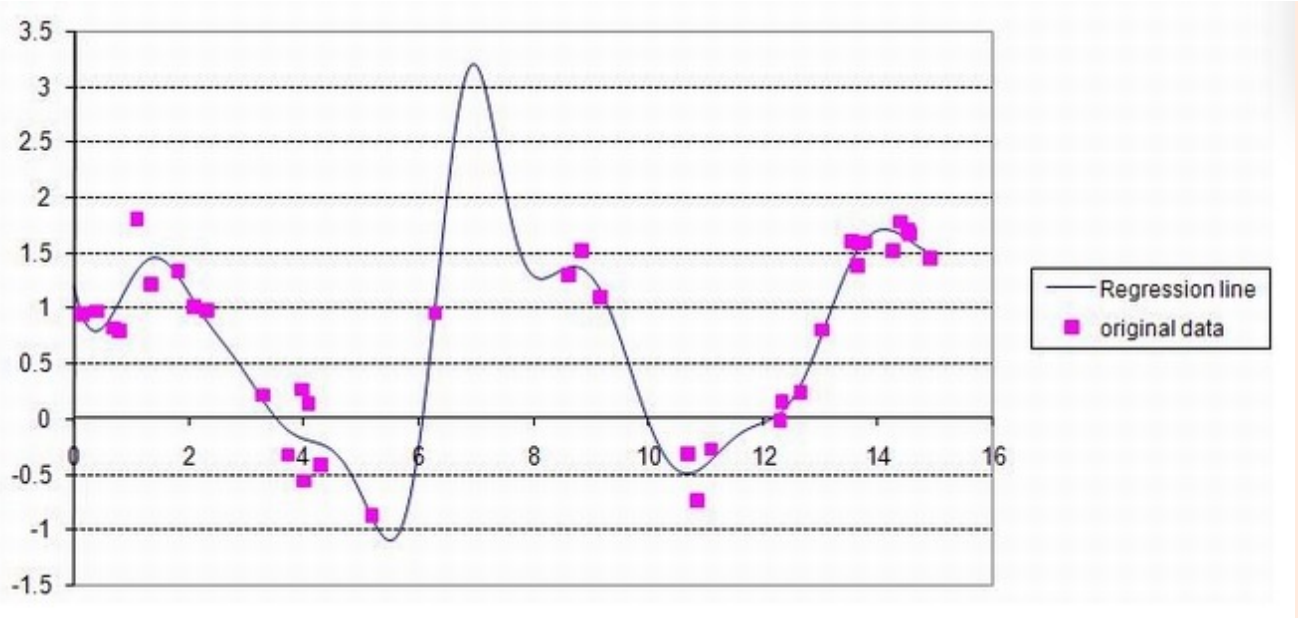
• Estimation is often done by constraining an objective function.

• Must be tested for significance, confidence.

Regression



Regression - when it gets complex…



Predicting Continuous Outcomes

- Model relationships between continuous variables
- Predict a target (dependent) variable based on one or more independent variables
- Understand the significance of independent variables
- Test hypotheses
- Quantify uncertainty

The Foundation

- Simple or multiple linear regression – one / many predictor(s)
- Assumptions: linearity, independence, homoscedasticity
- Finding best fit line
- Interpretation of coefficients (variable significance)
- Evaluating R-squared and model fit

Real-World Uses

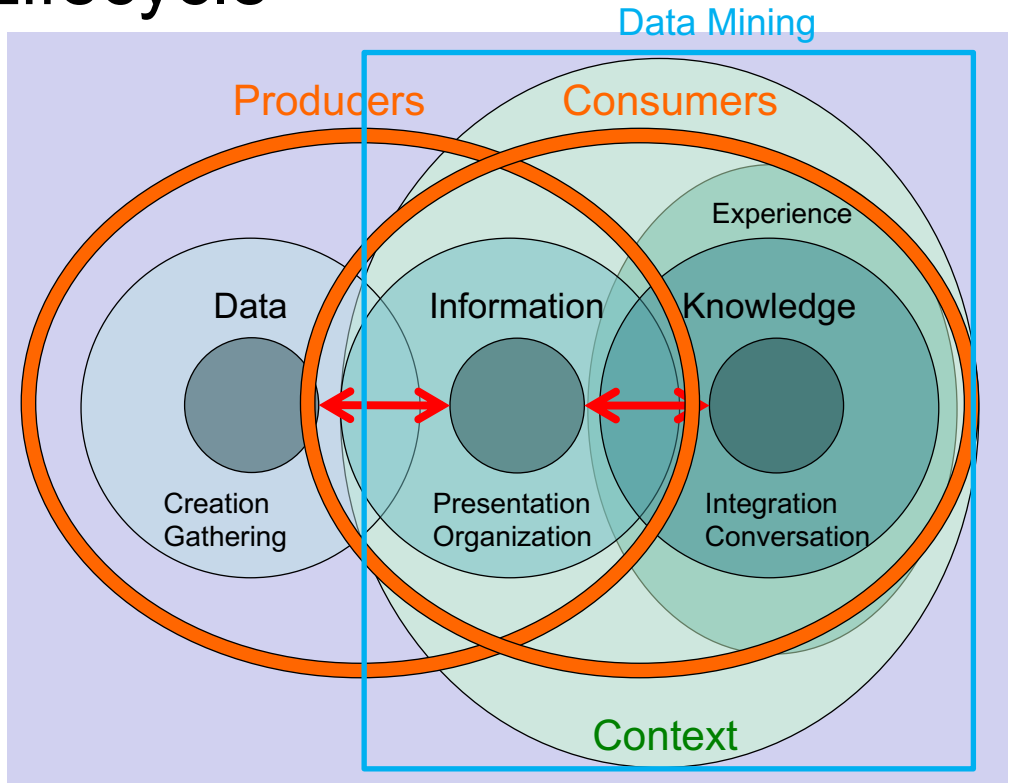
- Sales forecasting
- Price prediction
- Demand estimation
- Performance modeling

Data Mining

Discovering Knowledge in Data

- Extract patterns from large datasets
- Automated or semi-automated
- Combine statistics, machine learning, and databases
- Transform data into actionable insights
- Handle the abundance of data of varying quality

Data Lifecycle



Evolution of the Field

- 1960s: Early pattern recognition
- 1980s: Knowledge discovery in databases
    - Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)
- 1990s: Data mining term emerges
- 2000s: Data Deluge (Big Data, etc.)
- Today: Deep Neural Networks (Deep Learning, LLMs, etc.) - surge in AI
- Next: skynet .. jk

Data Mining = Finding Patterns

Classification (Supervised Learning)
– Classifiers are created using labeled training samples – Training samples created by ground truth / experts
– Classifier later used to classify unknown samples

• Clustering (Unsupervised Learning)
– Grouping objects into clusters so that similar objects are in the same cluster and dissimilar objects are in different clusters
– Discover overall distribution patterns and relationships between attributes

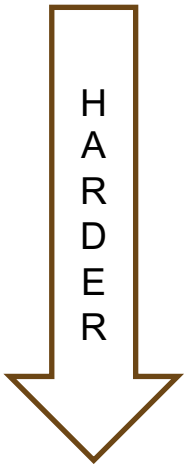
• Association Rule Mining
– Initially developed for market basket analysis
– Goal is to discover relationships between attributes

• Other Types of Mining
– Outlier Analysis
– Concept / Class Description
– Time Series Analysis

Models/ types

- Trade-off between Accuracy and Understandability
- Models range from “easy to understand” to incomprehensible

– Decision trees
– Rule induction
– Multi-variate Regression models
– Neural Networks
– Deep Learning



Supervised Learning

Learning with Labels

- Training data includes true class labels
- Algorithms learn input-output mapping
- Binary classification

e.g. spam detection, medical diagnosis

- Multiclass classification

e.g. object in image (cat, dog, person, etc.)

k-nearest neighbors (kNN)

• Can be used in both regression and classification (“non-parametric”)
– Supervised learning, i.e. model is trained

• kNN is a method for classifying objects based on the closest training examples in the feature space. ~ "Birds of a feather flock together"

• **An object is classified by a majority vote of its neighbors. k is always a positive integer**. The neighbors are taken from a set of objects for which the correct classification is known.

• It is usual to use the Euclidean distance, though other distance measures such as the Manhattan distance could in principle be used instead.

Algorithm

- The algorithm on how to compute the k-nearest neighbors is as follows:

– Determine parameter *k* = number of nearest neighbors, beforehand. This value **is up to you**.

– Calculate the distance between the query-instance and all the training samples. You can use **any distance** algorithm.

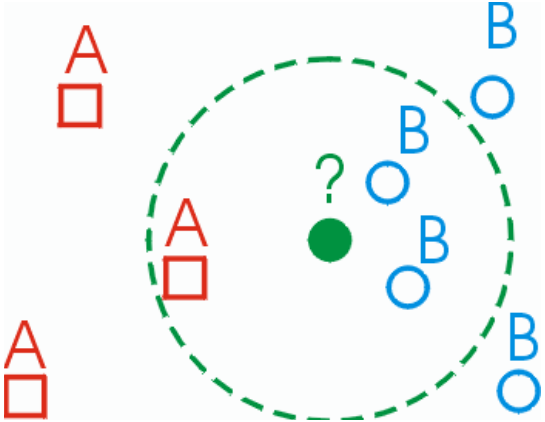
– Sort the distances for all the training samples and determine the nearest neighbors based on the *k* shortest distances.

– Since this is supervised learning, get the classes for the *k* nearest neighbors from the training set.

– Use the majority of nearest neighbors as the prediction value.

Choice of k?

- Don’t you hate it when the instructions read: the choice of ‘k’ is all up to you ??
- Loop over different k, evaluate results...



Distance metrics

• **Euclidean** distance is the most common use of distance. Euclidean distance, or simply 'distance', examines the root of the sum of square differences between the coordinates of a pair of objects. This is most generally known as the Pythagorean theorem.

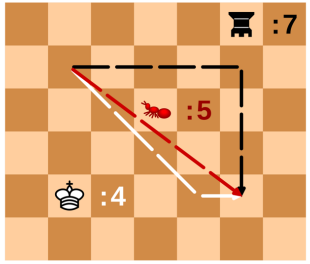
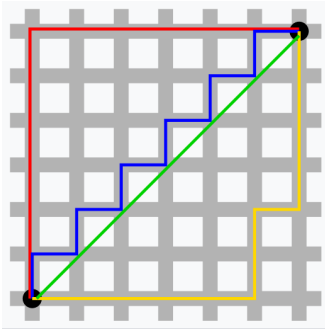


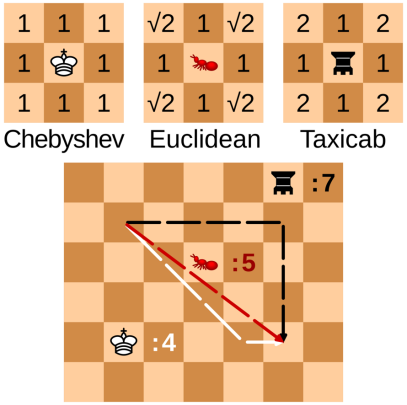
Image credit: Cmglee
license: CC BY-SA 4.0 – no changes

• The **taxicab** metric is also known as **rectilinear** distance, L1 distance or L1 norm, city block distance, **Manhattan** distance, or Manhattan length, with the corresponding variations in the name of the geometry. It represents the distance between points in a city road grid. It examines the absolute differences between the coordinates of a pair of objects.



https://en.wikipedia.org/wiki/Taxicab_geometry

More generally



• **Chebyshev** distance is also called the Maximum value distance, defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. In other words, it examines the absolute magnitude of the differences between the coordinates of a pair of objects.
https://en.wikipedia.org/wiki/Chebyshev_distance

• The general metric for distance is the **Minkowski** distance. When *p* is equal to 1, it becomes the city block distance, and when *p* is equal to 2, it becomes the Euclidean distance. The special case is when *p* is equal to infinity (taking a limit), where it is considered as the Chebyshev distance.

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Pros

- Simple to understand and implement
- No “training” phase
- Robust to noisy training data

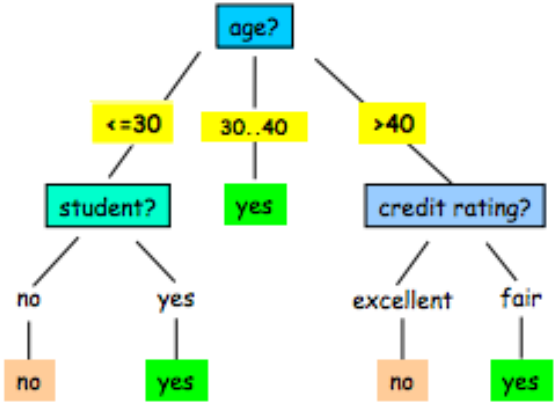
Cons

- Need to determine value of parameter *k*
- Computationally expensive with large datasets
- Sensitive to feature scaling

Decision tree classifier

Classification by Decision Tree Induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



buys_computer ?

Key Characteristics of Decision Trees

- Flowchart-like structure for decisions
- Root node → Internal nodes → Leaf nodes
    - Internal node: test on attribute
    - Branch: outcome of test
    - Leaf: class label or value

Pros

- Handle both classification and regression
- Interpretable and easy to visualize
- Handles non-linear relationships

Cons

- Prone to overfitting
- Sensitive to small data changes in the data

Unsupervised Learning

Discovering Hidden Structure

- Labeled data are not required
- Algorithms find naturally occurring structure in the data
- Used to find clusters of similar observations in feature space
- **Clustering** and **Dimensionality Reduction**

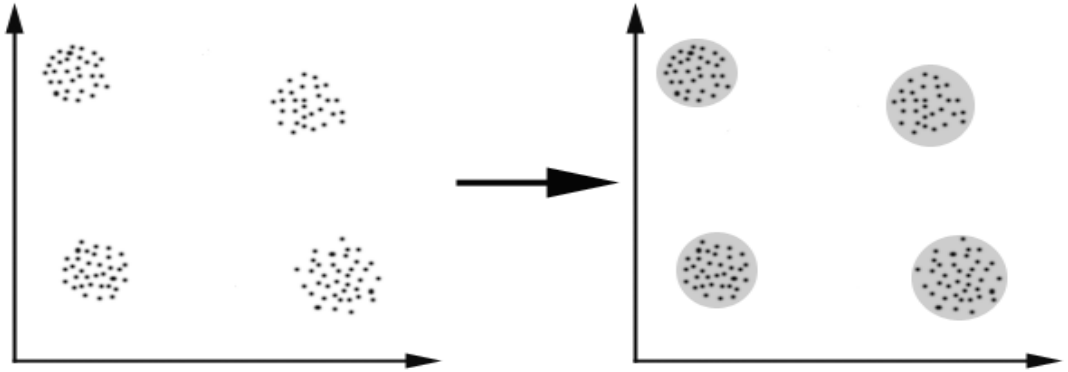
e.g. image segmentation, anomaly detection,

Clustering

- Partition dataset into clusters by optimizing some distance / similarity measure.
- Observations in a cluster are closer to each other in feature space than they are to other clusters.

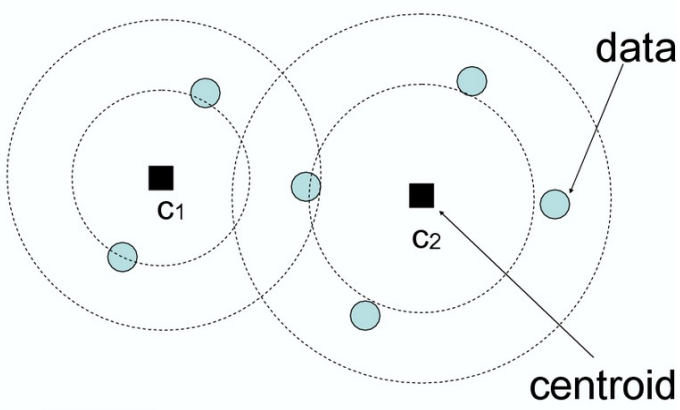
Similarity/Distance Measure

- Clustering is about finding “similarity”.
- To find how similar two objects are, one needs a “distance” measure.
- Similar objects (same cluster) should be close to one another (short distance).



k-Means Clustering

- Separate the objects (datapoints) into *k* clusters.
- Cluster center (centroid) = the mean (average) of all the data points in the cluster.
- Assigns each data point to the cluster whose centroid is nearest (using distance function).



K-Means Algorithm

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

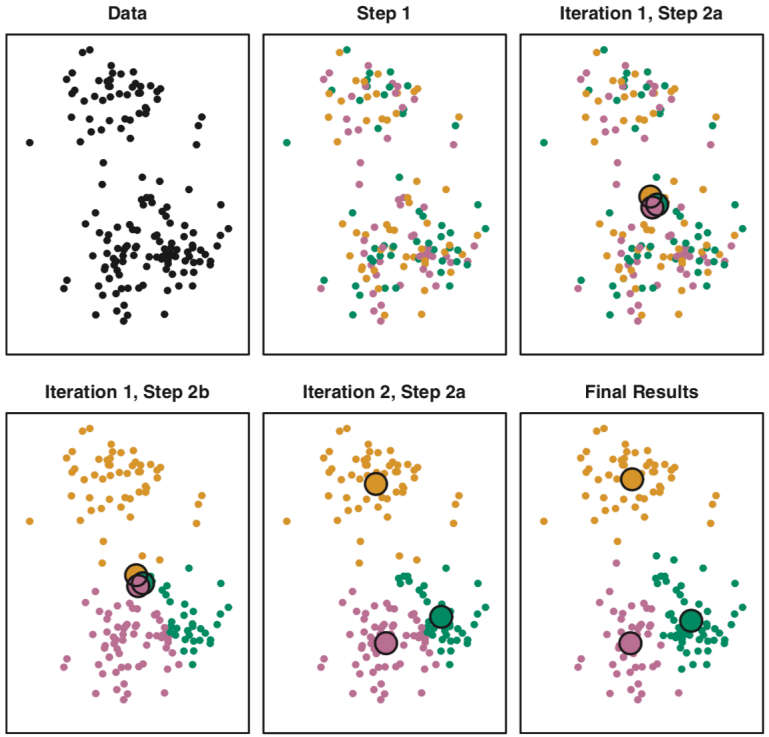
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans

K-Means Algorithm

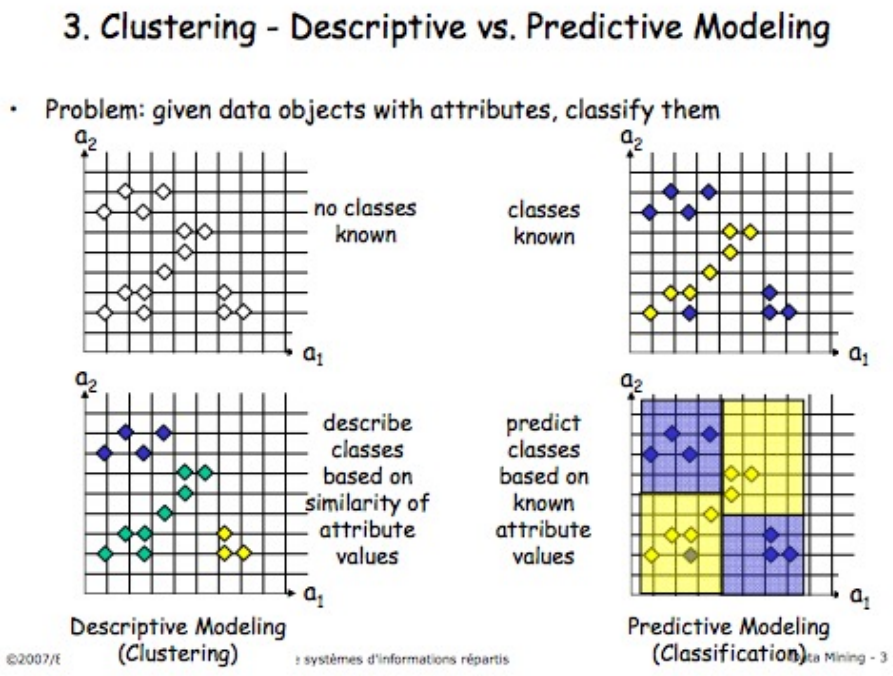
Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans

Describe v. Predict



Dimensionality Reduction

- Dimensionality Reduction is an unsupervised technique used to transform high-dimensional data into lower-dimensional space.

- Why?
  - Reduce number of features/variables in dataset
  - Preserve important information while removing noise
  - Make data easier to visualize and analyze
  - Reduce computational costs

Dimensionality Reduction Methods

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- t-SNE (t-Distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection)

Association Rules

- Find if-then patterns
- Association rule learning / mining
- Market basket analysis
- Support, confidence, lift metrics are used to evaluate rules
- Applications include recommender systems

Accuracy vs. Interpretability - Balancing Act

- Complex models: higher accuracy, lower interpretability
- Simple models: lower accuracy, higher interpretability
- Choose based on research / application needs as well as data availability
- Consider computational / project / institutional constraints
- Document model decisions

Prescriptive Analytics - Beyond Prediction

- Descriptive: What happened?
- Predictive: What will happen?
- **Prescriptive: What should we do?**
- Recommend optimal actions
- Incorporate constraints and objectives into decision making

Dataset search

• If you do not have a dataset in mind for your project, please search online and select datasets using search tools such as
https://datasetsearch.research.google.com/

• If you need help choosing a dataset, please come and talk to me during the class time or during virtual office hours, so that I can guide/help you to select datasets.

• **NOTE:6000-Level students MUST have TWO datasets (minimum two datasets) used during final project.**

More places to find data:

- US Government Data: https://www.data.gov/
- US Department of Agriculture: https://www.nass.usda.gov/Data_and_Statistics/index.php
- Center of Disease Control (CDC): https://www.cdc.gov/datastatistics/index.html
- US Financial Data: https://www.federalreserve.gov/data.htm
- European Union Open Data Portal: https://data.europa.eu/euodp/en/data/
- Nasa: https://data.nasa.gov/
- United Nations: https://data.un.org/

Next Class: Friday January 29th

Lab 2

Thanks!