ECE 657A: Lecture 1

Data and Knowledge Modelling and Analysis

Haitham Amar

January 7, 2019

Today's Class

Part I - Course Information

Course Admin

- Announcements
- Evaluation
- Schedule

Course Goals

- Background
- Learning Objectives
- Topics
- Tools and Resources

Data and Knowledge Modelling and Analysis

- Instructor: Haitham Amar- hamar@uwaterloo.ca
- Material: learn.uwaterloo.ca
- Lectures: Monday 5:30pm-8:20pm
- TA: Iman Fadakar- ifadakar@uwaterloo.ca
- TA Office Hours: email Iman to arrange time for that week

Announcements

- Registering/Waiting List - course is full
- Log on to learn.uwaterloo.ca
  - enable email notifications
  - use the message boards, let me know if you want specific groups or categories, I can create them
  - talk to each other

Work load and Evaluation

- Homework 5%
- Assignments 15%
- Final Exam 50 - (Closed book but few cheatsheets allowed)
- Project 30%
    - Proposal 5%
    - Presentation 10%
    - Report 15%
- Assignments and Projects can be done in groups of 3

Weekly Homework - 5% of Final Grade

- Every once in awhile you'll have one or two questions to apply a concept from class to a given dataset.
- Due date found on the homework handout.
- Grading scheme:
  0 Did not hand in.
  1 Handed in, missing some major part, only partially done.
  2 Full answer attempted, correct or mostly correct.
- Handed in electronically as a PDF or a python/R notebook.

Assignments

- Two or three assignments (We'll adjust as the term progresses).
- A few datasets and some specific data cleaning, analysis, experiments to carry out and report back.
- Three weeks to complete.
- Should be in groups, same group as project.
- Report written as a PDF (Word, OpenOffice, **LaTex**)

Project

- Carry out an end-to-end data analysis project.
- **Application Oriented:** You have a problem, perhaps in your field of research, that you would like to analyze using the concepts and algorithms of this course.
- **Algorithm Oriented:** You select an interesting data analysis/machine learning technique that you want to learn more about. Then you find multiple datasets to test out the algorithm on and compare its performance against other algorithms.
- Groups: Projects should be worked on in groups of 2-3 people. Some can do on their own but you need to come to me and make a strong case for it.
- Detailed description: see project description on LEARN.
- Turnitin option will be turned on by default for the report to check report originality. You can opt out of this option. However, other originality checking methods will be deployed.

Course Dates

First Class	January 7, 2019
Last Class	Apr 1, 2019
Final Exam	To Be Determined

Important Dates (Subject to change - very much so. Really, subject to change)

Task	**Issued**	**Due**
Assignment 1	January 21	February 11
Project Proposal	soon	February 15
Assignment 2	February 25	March 18
Feedback on Proposal	February 25	
Project Presentations		March 18, 25, apr1?
Project Report		Apr 5

Course Admin

- Announcements
- Evaluation
- Schedule

Course Goals

- Background
- Learning Objectives
- Topics
- Tools and Resources

What are the goals of this course?

- Everyone has data to process, many tools and best practices already exist to do this.
- Data could come from experiments, databases, the internet, sensors or any other files.

This course aims to

- provide engineering graduate students with essential knowledge of data representation, grouping, mining and knowledge discovery.
- Level the playing field on data representation, processing, basic statistics, analysis, data mining.
- Introduce basic Machine Learning techniques.

Required Background

- Math and Linear Algebra : sets, marticies, transpose, cross product, dot product, matrix multiplication, solving system of linear equations
- Programming :
  - You should be comfortable programming in some language, not large software application but lots of calculations, plotting, etc.
- Writing and Presenting :
  - Assignments and Project require written reports, suggested tools : **LaTex** (local or shareLaTeX), Word, Google Doc
  - Presentation : Latex Beamer, Powerpoint, Keynote
- Probability and Statistics : (not required, we will define or review these, but it would help)
  - definition of probability, Bayes theorem, information, entropy, KL-divergence, probability distributions (Gaussian, Bernoulli, Poisson, . . .)
  - hypothesis testing, chi-squared

Learning Objectives

By the end of this course you will...

- Explain the sources and nature of data.
- Demonstrate how to best represent given data, summarize it, select proper metrics to evaluate the quality of the data and preprocess it for full analysis.
- Demonstrate ability to process data to extract useful information and knowledge.

Answering Questions

- What is Data?
- How do I prepare data for analysis to remove sources of error, bias, noise?
- What can I say about the questions I can answer using a given dataset?
- How do I train, test and evaluate my hypotheses using data?
- What algorithm are the most appropriate answer my questions using this data?

Topics to be covered

1. Data types, sources, nature, scales and distributions
2. Data representations, transformation, dimensionality reduction and normalization
3. Classification: Statistical based, Distance based, Decision based, Deep Learning.
4. Clustering: Partitional, Hierarchical, Model and Density based, others.
5. Retrieval and Mining: Similarity measures and matching techniques.
6. Reinforcement Learning: Classification, Control and learning patterns over time.
7. Knowledge discovery in data: Rule induction, Association rules mining, text mining.

Computing Resources

- Course Website:
  - No personal website- You can refer to http://markcrowley.ca/teaching
  - See **Computing Resources** page on website with tips on servers/systems you can use on campus.
- Sharcnet/Compute Canada
  - research students could have supervisor sponsor them to use Sharcnet, no cost.
- If you find useful resources, add to them the resources discussion forum on LEARN.

Other Tools and Resources

- Mendeley.com Community - online resource for academic papers. Course Group join and post your own papers or comments.
- Kaggle Competition - https://www.kaggle.com/datasets
- Cloud Services - free to use for single user, single machine smaller runs.
  - These have everything we'll cover in this course, we'll learn how to use them, why they are used, to allow you to go beyond them
  - Amazon Web Services (AWS)
  - Azure Tools - Microsoft

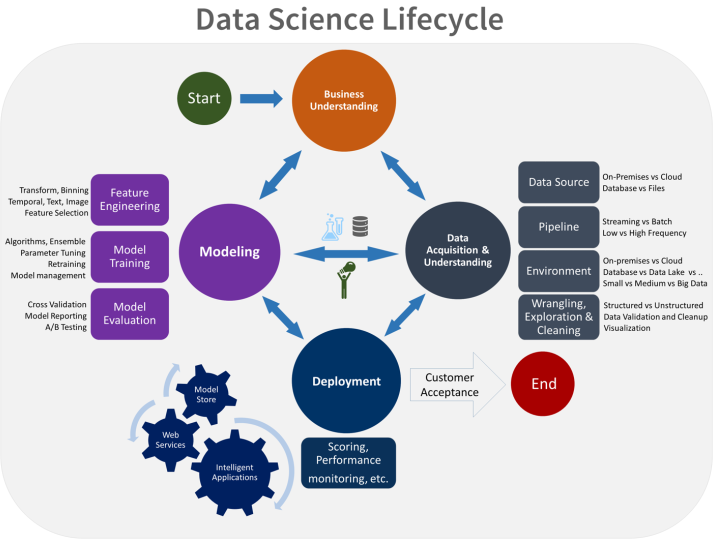
Tools for Data Management and Analysis

- **Only** two tools that you can to choose from ...
- python (The de facto programming choice for data scientists)
  - numpy, scipy, scikit-learn
  - Lots of resources online, communities, modules, new code tools all the time.
- R (Adoption has gone down)
  - The statistician's choice. Very powerful, less support from me, but large online community too.

Other Relevant Courses:

- ECE 657: Tools of Intelligent Systems Design
- ECE 750: Topic 5 - Distributed and Network-Centric Computing
- CS 489/698: Big Data Infrastructure
- CS 848/858: Models and Applications of Distributed Data Processing Systems
- CS 685: Machine Learning: Statistical and Computational Foundations
- STAT 841: Statistical Learning - Classification
- SYDE 675: Pattern Recognition (similar to this course)

Data science lifecycle:



Part II

Understanding and Preparing Data

Outline of Part II: Understanding and Preparing Data

Landscape of Topics in Data Analysis, AI, ML, Big Data...

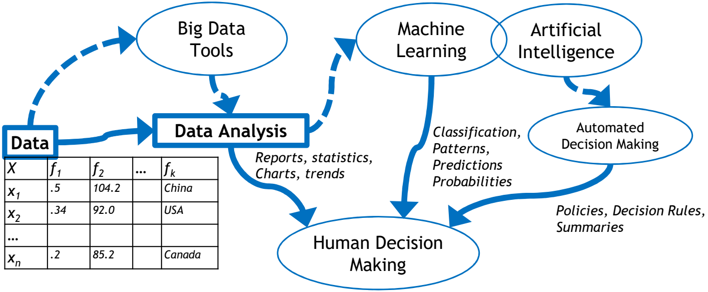
Data, Data Types and Information

- Types of Data
- Data Representations

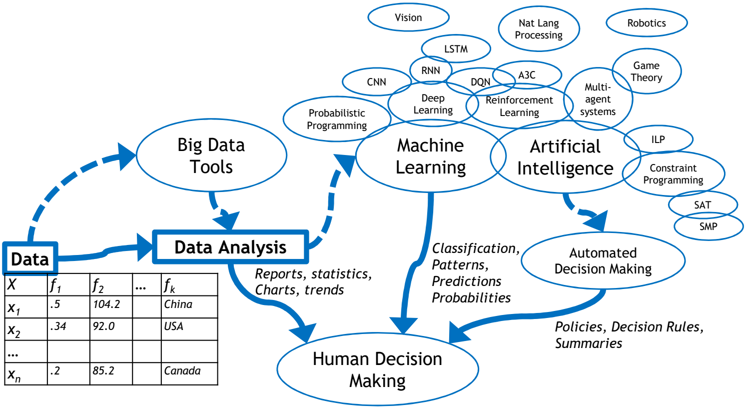
Summarizing Data

- Central Tendency
- Measures of Dispersion
- Multiple Variables

Data, Big Data, Machine Learning, AI, etc, etc,



Data, Big Data, Machine Learning, AI, etc, etc,



X	f_1	f_2	...	f_k
x_1	.5	104.2		China
x_2	.34	92.0		USA
...				
x_n	.2	85.2		Canada

Data Data Something Something. . .

- **Data Sources**: measurements from sensors, records, files, document, archives, transactions.
- **Data Modeling**: Creating a structure, organization, function or an abstract view of the data.
- **Data Analysis**: Transforming or operating on data to extract useful information, knowledge or conclusions.
- **Data Mining**: Carrying this further to discover unforeseen or hidden patterns in the data.

Big Data

Big data is about quality and performance given **huge** amounts of data, that is not primarily the focus of this course. But the tools and analysis methods we learn are part of the basis you need to deal with Big Data.

- Volume - Large amounts of data, social networks, phone, location, embedded systems, environmental, satellites, "full firehose"
- Velocity - Streaming, online data, arriving quickly, time series, real-time
- Variety - Heterogeneous (many types), many sources, category data, numerical data, continuous/discrete, text, images, audio, video

Big Data

Some people say it is also includes:

- **Veracity** - Solution requires: Accuracy, Confidence, Precision, Error
- **Variability** - Changes moment to moment, distribution can change at different times (seasonal, trends, fads, memes)
- **Complexity** - Combinatorial connections between entities in the data, form networks, hierarchies

Major Types/Areas of AI

Artificial Intellgience: some algorithm to enable computers to perform actions we define as requireing intelligence.

Major Types/Areas of AI

Artificial Intellgience: some algorithm to enable computers to perform actions we define as requireing intelligence. **This is a moving target.**

Major Types/Areas of AI

Artificial Intellgience: some algorithm to enable computers to perform actions we define as requireing intelligence. **This is a moving target.**

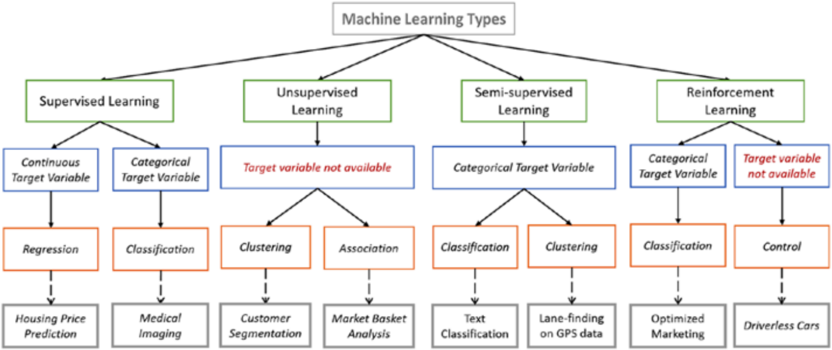
- Search Based Heuristic Optimization (A*)
- Evolutionary computation (genetic algorithms)
- Logic Programming (inductive logic programming, fuzzy logic)
- Probabilistic Reasoning Under Uncertainty (bayesian networks)
- Computer Vision
- Natural Language Processing
- Robotics
- **Machine Learning**

Food for Thought

- What do we mean by intelligence
- When do we say that a machine is intelligent
- Self reflection? Being the subject of their own thought?......

Types of Machines Learning

Machine Learning: *"Detect patterns in data, use the uncovered patterns to predict future data or other outcomes of interest"* – Kevin Murphy, Google Research.



Data and Information

One way to think about it...

- **Data** : Value that is measured (continuous, e.g 25, 108.3) or counted/observered (discrete, e.g male, married, 5). Data by itself does not have a meaning.
- **Information** : Interpreted data- adding meaning to data, understanding relations on data. e.g measured data is 25, measuring device is thermometer then the reading is temperature. The attribute temperature adds a meaning to the data.

What is Information?

Another way to think about it... **Entropy** measures the uncertainty that is resolved after observing a binary variable P_i.

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i$$

- If each trial is equally probable and independent then you can add them to get the cumulative entropy.
- If the next outcome is certain, then entropy is 0.
- Outcome of a coin flip provides 1 bit of information.

Types of Data Attributes

A data point has a set of *attributes* (also called dimensions, features or variables):

- 25, 30, -1.282, 8.3e5
- 1st, 3rd
- blue, red, green
- hi, med, lo

Types of Data (Qualitative)

- **Nominal**: no implication for quantity (qualitative)
  e.g occupation: engineer, teacher, dentist, bus driver. Or Color value: blue, red, green, Binary is a special case 0,1. (=, !=)
- **Ordinal**: relative ranking among values
  (i.e. order inNominal Ordinal relation to each other.
  e.g (hi,med,lo), (disagree,neutral, agree),(5,3,1), ($<$, $>$)

*[From [**?**] Chp 2]*

Intervals

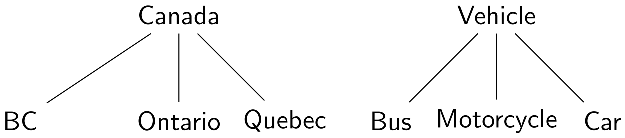
- **Interval:** like the ordinal type but on a scale of equal-size units i.e. a unit of measurement exists.
- Interpretation of numbers depends on the unit. The interval (range) is important for the interpretation. (+,-)
- E.g the significance of a mark of 10 will be different if the interval is 0-10 from that of interval of 0-100.
- Interval numbers represent differences between values, not absolute quantities.
- Temperature in C or F is an interval attribute

Ratios

- **Ratio:** like the interval in terms of order and uniformity but the scale has an inherent zero-point. (*, /)
- e.g Kelvin temperature scale for heat. 0 K means no heat, 50K is double the heat of 25 K.
- Cant say that for Celsius scale, 0 C is the amount of heat at freezing point.
- Used for physical quantities: height, weight, length etc.
- Also locations, distance, money

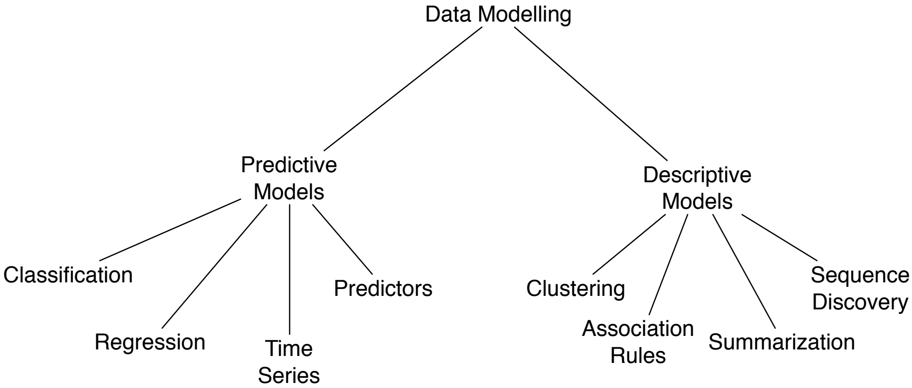
Structural Data

- Values are represented in a tree or hierarchy or graph
- Hierarchical structure could be whole-part, abstract-specific, classes-instances



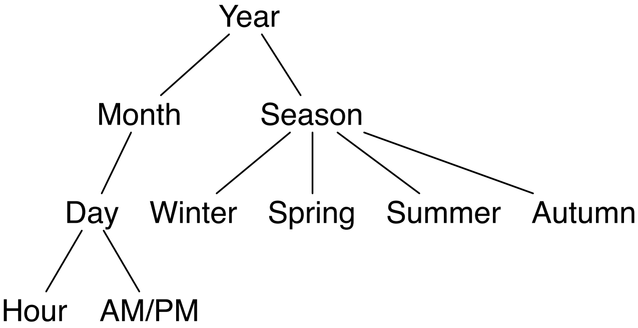
Examples of tree structured data.

Structural Data



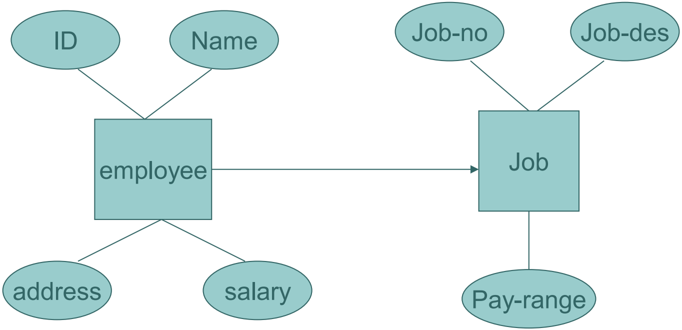
Examples of *class-instance* tree structured data.

Graphs or Trees



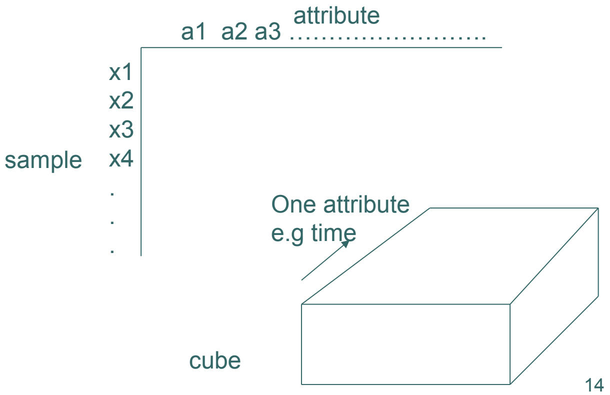
Databases

- data that has the same structure (schema) or abstract view independent of the physical layer



Lists/Vector/Matrix/Data Cube

- mainly table or vector or attribute-sample matrix which provides relational view



Summarizing Data

We have data we need to find patterns in it.

- Simplest pattern is a summary of the data.

Summarizing A Single Variable

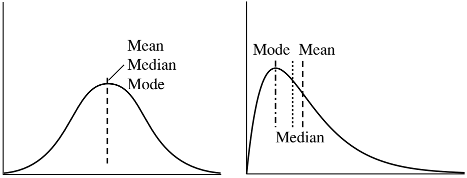
- Given a univariate sample $X_1, \ldots, X_n$ (could be Real, Natural, Integers)
- Goal: Summarize the variable compactly with a few numbers:
  - We want to summarize properties like spread, variation, range. Anything that can provide a summary statistic for the variable.
- Average : simplest and most common and estimate of central tendency.

$$\underline{\texttt{mean(x)}}) = \mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - **Pro:** If the samples come from a normal distribution then the average is the optimal estimate.
  - **Con:** Sensitive to outliers. (could be noise, data entry error, actual outliers)

Summarizing A Single Variable

- **Median:** If the samples are sorted then the median is the value that splits the list into half
- **Mode:** is the most common value in the list of samples (data can be bimodal or more)
- **Skew:** (third moment) high skew means the bulk of the data is at one end. Result: *Median* will be a better measure than mean.
- **Kurtosis:** (fourth moment) A measure of the heaviness of the tail of the distribution with respect to a set of points with a normal/Gaussian distribution and the same variance.



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Central Moments of a Set of Points

Mean(1), Variance(2), Skew(3) and Kurtosis(4) are unified by a single type of calculation on the n data points.

$$\mu_k \approx \int_{-\infty}^{\infty} (x-c)^n f(x) dx$$

$$\mu_k \approx \frac{1}{n-k+1} \sum_{i=1}^{n} (X_i - \mu_{k-1})^k$$

The 3rd and 4th moments are usually normazlied by s^k just as Standard Deviation is.

Types of Mean Functions

- **Trimmed Mean:** ignoring small percentage of highest and lowest values
- **Geometric Mean:**

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \leq \texttt{Mean} \tag{1}$$

$$= \exp\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right] \tag{2}$$

  - Arithmetic mean of logarithm transformed x
  - Good for positive values and output of growth rates
  - Most appropriate for ranking normalized results (different normalization can alter ordering for arithmetic or hamonic means)

Types of Mean Functions

- **Harmonic mean:** average of *rates*

$$H = \frac{n}{1/x_1 + 1/x_2 + \cdots + 1/x_n}$$

  - It is the reciprocal of arithmetic mean of the reciprocals of the sample points.
  - Appropriate for values that are inversely proportional to time such as "speedup".

Mean Examples

Data: `X=[1,1,1,1,1,1,100]`

- $n = 7$
- Mean=`sum(X)/n`=106/7=15.4
- Median=`median(X)`=1
- Mode=`Mode(X)`=1
- Trimmed mean(25%)=1
- Geometric Mean=1.9307
- Harmonic mean=1.1647

Measures of Dispersion: Variance and Deviation

- measure the spread of the data range
- **Standard Deviation:**

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - **Pro:** Same units as the data
  - **Con:** Sensitive to outliers
  - std(x)
- **Variance:**

$$var(x) = \sigma^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Variance and Deviation

- **Mean Absolute Deviation (MAD)**

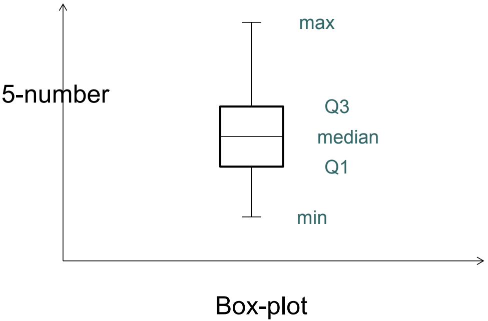
$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

  - Less sensitive to outliers than STD
- **Interquartile Range (IQR):** Difference between 75th (Q3) and 25th (Q1) percentile of data

Deviation Examples

Data: `X=[1,1,1,1,1,1,100]`

- $n = 7$
- Range=`range(X)/n`=99
- Std=`std(X)`=37.42
- MAD=`mad(X)`=24.24
- IQR=0



The **Pearson Correlation Coefficient (PCC)** is slightly more complicated way to analyse the relation between two attributes.

Pearson Correlation Coefficient (PCC)

- PCC measures of how strongly one attribute implies another

$$r = cov(v_1, v_2)/s_1 s_2$$

$$cov(v_1, v_2) = \frac{1}{n}\{(v_1 - \bar{v_1})(v_2 - \bar{v_2})^T\}$$

- **Interpretation:**
  - $-1 \leq r \leq 1$
  - -1 corresponds to negative correlation
  - $+1$ corresponds to positive correlation
  - Variance is a special case of covariance where $v_1 = v_2$
  - $r \neq 0$ implies dependency
- Independence implies covariance or correlation $=0$
- However, in general covariance or r=0 doesn't necessarily imply independence

PCC Examples

$$r = cov(v_1, v_2)/s_1 s_2$$

$$cov(v_1, v_2) = \frac{1}{n}\{(v_1 - \bar{v_1})(v_2 - \bar{v_2})^T\}$$

$$X = (2, 1, 3)$$

$$\bar{X} = 2 \quad S_X^2 = \frac{2}{3}$$

$$X - \bar{X} = (0, -1, 1)$$

$$Y = (1, 3, 2)$$

$$\bar{Y} = 2 \quad S_Y^2 = \frac{2}{3}$$

$$Y - \bar{Y} = (-1, 1, 0)$$

$$r = \left(\frac{1}{3}\right)\left(\frac{-1}{2/3}\right) = -0.5$$

PCC Examples

X=(2,1,3)	Y=(1,3,2)	r= -0.5	weak negative correlation
X=(2,1,2)	Y=(1,3,1)	r= -1	strong negative correlation
X=(2,1,2)	Y=(4,2,4)	r= 1	strong positive correlation
X=(2,1,2)	Y=(5,6,7)	r= 0	independent (really?)

Table: Some PCC exmaples

Cross Correlation

- Between two time series: association between values in the same time series separated by some lag $v_1(i), v_2(i)$
- Measures similarity between them by applying a time lag to one of them.
- It can be used to find repeated pattern or periodic nature so it can be used for prediction.
- Correlation coefficient r
- **Autocorrelation:** cross-correlation between two values at different points in time in the **same time series** (also called autocovariance)
  - series separated by some lag $v_1(i), v_1(i + lag)$
  - it can be used to find repeated pattern or periodic nature so it can be used for prediction.

$$R(s,t) = \frac{E[(X_t - \bar{x})(X_s - \bar{x})]}{\sigma_t \sigma_s}$$

Multivariate Data Representation

- Most common is sample-attribute matrix (pattern matrix or feature matrix or observation matrix)
- others: linked list, hierarchical

	attribute1	attribute2	attribute3………		attribute d
sample1	value11	value12	value13	....	value 1d
sample2	……	…..	…..	………	……
sample3	……	…..	…..	………	……
.					
.					
.					
.					
.					
.					
sample n	value 1n	value 2n	value 3n	………..	value nd

Part III

Data Preprocessing

Outline of Part III: Data Preprocessing

Data Examination and Cleaning

- Accuracy, Completeness, Consistency, Interpretability

Data Transformation

- Filling in Missing Data
- Smoothing with Bins
- Smoothing with Windows
- Normalization - Feature Scaling
- Scaling

Data Reduction

- via Sampling

Data Prepocessing Overview

- Examination of Data: Quality
  - accuracy, completeness, consistency, interpretability
- Data Cleaning: missing values, outliers, noise
- Transformation:
  - Smoothing with bins and windows
  - Normalization and Scaling
- Data Reduction: dimensionality, numerosity

Data examination

Data Quality:

- Accuracy: incorrect (eg.birthdates), inaccurate, transmission errors, duplicates
- Completeness: not recorded values, unavailable, ..
- Consistency: delete inconsistent data? acquire more data? average?
- Interpretability: how easily the data can be understood, correction of errors or removal of inconsistent data could make it harder to interpret

Data Cleaning

Examining the data to correct for:

- missing values
- outliers
- noise

Outline of Data Preprocessing

Data Examination and Cleaning

- Accuracy, Completeness, Consistency, Interpretability

Data Transformation

- Filling in Missing Data
- Smoothing with Bins
- Smoothing with Windows
- Normalization - Feature Scaling
- Scaling

Data Reduction

- via Sampling

Missing Values

- Use attribute mean (or majority nominal value) to fill in missing values
- If there are classes, use *attribute mean* (or majority nominal value) for all samples in the same class
- Can use prediction or interpolation to fill in missing values : linear, polynomial, spline
- Can remove the samples that have too many values missing

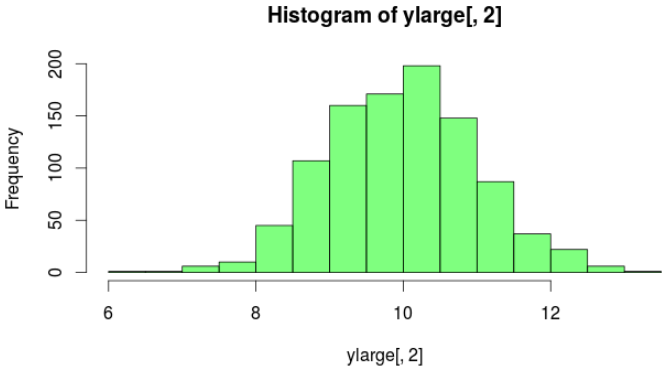
Dealing with Outliers or Noise

- **Detection:**
  - Use histograms to detect outliers
  - Use difference between mean, mode, median to indicate outliers
  - Use clustering to detect outliers.
  - Observe fluctuation in the values
  - Inconsistent values (negative values for positive attributes)
- **Fixing:**
  - Remove samples that are way out of range.
  - Smoothing the data to get rid of fluctuations.
  - Use logic check to correct inconsistency.
  - Use prediction methods or fitting.

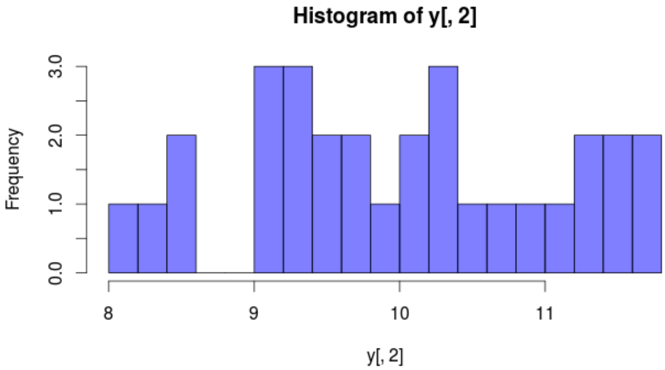
Anomaly Detection: Related, what if the outlier is what you are looking for?

Histograms for Outlier Detection

- Histogram is a bar plot of values vs frequency
- Values divided into **bins** (ranges of values)



Histograms for Outlier Detection



Outline of Data Preprocessing

Data Examination and Cleaning

- Accuracy, Completeness, Consistency, Interpretability

Data Transformation

- Filling in Missing Data
- Smoothing with Bins
- Smoothing with Windows
- Normalization - Feature Scaling
- Scaling

Data Reduction

- via Sampling

Binning Methods for Smoothing

Data Smoothing: Focus here is not correcting the data but softening it.

- Sort data and partition into bins
  - equal width
  - equal frequency by number of samples
- Smooth the values in each bin by:
  - replacing with the mean or median
  - replacing with the nearest bin boundary value

Binning Example

Sorted data: `[4,8,9,15,21,21,24,25,26,28,29,34]`
Using 3 bins of equal 4 samples.

	Bin 1	Bin 2	Bin 3

Binning Example

Sorted data: `[4,8,9,15,21,21,24,25,26,28,29,34]`
Using 3 bins of equal 4 samples.

	Bin 1	Bin 2	Bin 3
Binned Data:	4,8,9,15	21,21,24,25	26,28,29,34

Binning Example

Sorted data: `[4,8,9,15,21,21,24,25,26,28,29,34]`
Using 3 bins of equal 4 samples.

	Bin 1	Bin 2	Bin 3
Binned Data:	4,8,9,15	21,21,24,25	26,28,29,34
means	9,9,9,9	23,23,23,23	29,29,29,29

Binning Example

Sorted data: `[4,8,9,15,21,21,24,25,26,28,29,34]`
Using 3 bins of equal 4 samples.

	Bin 1	Bin 2	Bin 3
Binned Data:	4,8,9,15	21,21,24,25	26,28,29,34
means	9,9,9,9	23,23,23,23	29,29,29,29
boundaries	4,4,4,15	21,21,25,25	26,26,26,34

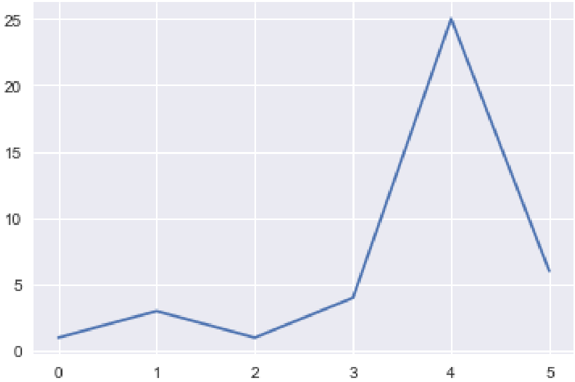
Smoothing within a Window

- If the values fluctuate so rapidly, we can do smoothing.
- Smoothing within a window using a moving average
- For example, for window size 3, using the median or mean to smooth
- i.e. mean or median of 3 consecutive values

Smoothing within a Window

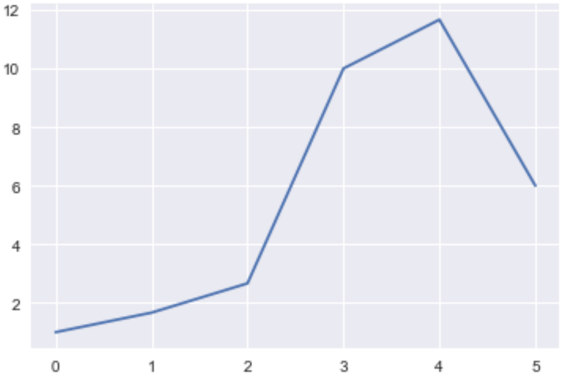
Example

X: [0,1,2,3,4,5,6]
Y: [1,3,1,4,25,6]



Smoothing within a Window

X: [0,1,2,3,4,5,6]
Y: [1,3,1,4,25,6]
smoothed Y : [1.67, 2.67, 10, 10.67, 11.33, 6]



Outline of Data Preprocessing

Data Examination and Cleaning

- Accuracy, Completeness, Consistency, Interpretability

Data Transformation

- Filling in Missing Data
- Smoothing with Bins
- Smoothing with Windows
- Normalization - Feature Scaling
- Scaling

Data Reduction

- via Sampling

Normalization

- Map the values $x_1, x_2, \ldots, x_n$ of the attribute A to a new value x_i' in the interval [0,1] (or any other interval).
- **Min-Max normalization:**

$$x_i' = \frac{x_i - \min_i x_i}{\max_i x_i - \min_i x_i}$$

  PRO: This makes the values invariant to rigid displacement of coordinates.
  CON: It will encounter an out-of-bounds error if a future input case for normalization falls outside of the original data range for A
- **Subtract the mean:** $x_i' = (x_i - \bar{A})$

[From [?] Chp 3.5]

Z-Score Normalization

- Z-score (standard score, standarization) normalization: Scale by mean and standard deviation

$$x_i' = (x_i - \bar{A})/\sigma_A \tag{3}$$

- Positive means value is above the mean, Negative means it is below the mean.
- Data is modified so that it now has the aggregate properties of a standard normal distribution $\mu = 0$ and $\sigma = 1$
- Many analysis and machine learning algorithms benefit from normalizing your data (decision tree methods are the main exception)

Pros and Cons of Standardization

- *Pro:* This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization
- Could be a good thing if you don't mind similar points treated the same (grades)
- *Con:* Normalization may or may not be desirable in some cases. It may make samples that are dispersed in space closer to each other and hence are difficult to separate.
- This could be unacceptable when small differences matter - safety, money in millions

Normalization Examples

$A = (-200, 400, 600, 800)$
Min$= -200$, Max$= 800$, max-min$= 1000$

Min-Max Normalization

$X' = (0, 0.6, 0.8, 1)$
mean$' = 0.6, \sigma_{X'} = 0.374$

Subtracting Mean Normalization

mean$=400$
$X'' = (-600, 0, 200, 400)$
mean$'' = 0$, $\sigma_{X''} = 100\sqrt{14}$

Z-score normalization

$X''' = (\frac{-6}{\sqrt{14}}, 0, \frac{2}{\sqrt{14}}, \frac{4}{\sqrt{14}})$
mean$''' = 0$, $\sigma_{X'''} = 1$

Normalization By Data Scaling

$$x' = \frac{x}{10^j}$$

- Where j is the smallest integer such that $\max|x'| < 1$
- Example:
  - if $x \in [-986, 917]$ then $max|x| = 986$ then $x' = 1000$
  - So -986 will normalize to -0.986 and 917 to .917
- *Note:* normalization can change the characteristics of the original data. But it is good for comparing values of different scales and reduces influence large numbers in the data.

Normalization of Matrix Data

- If A is the sample-feature matrix in terms of normalized data then let

$$R = \frac{1}{n} A^T A$$
$$r = \frac{1}{n} \sum_{k=1}^{n} x_{ki} x_{kj}$$

- Under subtraction normalization, R is a covariance matrix
- Under z-score normalization r_{ij} becomes the correlation coefficient between features i and j and $r_{jj} = 1$ for all j
- R is then called the correlation matrix.

Outline of Data Preprocessing

Data Examination and Cleaning

- Accuracy, Completeness, Consistency, Interpretability

Data Transformation

- Filling in Missing Data
- Smoothing with Bins
- Smoothing with Windows
- Normalization - Feature Scaling
- Scaling

Data Reduction

- via Sampling

Data Reduction

- **Goal:** improve performance, without hurting accuracy too much
- Numerosity Reduction
  - regression - replace many points with smaller number of predictions or function
  - clustering - much more on time on this later
  - sampling - to reduce data size
- Dimensionality Reduction
  - wavelet transforms
  - principle component analysis (more on this later)

Sampling for Data Reduction

- **Sampling:** obtaining a small subset of datapoints s to represent the whole data set n
- Allows a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the present of skew
  - Use adaptive sampling methods such as stratified sampling

[From [?] Chp 3.4.8]

Types of Sampling

- **Simple random sampling:** Draw a random number form the sample indices and select the object. Treats all sample as equally likely to be selected.
- **Sampling without replacement:** Remove the objects you select from the remaining samples. Original sample gets reduced every time you make a selection.
- **Sampling with replacement:** A selected object is put back in the original sample. You may select the same object more than one time.
- **Stratified sampling:** Similar to binning where the data set is partitioned and samples are selected from each partition. Good for skewed data.

Wrap-Up