# Applied Data Mining

Rockhurst University

Intro, Naive Bayes, and KNN

Section 1

# Intro

# Why data mining?

- The Explosive Growth of Data: from terabytes to petabytes
- Data collection and data availability
- Automated data collection tools, database systems, Web, computerized society
- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, ?
  - Science: Remote sensing, bioinformatics, scientific simulation, ?
  - Society and everyone: news, digital cameras, YouTube, Twitter, etc
- **We are drowning in data, but starving for knowledge!**
- "Necessity is the mother of invention" — Data mining — Automated analysis of massive data sets

# Supervised learning

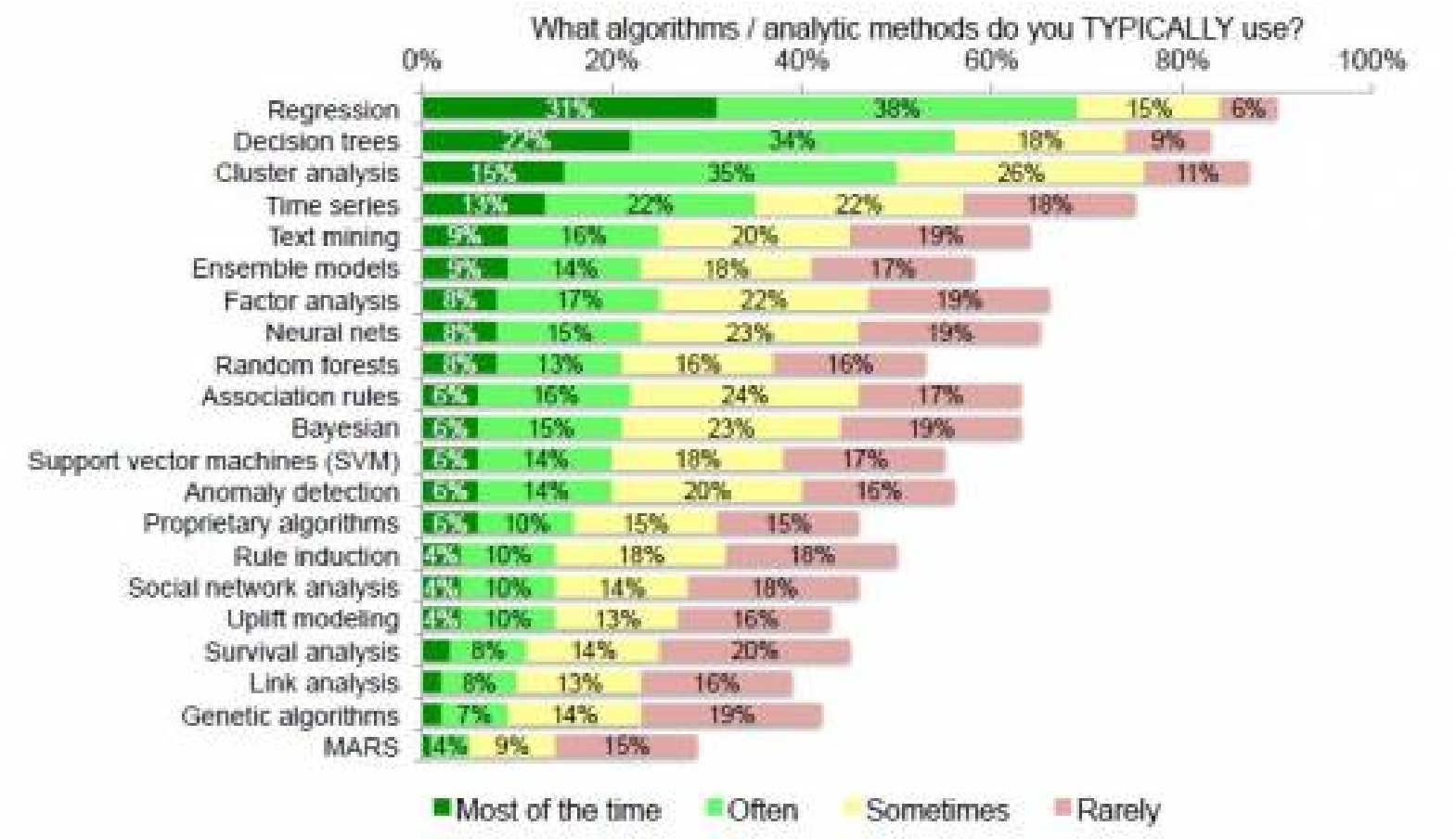
- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
  - Predict some unknown class labels
- Typical methods
  - Linear regression, logistic regression (BIA 6309)
  - K nearest neighbors, nave Bayesian classification (tonight)
  - Decision trees, rule-based and pattern-based classification
  - Support vector machines, neural networks, ... (BIA 6303)
- Typical applications: Credit card fraud detection, direct marketing, classifying stars, diseases, web-page

# Unsupervised learning

- Class label or target variable is unknown
- Group data to form new categories or find frequent patterns (or frequent itemsets)
- Principle: Maximizing intra-class similarity and minimizing interclass similarity and looking for correlation, finding items that are highly correlated
- Typical methods: K-means, hierarchical clustering, association rules, correlation analysis
- Typical applications: Marketing, image segmentation, recommender systems

# Lots of tools

1




[1] Source: Source: http://gerardnico.com/wiki/data_mining/algorithm

# Remember: regression analysis

Remember: linear regression tries to fit a line to the observations to minimize the sum of squared residuals.

Let's look at a simple example: ▸ Interpreting Model Output in R

# Remember: regression analysis

Remember: logistic regression is also a linear model but one where the dependent variable is a class variable, not a continuous variable.

Let's look at a simple example: ▸ The Grad School example

# Classification Vs. Numeric Prediction

- Numeric Prediction:
  - models continuous-valued functions.
  - Example: linear regression
- Classification :
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
  - Examples: Logistic regression, KNN, Naive Bayes, Decision trees, etc.

# Classification - A two step process

Model construction: describing a set of predetermined classes

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
- The set of tuples used for model construction is training set
- Choose the classification algorithm that is going to best fit your training set.

Model usage: for classifying future or unknown objects

- Run classification algorithm on validation (test) set
- Estimate accuracy of the model
- The known label of test sample is compared with the classified result from the model
- Accuracy rate is the percentage of test set samples that are correctly classified by the model

Section 2

# Naïve Bayes

# Probability refresher

- **Prior probability**: P(A)-probability of event A occurring
- **Joint probability**: P(A ∩ B) or P(A,B)-probability of event A and B occurring
- **Conditional probability**: P(A | B)-probability of event A occurring given that event B has occurred. Not necessary the same as P(B | A)
- **Independence**: If A and B are independent, then P(A | B) = P(A) and P(B | A) = P(B)
- Altogether, we have:
  **P(A,B) = P(B | A) P(A) = P(A | B) P(B)**

# Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A|B)$: probability of instance B being in class A. This is what we are trying to compute.
- $P(B|A)$ : probability of generating instance B given class A. Being in class A "causes" you to have feature B with some probability.
- $P(A)$: probability of occurence of class A - how frequent is class A in our data set.
- $P(B)$: probability of occurence of class B - how frequent is class B in our data set.

# Customer Named "Drew"[2]

We need to call a potential customer named "Drew Smith". Do we ask for Mr. Smith or Ms. Smith?

$$P(Male|Drew) = \frac{P(Drew|Male)P(Male)}{P(Drew)}$$

$$P(Female|Drew) = \frac{P(Drew|Female)P(Female)}{P(Drew)}$$

Note: the $P(Drew)$ in the denominator is actually irrelevant since it is the same for Male or Female. For shorthand this is often left out.

[2] Source: http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf

# Data

We'll use data on our existing customers to see what we can learn.

| Name | Sex | Name | Sex |
|---|---|---|---|
| Drew | Male | Alberto | Male |
| Claudia | Female | Karin | Female |
| Drew | Female | Nina | Female |
| Drew | Female | Sergio | Male |

| | DREW | NOT DREW | Total |
|---|---|---|---|
| MALE | 1 | 2 | 3 |
| FEMALE | 2 | 3 | 5 |
| Total | 3 | 5 | 8 |

# Question

Is "Drew Smith" more likely to be a male or a female?

- The probability that they are male knowing that they are named Drew: $P(Male|Drew) = \frac{P(Drew|Male)P(Male)}{P(Drew)} = \frac{\frac{1}{3} * \frac{3}{8}}{\frac{3}{8}} = \frac{1}{3}$
  (or $\frac{1}{8} = 0.125$ if we ignore the denominator)
- The probability that they are female knowing that they are named Drew: $P(Female|Drew) = \frac{P(Drew|Female)P(Female)}{P(Drew)} = \frac{\frac{2}{3} * \frac{5}{8}}{\frac{3}{8}} = \frac{2}{3}$
  (or $\frac{2}{8} = 0.250$ if we ignore the denominator)

"Drew Smith" is more likely to be female!

# New data on Drew

So far we have only considered one aspect or "feature" of the data (Name) to determine the sex. What if we have more information? How do we use all the features?

| Name | Over 5' 7" | Eyes | Hair | Sex |
|---|---|---|---|---|
| Drew | No | Blue | Short | Male |
| Claudia | Yes | Brown | Long | Female |
| Drew | No | Blue | Long | Female |
| Drew | No | Blue | Long | Female |
| Alberto | Yes | Brown | Short | Male |
| Karin | No | Blue | Long | Female |
| Nina | Yes | Brown | Short | Female |
| Sergio | Yes | Blue | Long | Male |

# Assume independence

To simplify this process, Naïve Bayes classifiers assumes that attributes have independent distributions so therefore:

$$P(Name|Sex) = P(Height|Sex) * P(Eyes|Sex) * P(Hair|Sex)$$

What if we know that Drew is over 5' 7", has blue eyes and long hair? What sex would we guess then?

Our independence assumption means that we can calculate

$$P(Drew|Male) = P(over5'7'|Male) * P(Blue|Male) * P(Long|Male)$$

$$P(Drew|Female) = P(over5'7"|Female) * P(Blue|Female) * P(Long|Female)$$

The higher probability wins.

# Frequency tables

In addition to our Drew counts, we need the probabilities for the other attributes.

| | **Over 5' 7" = Yes** | **Over 5' 7" = No** | **Total** |
|---|---|---|---|
| MALE | 2 | 1 | 3 |
| FEMALE | 2 | 3 | 5 |

| | **Blue Eyes = Yes** | **Blue Eyes = No** | **Total** |
|---|---|---|---|
| MALE | 2 | 1 | 3 |
| FEMALE | 3 | 2 | 5 |

| | **Long Hair = Yes** | **Long Hair = No** | **Total** |
|---|---|---|---|
| MALE | 1 | 2 | 3 |
| FEMALE | 4 | 1 | 5 |

# Naïve Bayes Calculations

We find out that potential customer "Drew" is over 5' 7", has blue eyes, and long hair:

- $P(Drew|Male) = P(over5'7"|Male) * P(Blue|Male) * P(Long|Male)$

  $P(Drew|Male) = 2/3 * 2/3 * 1/3 = 4/27 = 14.8\%$

- $P(Drew|Female) =$
  $P(over5'7"|Female) * P(Blue|Female) * P(Long|Female)$

  $P(Drew|Female) = 2/5 * 3/5 * 4/5 = 24/125 = 19.2\%$

Guess Drew is female

# Pros and Cons

- Pro: Naïve Bayes isn't sensitive to irrelevant features. For example, eye color isn?t relevant for gender. Your prediction would be the same if you include it or leave it out.
  - $P(Male|Drew)$ — with eye color: 14.8%
  - $P(Male|Drew)$ — without eye color: 22.2%
  - $P(Female|Drew)$ — with eye color: 19.2%
  - $P(Female|Drew)$ — without eye color: 32%
- Con: Assumption of independence can be problematic. Is height **really** independent of gender?

Section 3

# K Nearest Neighbor

# KNN

- kNearest Neighbor works by looking at the k nearest neighbor to the chosen observation and classifies the data point to the closest neighbor.
- “Tell me who your neighbors are, and I’ll tell you who you are”
- Classified as a ”lazy learner” — there is no underlying model. Lazy learners don’t know how they will classify the next observation until they see it.
- Choosing K is an art.
- Hates missing values!

# Measuring distance

In order to know which neighbors are nearest, you need to be able to measure the distance between points. There are several ways to do this.

Some frequently used distance functions.

Camberra :

$$d(x,y) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|} \quad (2)$$

Minkowsky :

$$d(x,y) = \left( \sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r} \quad (3)$$

Chebychev :

$$d(x,y) = \max_{i=1}^{m} |x_i - y_i| \quad (4)$$

Euclidean :

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \quad (5)$$

Manhattan / city - block :

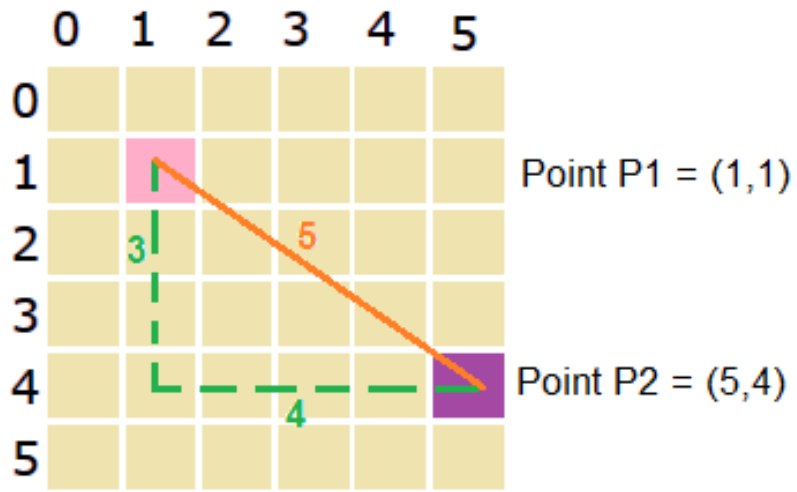
$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i| \quad (6)$$

[3]

[3] Source: https://medium.com/data-science-group-iitr/k-nearest-neighbors-knn-500f0d17c8f1

# Most common distances

The most common measures are Euclidean and Manhattan distances.




Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

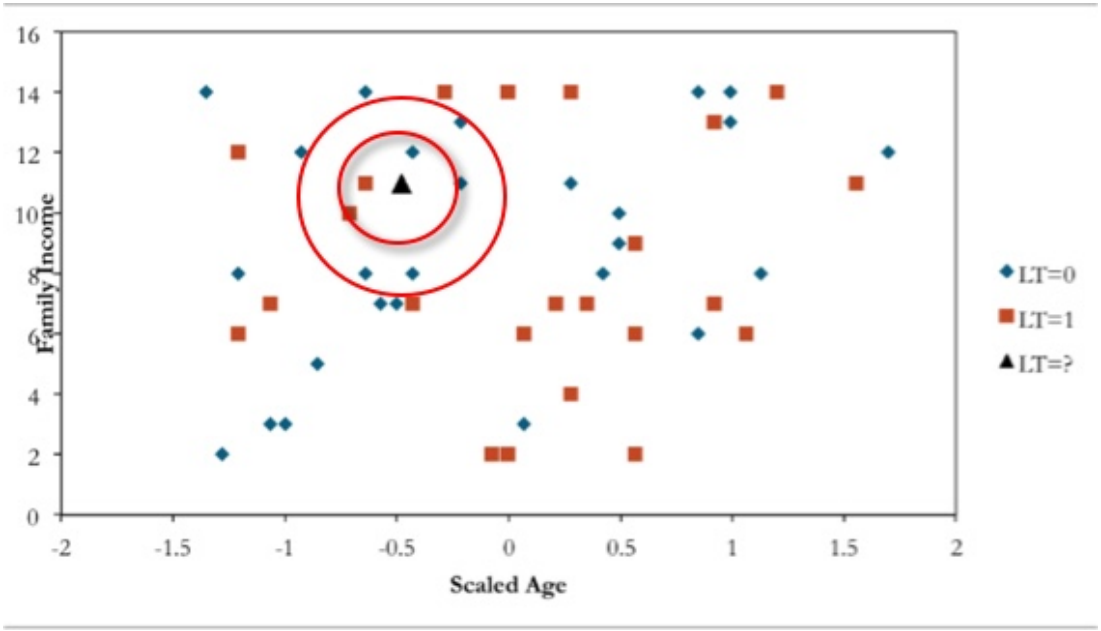
Manhattan distance = $|5-1| + |4-1| = 7$

[4]

[4] Source: https://prismoskills.appspot.com/lessons/2D_and_3D_Puzzles/Chapter_05_-_Distance_between_points.jsp

# Choosing K

How many neighbors should you choose?




▶ More info

# Things to consider

- K needs to be large enough to minimize error rate
- K can't be too large or boundaries will be too smooth
- Distance calculations are affected by scale: data should be standardized. ▸ Example
- Doesn't work with factors: you can't calculate distance with factors.

# Wrap up

Three classification models:

- Logistic regression: linear model provides coefficients for hypothesis testing but log odd interpretation can be confusing.
- K Nearest Neighbor: simple and handles large data but choosing K and distance measure need thought.
- Naïve Bayes: probability basis is appealing but independence assumption can cause trouble.