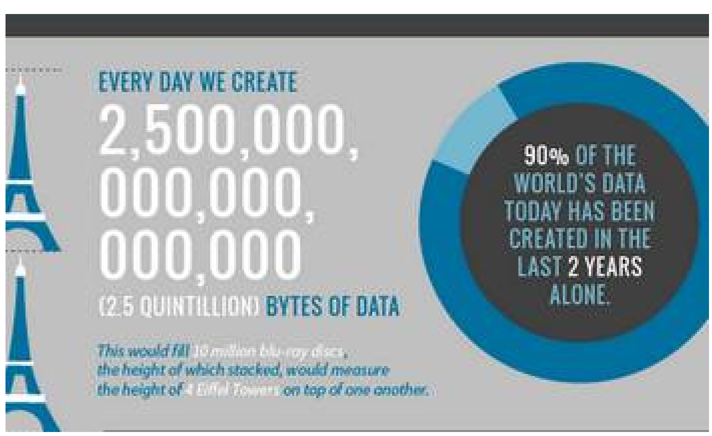
Introduction to Data Science

Classification (Supervised Learning)

(Created by HL Viktor: Based on subset of Chapters 8, 9 of Han et. al.)

1

(Machine) Intelligence Revolution?

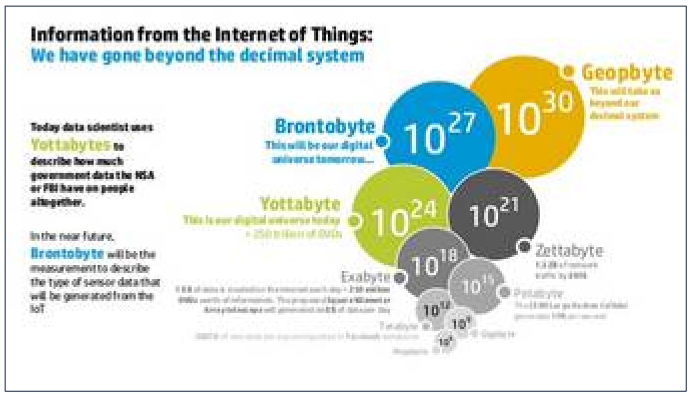


- http://dncapital.com/thoughts/beyond-big-data-to-data-driven-decisions/
- https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/human-capital/ca-EN-HC-The-Intelligence-Revolution-FINAL-AODA.pdf
- https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IML14576USEN

2

Data mining + Machine Learning

- Data driven discovery: making sense of the data deluge



Data mining + Machine learning

- Introduction and definitions to supervised learning
- KDD lifecycle
- Data mining example
- Data preprocessing
- Evaluation of results

Classification and Prediction

- Examples of “Supervised learning”
- We have historic data and the outcome is known
  - Past home owners with a home loan (mortgage):
    - mortgage paid on time (class 0: good)
    - house repossessed by bank (class 1: bad)
  - Heart Surgery patients in a hospital:
    - Back at home (class 0: good)
    - in general ward (class 1: recovering)
    - in Intensive Care (class 2: seriously ill)
    - Deceased (class 3: bad)

The goal of classification

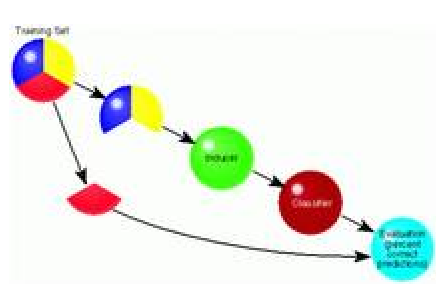
- We organize and categorize data in distinct classes
- We know the class labels and the number of classes
- E.g. Past Labor Negotiations (did they go no strike (or not))
- A model is created based on the data distribution
- The model is then used to classify new data
- Classification is used for the prediction of discrete and nominal values
  - Typically with classification, I aim to predict in which bucket to put the ball, not the exact weight of the ball.

The goal of prediction

- We aim to forecast the value of an attribute based on values of other attributes.
- E.g. Exchange Rate of Canadian Dollar to Euro
- A model is first created based on the data distribution.
- The model is then used to predict future or unknown values.
- Prediction is used for the prediction of numeric
    - Typically with prediction, I aim to predict the exact weight of the ball.

The phases of building a classifier (for now)

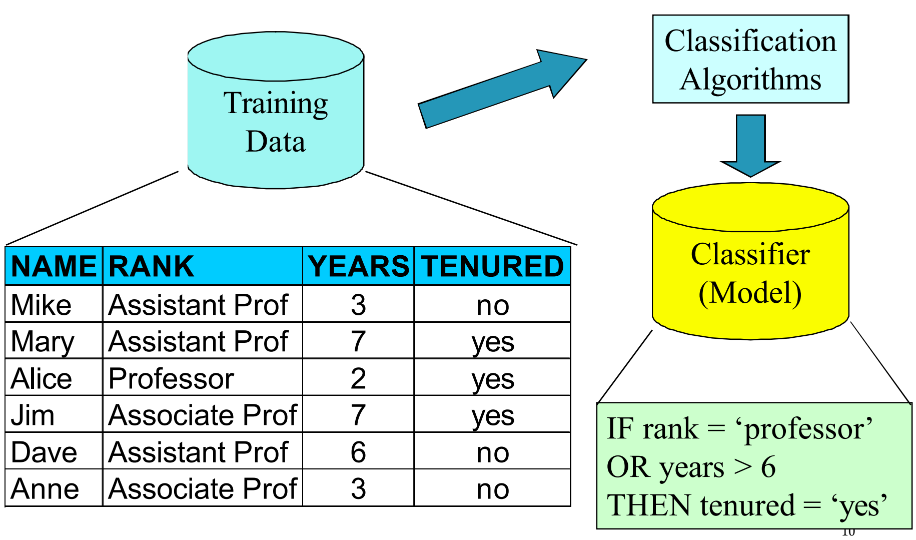
1. Divide the data into training and test data
2. Induce a classifier (model construction)
3. Test (model evaluation)
4. Use to predict new values (use model)



Classification—A Two-Step Process

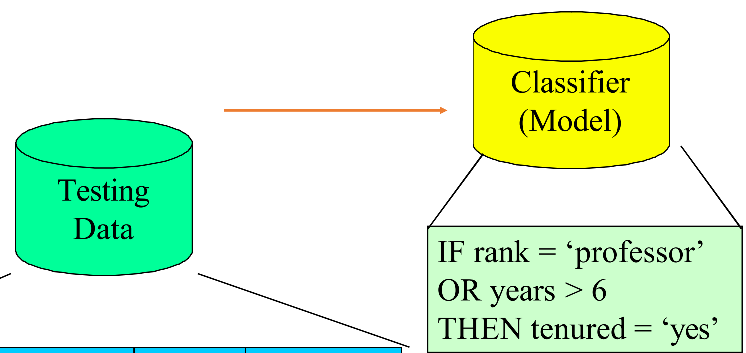
- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
  - 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
    - Test set is independent of training set, otherwise over-fitting will occur
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Alice	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

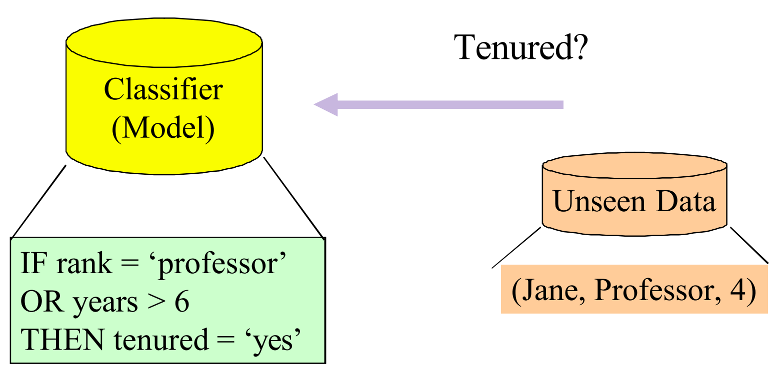
Process (2): Testing the Model against other data



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
George	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

11

Process (3): Using the Model in Prediction



12

Two important Issues

1. Data preparation
2. Evaluation

13

Preparing data for classification

Data transformation:

- Discretization of continuous data
- Normalization to [-1..1], [0..1], [0.1..0.9], z-score…
- Data Cleaning
- Smoothing to reduce noise

Relevance Analysis:

- Feature selection to eliminate irrelevant attributes

14

User Expectations versus Data Reality

- Decisions
  - Do we have enough data?
  - Do we have enough high quality data?
  - Do we have the ability to get enough high quality data soon?

  - Biggest risk → underestimating the difficulty to source your data
  - List success criteria: specific, measurable

15

Types of Data Sets and Data

- **Records:**
  - **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
  - Time series
  - Sequential Data: transaction sequences
  - Data streams
- Spatial, image and multimedia

	team	coach	pla y	ball	score	game	wi n	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

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

16

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

17

Databases and Data Objects

- Databases are made up of data objects ☺
- A **data object** represents an **entity**; with **relationships (1:M, N:M, 1:1)**
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples.*
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.

18

A word about Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

19

Attribute Types and Analytics

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
  - Issue: measuring "distance"
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender
  - <u>Asymmetric binary</u>: outcomes not equally important
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., Cancer positive)

20

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚ or F˚, calendar dates*
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *length, counts, monetary quantities*

21

Discrete vs. Continuous Attributes

- **Discrete Attributes**
  - Has only a finite or countably infinite set of values
    - E.g., postal codes, profession, or the set of words in a collection of documents
    - Many ML algorithms struggle with these (more later)
- **Continuous Attributes**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits

  Often we convert these to attribute bands, for data analysis

22

Attribute types: Questions

Issue: Some data mining techniques "favors" numeric versus nominal data, and vice versa

Initial Questions:

- Do we need to convert an attribute type (age to age-range)?
- Do we have an ordering (city → province → country)?
- Do we need to aggregate (individual sales to daily sales)?
- Do we need to combine values (auburn and brown hair)?
- How do we measure distance

Approaches

- Ask your users!!!!
- Done during data preprocessing once we got a feel of our data

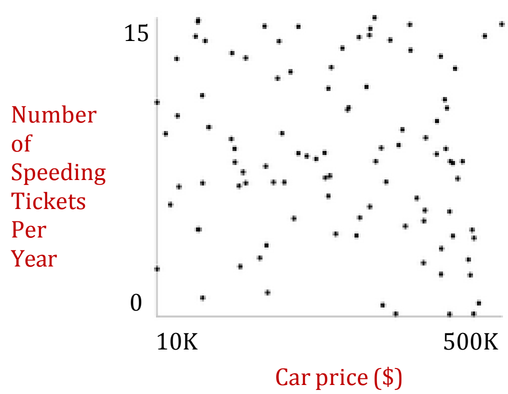
Descriptive data summarization

General idea: Get an overall picture of your data

See how it is distributed; if there is skew, if it has a high variance, and so on

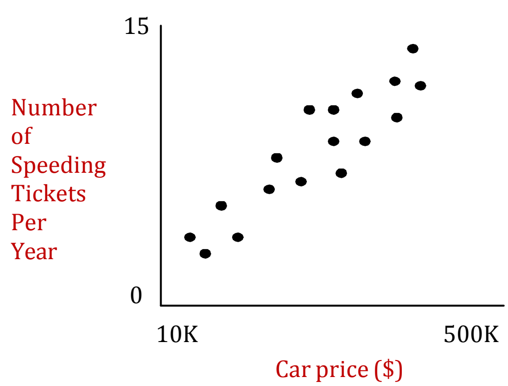
- Central tendencies
- Dispersion of data

Getting to know your data…



25

Getting to know your data…



26

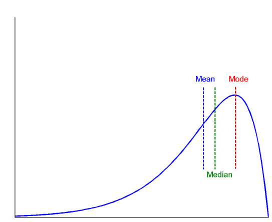
Measuring the Central Tendency

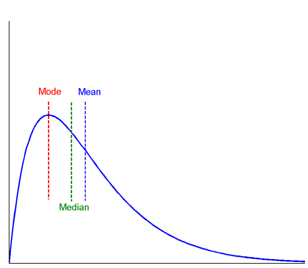
- Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: chopping extreme values
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*): $median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

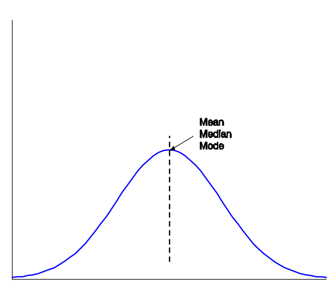
27

Symmetric vs. Skewed Data

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







28

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ*)
  - Variance: (algebraic, scalable computation)

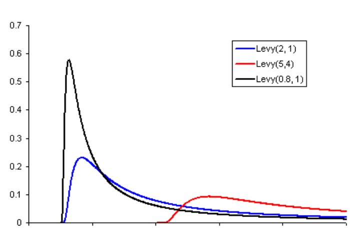
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

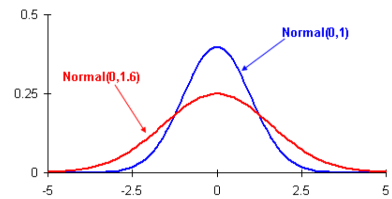
  - Standard deviation *s (or σ)* is the square root of variance *s^2 (or σ^2)*

29

Normal distribution: A strong assumption?

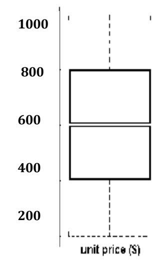
- Very often, we assume a normal distribution
- What if it is not? (e.g. earthquake, financial markets, ketchup sales…)





30

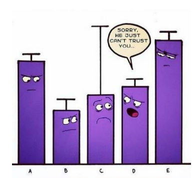
Boxplot Analysis



- Five-number summary of a distribution:

  Minimum, Q1, M, Q3, Maximum

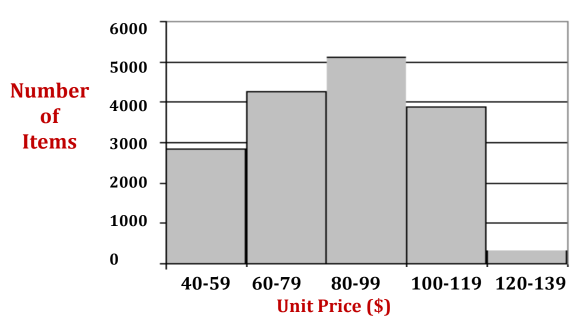
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum



31

Histogram Analysis

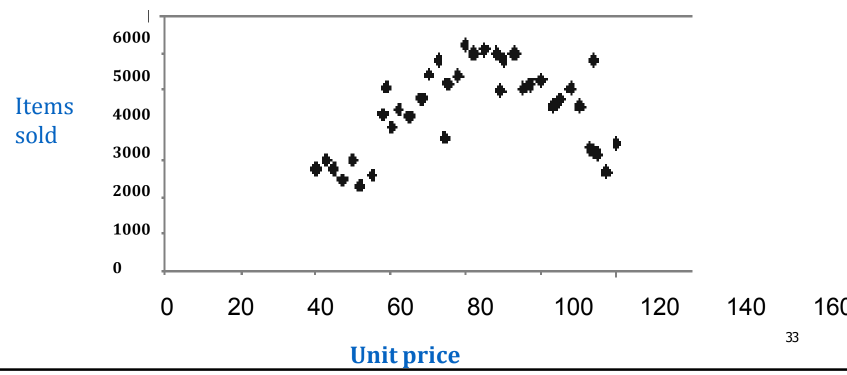
- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



32

Scatter plot: Often used

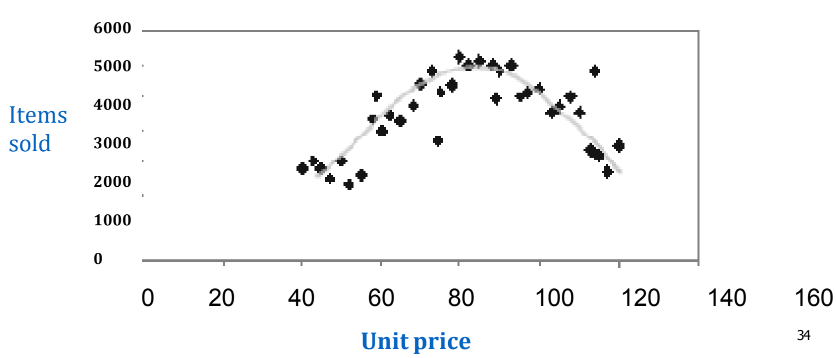
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



33

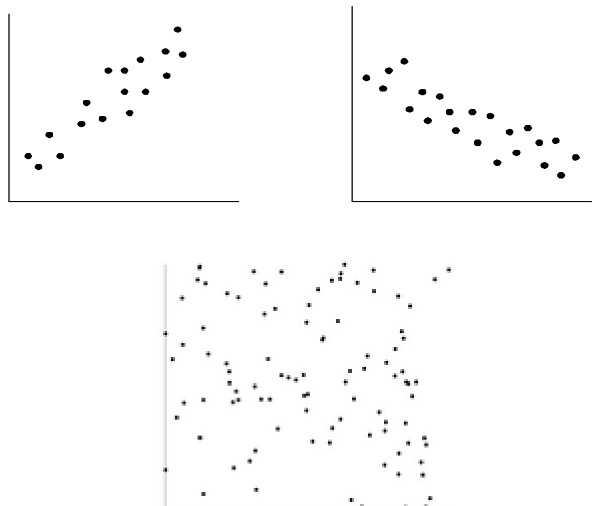
Loess (local regression) Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



34

Positively, Negatively and Uncorrelated Data



35

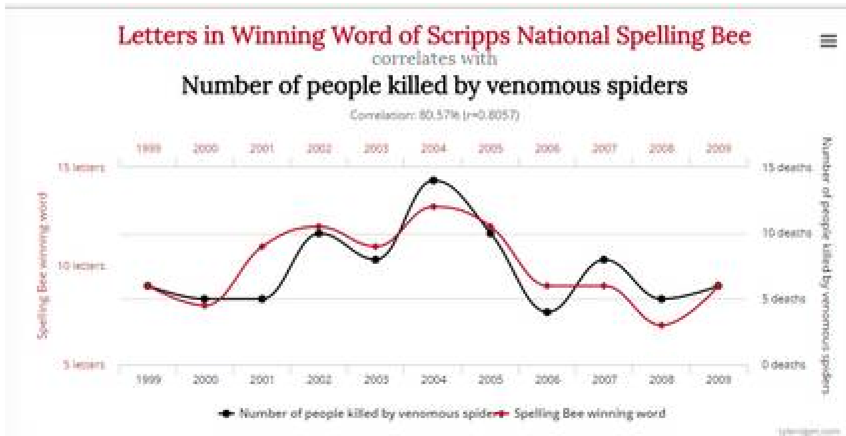
Evaluating Classification Methods

- Accuracy
  - classifier accuracy: predicting class label
  - predictor accuracy: guessing value of predicted attributes
- Speed
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- Robustness: handling noise and missing values
- Scalability: efficiency in disk-resident databases
- Interpretability
  - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules
- More later…

36

A word of caution…

- http://www.tylervigen.com/spurious-correlations
- We need to use our common sense!!!



37

Next…
Classification algorithms

38