Data Mining?

Data Mining – *Chen et al, 1996*

...a process of nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

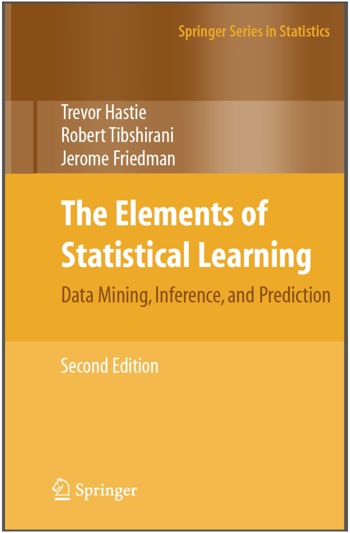
Data Mining – *Cios et al, 2007*

The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain.

- Also called *knowledge discovery in databases*
- How does data mining relate to machine learning, artificial intelligence, statistics, data science, analytics, Big Data?
- No definite agreement on the meaning of this term

What about textbooks on Data Mining?

- There are roughly three styles of textbook on Data Mining: academic, practical, and business
- **Academic**: collects machine learning techniques and statistics (ex. *Hastie* et al is probably overall most popular textbook on data mining)
- **Practical**: explains a specific technology such as Apache Hadoop (ex. any Oreilly book)
- **Business**: Attempts to explain to non-technical audience why *analytics* are important



Learning, Machine Learning, and Data Mining

- What is learning and how do we learn?
  - Learn from experience
- Machine Learning: computer programs that solve problems without being explicitly programmed (*solution is learned from data*)
- Machine Learning is a core requirement of Data Mining (and other fields such as artificial intelligence)
- Data Mining and Data Science also include broader questions related to data

Learning – *Tom Mitchell*

Learning is improving with experience at a task. Improve over task T, with respect to a performance measure P, based on experience E

Learning, Machine Learning, and Data Mining

- What is learning and how do we learn?
  - Learn from experience
- Machine Learning: computer programs that solve problems without being explicitly programmed (*solution is learned from data*)
- Machine Learning is a core requirement of Data Mining (and other fields such as artificial intelligence)
- Data Mining and Data Science also include broader questions related to data

Learning – *Tom Mitchell*

Learning is improving with experience at a task. Improve over task T, with respect to a performance measure P, based on experience E

Learning, Machine Learning, and Data Mining

- What is learning and how do we learn?
  - Learn from experience
- Machine Learning: computer programs that solve problems without being explicitly programmed (*solution is learned from data*)
- Machine Learning is a core requirement of Data Mining (and other fields such as artificial intelligence)
- Data Mining and Data Science also include broader questions related to data

Learning – *Tom Mitchell*

Learning is improving with experience at a task. Improve over task T, with respect to a performance measure P, based on experience E

Learning, Machine Learning, and Data Mining

- What is learning and how do we learn?
  - Learn from experience
- Machine Learning: computer programs that solve problems without being explicitly programmed (*solution is learned from data*)
- Machine Learning is a core requirement of Data Mining (and other fields such as artificial intelligence)
- Data Mining and Data Science also include broader questions related to data

Learning – *Tom Mitchell*

Learning is improving with experience at a task. Improve over task T, with respect to a performance measure P, based on experience E

Machine Learning Example: The Iris Dataset

- *Observational Model*: Set of objects (instances) produce a vector of features
- Machine Learning algorithms attempt to discover the function between the features and the outcome
- In typical notation rows are instances and columns are features

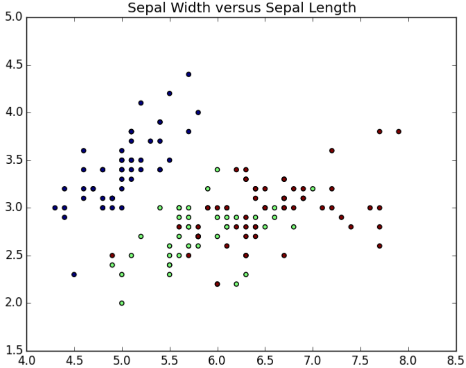
Iris	Sepal Len.	Sepal Wid.	Petal Len.	Petal Wid.	Target
# 1	5.1	3.5	1.4	0.2	0
# 2	4.9	3.0	1.4	0.2	0
⋮	⋮	⋮	⋮	⋮	⋮

$$\text{Target} = f(SL, SW, PL, PW) \tag{1}$$

Properties of Functions and Inductive Learning

- A *function* is defined by its domain and range and also by being single-valued
- A function is equally well defined by its *graph*: the relation defined in the product space **Domain** $\times$ **Range**
  - Form all pairs $(x, f(x))$
  - Domain: independent variables, Range: dependent variables
- Practically, we can only collect a finite number of samples as above
- The *inductive learning hypothesis* says that a model of f that performs well on a training set will also perform well on the entire space

Visualizing Data and Functions, Scatterplots

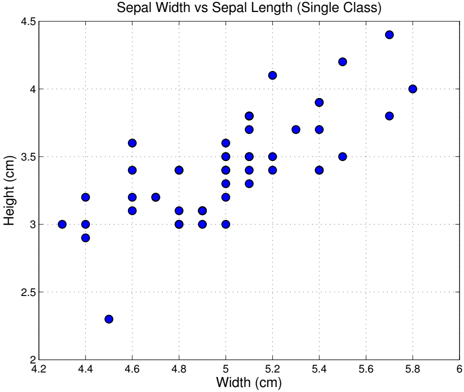


- A *scatter plot* is a figure where two variables are plotted against one another and each datapoint is represented by a single marker
- Effective way to visualize the relationship of two-variables

Categorical vs. Numerical Data

- The *Iris* dataset contains examples of both categorical and numeric data
  - The dependent variable (target) is the three types iris: *setosa, veriscolor, virginica*
  - The independent variables are numeric, they are lengths measured in centimeters
- *Supervised* learning occurs when our data is labelled, each instance is accompanied by some outcome variable
- *Classification*: Outcome is a categorical variable
- *Regression*: Outcome is a numeric variable
- Our model typically includes some assumptions about the form of f : Features $\rightarrow$ Outcome

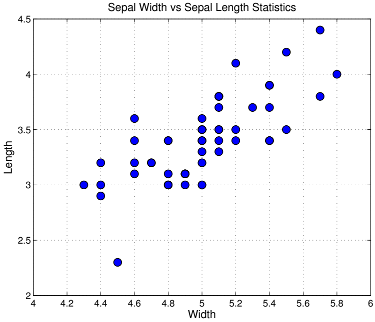
Subset of the Iris Dataset



- Examine just one type of iris and two features
- Allows us to look at some statistics and curve fitting

Characterizing the Data with Statistics

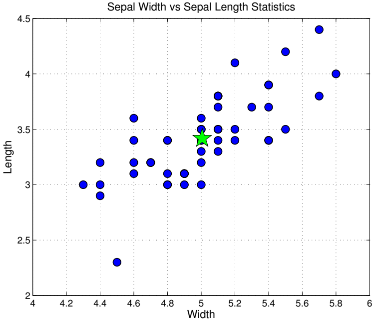
- Characterize the data in terms of *central tendency* and *dispersion* (x_i, y_i)
- Sample mean in both dimensions $\mu_x = \frac{1}{N}\sum_i^N x_i$, ...
- Standard deviation in the x dimension, $(\mu_x - \sigma_x, \mu_x + \sigma_x)$
- ... $(\mu_y - \sigma_y, \mu_y + \sigma_y)$



$$\sigma_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \mu_x)^2 \tag{2}$$

Characterizing the Data with Statistics

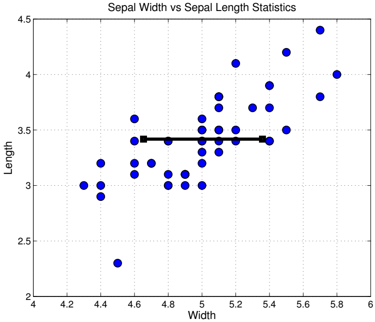
- Characterize the data in terms of *central tendency* and *dispersion* (x_i, y_i)
- Sample mean in both dimensions $\mu_x = \frac{1}{N}\sum_i^N x_i$, ...
- Standard deviation in the x dimension, $(\mu_x - \sigma_x, \mu_x + \sigma_x)$
- ... $(\mu_y - \sigma_y, \mu_y + \sigma_y)$



$$\sigma_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \mu_x)^2 \tag{2}$$

Characterizing the Data with Statistics

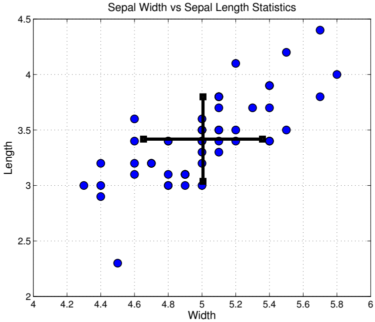
- Characterize the data in terms of *central tendency* and *dispersion* (x_i, y_i)
- Sample mean in both dimensions $\mu_x = \frac{1}{N}\sum_i^N x_i$, ...
- Standard deviation in the x dimension, $(\mu_x - \sigma_x, \mu_x + \sigma_x)$
- ... $(\mu_y - \sigma_y, \mu_y + \sigma_y)$



$$\sigma_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \mu_x)^2 \tag{2}$$

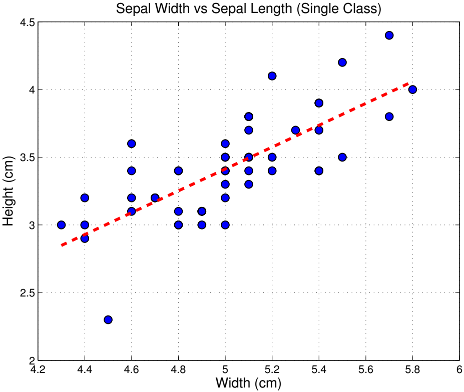
Characterizing the Data with Statistics

- Characterize the data in terms of *central tendency* and *dispersion* (x_i, y_i)
- Sample mean in both dimensions $\mu_x = \frac{1}{N}\sum_i^N x_i$, ...
- Standard deviation in the x dimension, $(\mu_x - \sigma_x, \mu_x + \sigma_x)$
- ... $(\mu_y - \sigma_y, \mu_y + \sigma_y)$



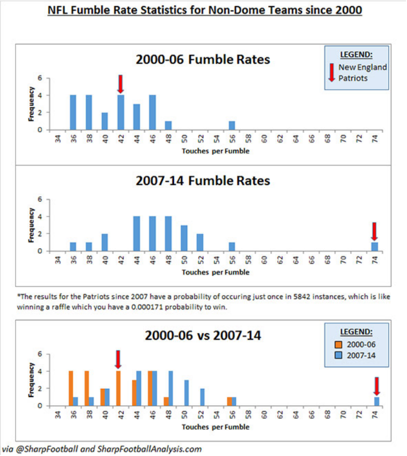
$$\sigma_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \mu_x)^2 \quad (2)$$

Linear Regression



- Clearly, there is a trend in the data
- The width and lengths change together
- Using the *method of least squares* we can find a line that fits the data

DeflateGate Scandal



- Real world (sorta) application of data-mining

Data-Mining: Motivation

- Humans generate increasingly huge amounts of data
- Sources of data: social media, server logs, point-of-sale terminals, medical records, etc.
- Potentially useful resource
- Caveat: large amount of raw-data is of little value without some automated techniques to extract information from it
- Distinction between *data* and *knowledge/information*

Data-Mining: Definition

- Extracting previously unknown and useful information from a corpus of data
- Accomplished by creating computer programs that can discover patterns and regularities in the data
- Problems: patterns may be uninteresting or spurious (artifact of the particular dataset), missing or corrupt values

Relationship to other fields

- Data mining is closely related to a number of other fields including...
- Statistics
- Machine Learning
- Detection and Estimation Theory
- Signals and Systems
- *Difficult to draw a precise distinction between these areas*

Distinguishing Features of Data-Mining

- High volume data
- Primarily *unsupervised* machine learning problems
- Concerned with how our solution will scale and genuinely seeking to discover new knowledge (hence *mining*)

Machine Learning

"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" Tom Mitchell

- Example: Classification, assignment of correct labels to previously unseen data

Classification and Standard Notation

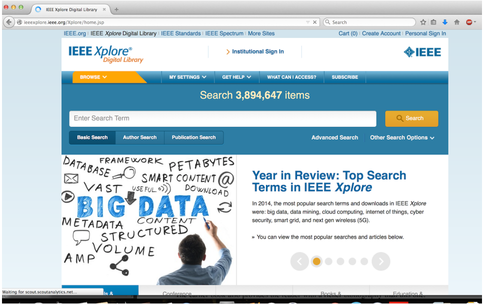
- Classification is ultimately about discovering some function $f(x)$ that maps observations to classes
- What is the precise nature of the argument function and how is the data organized?
- Data instances: record of a database, row of a matrix
- Features: fields in a record, columns of a matrix
- Matrix view: row-dimension is the number of observations and the column dimension is the number of features

Supervised versus Unsupervised Learning Problems

- A dataset may or may not include labels that tell us what categories or classes our observations belong to
- *Supervised Learning*: typical problems are classification and regression
- *Unsupervised Learning*: typical problems are clustering or learning association rules

Why do we care about this now?

- *Big Data* is a term that is increasingly becoming a buzzword
- *Data Science* is claimed to be one of the fastest growing and in-demand professions
- Why is this?



Answer

- Algorithms
- Infrastructure
- Data

What is the *process* for approaching a data mining or data science problem?

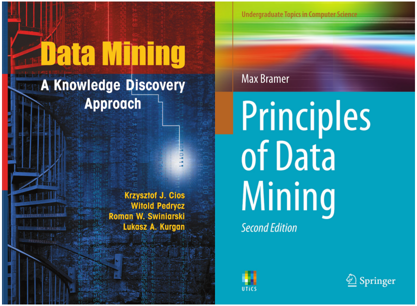
- One process due to H. Mason and C. Wiggins is called OSEMN (pronounced "awesome") which is an mnemonic for *Obtain, Scrub, Explore, Model, iNterpret* (see Janssens, Jeroen. "Data Science at the Command Line." (2014))
- Organizes the activities of data science into a roughly serial process
- Scrub refers to *Data Cleaning* where data is prepared for automated techniques (taking care of NaNs, accounting for missing values, standardizing some records, selecting initial features)

Additional Topics to Cover Today

- Overview of how the course will run
- Begin to review some basic concepts of probability and talk about *Bayes Theorem*

Textbooks for the Course

- Intend to teach out of two textbooks primarily
- "Data Mining: A Knowledge Discovery Approach" Cios *et al*
- "Principles of Data Mining" Bramer
- Both of these books are available as a pdf download from link.springer.com from within NU



High-level Overview of Topical Coverage

- Review of some essential background material: primarily linear algebra and probability theory
- Data Science Process
- Section on machine learning covering supervised and unsupervised learning algorithms: least squares, partial least squares, nearest neighbor, support vector machines, k-means classification, linear and logistic regression, linear discriminant analysis
- Data mining applications in a practical setting: large dataset management with cloud computing tools such as Apache Hadoop
- Data Visualizations

Technology

- **Programming Languages**
- Python with some standard libraries for numerical computations, plotting, etc: numpy, scipy, matplotlib
- Javascript with d3 and jquery for visualization
- **Other Tools**
- Apache Spark
- Amazon EC2 (??)
- `curl` and REST Apis for data collection

Other Resources

- Machine Learning Repository at UCI
  http://archive.ics.uci.edu/ml/, currently warehouses 307 datasets in a variety of domains
- IEEE Transactions: *Pattern Analysis and Machine Intelligence*, *Knowledge and Data Engineering*, *Signal Processing*

Assessment

- Assessment will be based on:
- mid-term and final exams
- 5-6 Homework Assignments
- 3-4 Quizzes
- 0-2 Presentations
- Class Participation

What is Probability?

- Probability that A is true is denoted $P(A)$ generally
- Two views of probability *classical* view and the *frequentist* view

$$P_{\text{Classical}}(A) = \frac{N_A}{N} \tag{3}$$

$$P_{\text{Frequentist}}(A) = \lim_{n \to \infty} \frac{n_A}{n} \tag{4}$$

Axioms of Probability

- Implicit in probability the notion of a sample space, all possible outcomes
- Integrating P over the sample space results in 1
- $P(\top) = 1$
- $P(\bot) = 0$
- $0 \leq P(A) \leq 1$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- $P(A \vee \neg A) = 1$
- $P(A \wedge \neg A) = 0$

Conditional Probability

- The probability that A is true given that we know B is true
- $P(A|B) = \frac{P(A \wedge B)}{P(B)}$
- $P(B|A) = \frac{P(A \wedge B)}{P(A)}$
- In the context of machine learning we should consider conditional probabilities relating to observations
- Conditional probabilities are sometimes called likelihoods

Bayes Law

- *Bayes Law* gives us a formula for reversing a conditional probability
- Observe that $P(A \wedge B)$ occurred in both conditional formulas from the previous slide

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{5}$$

Rare Disease Example

- Assume that in a total population of 10,000 people 1% are afflicted with a rare condition
- Test is available where 99 % of sick patients test positive for disease and 99 % of healthy patients test negative
- Question: What is the probability that a person is sick if they test positive?
- I.e. what is $P(\text{sick}|\text{tested positive})$

Rare Disease Example cont.

- Solution using Bayes Law

$$P(\text{sick}|\text{tested positive}) = \frac{P(\text{tested positive}|\text{sick})P(\text{sick})}{P(\text{tested positive})} \tag{6}$$

$$P(\text{sick}|\text{tested positive}) = \frac{(99/100)(1/100)}{99/10000 + 99/10000} = 0.5 \tag{7}$$

Naive Bayes Classifier

- These observations can be used to create a (supervised) classification algorithm
- Naive Bayes has been applied extensively in text classification in particular for spam filtering
- In a text classification problem the features might be the presence or absence of certain words in the document
- Notice that this makes the problem entirely *categorical*

Naive Bayes Classifier cont.

- In the Naive Bayes algorithm we have a certain number of classes denoted C_k and observations $\mathbf{x}$ of some categorical variable
- From the training data we can determine both the prior probabilities $P(C_k)$ and the likelihoods
- $P(\mathbf{x}|C_k)$ which factors as a result of the independence (naive) assumption
- Given new data we calculate the posteriori probability for each of the classes $P(C_k|\mathbf{x})$ and assign the observation to the class with the largest posterior, this is called the MAP rule

Follow-up

- Please attempt to get a working python environment
- Download the textbooks for the course read the introductory chapters
- Follow-up chapter 3.2 in "Understanding Data Mining" to see a worked example of Naive Bayes Classification