Introduction

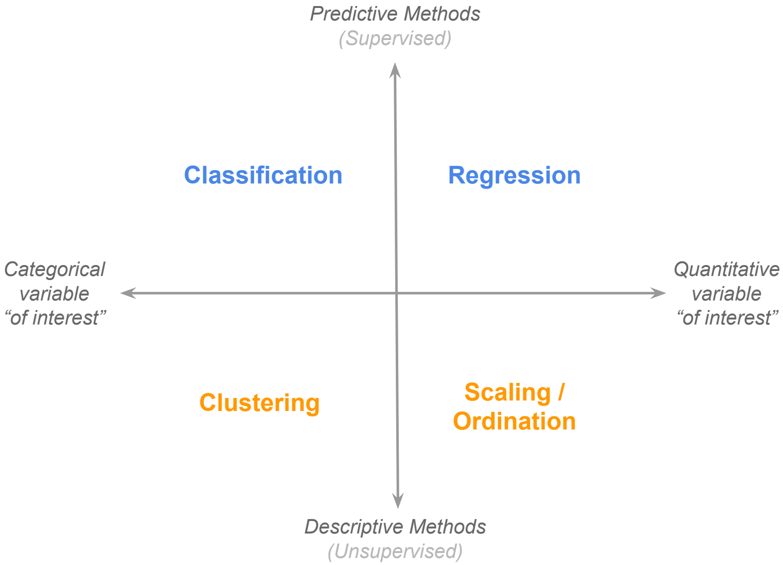
Intro 2 Statistical Learning

Gaston Sanchez

An introduction to Predictive Modeling and Statistical Learning

Statistical Learning Branches

Statistics	**Machine Learning**
Predictive methods	Supervised learning
Descriptive methods	Unsupervised learning



Predictive Methods
(Supervised)

Classification

Regression

Categorical variable “of interest”

Quantitative variable “of interest”

Clustering

Scaling / Ordination

Descriptive Methods
(Unsupervised)

A word of caution

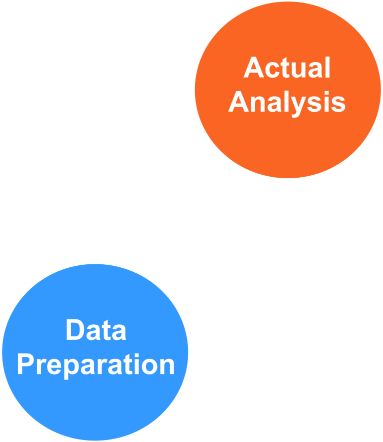
Sometimes there might not be a clear distinction between supervised and unsupervised learning. Often, a given method mixes both types of approaches.

Data Analysis Cycle (DAC)

Cycle of Data Anlaysis Projects

Data Preparation

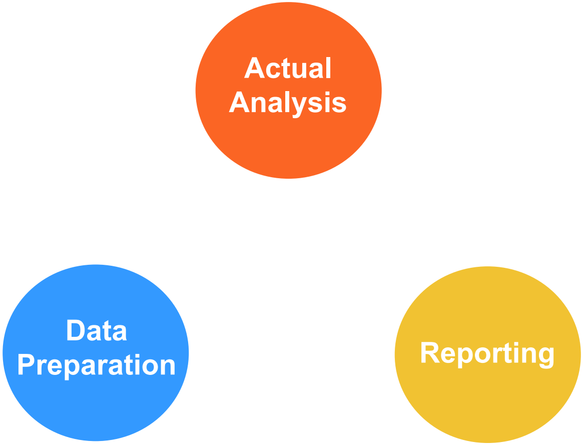
Cycle of Data Anlaysis Projects



Actual Analysis

Data Preparation

Cycle of Data Anlaysis Projects

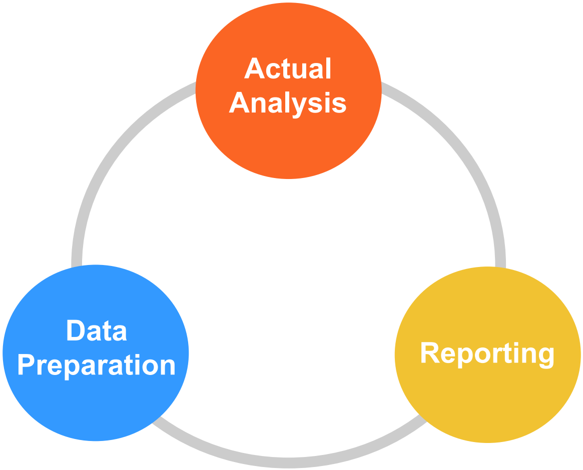


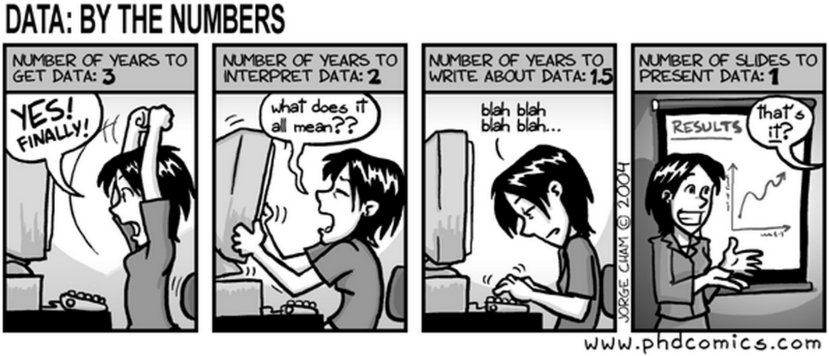
Actual Analysis

Data Preparation

Reporting

Cycle of Data Anlaysis Projects




DATA: BY THE NUMBERS
NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
JORGE CHAM © 2004
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
www.phdcomics.com

http://www.phdcomics.com/comics/archive.php/archive/tellafriend.php?comicid=462

Data Preparation

Core Data Analysis

Reporting

Communication

NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?

Keep in mind

NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!

Data

NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??

Analysis

NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...

Report

NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?

Communication

(Some) Major Data Analysis Tasks

- **Visualization**: to facilitate human discovery
- **Summarizing**: describing information
- **Deviation Detection**: finding changes
- **Profiling**: finding relevant characteristics of a group of individuals
- **Associations**: finding relationships, e.g. A & B & C occur frequently
- **Clustering**: finding groups in data

(Some) Major Data Analysis Tasks

- **Visualization**: to facilitate human discovery
- **Summarizing**: describing information
- **Deviation Detection**: finding changes
- **Profiling**: finding relevant characteristics of a group of individuals
- **Associations**: finding relationships, e.g. A & B & C occur frequently
- **Clustering**: finding groups in data
- **Prediction**

Keep in mind

Data

Analysis

Report

Communication

This is where predictive modeling activities tend to take place

Keep in mind

NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!

Data

NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??

Analysis

NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...

Report

NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?

Communication

In practice these are where we spend most of our time

Official Title

Modern Statistical Prediction?

Modern Statistical Prediction?

- Statistical Prediction is not a new task
- Predictive applications (least squares) date back to 18th-19th century (Andrien-Marie Legendre -vs- Carl Friedrich Gauss)
- Regression framework originated at the beginning of 20th century (Francis Galton, Karl Pearson, Udny Yule)
- Classification framework originated around the 1930s (Ronald Fisher, P.C. Mahalanobis, B.L. Welch)

So where does the “modern” part come from?

Modern Statistical Prediction

So where does the “modern” part come from?

- Model concept
- Data Sets
- Fields of Application
- Computing Tools
- Mathematical/Algorithmic Tweaks
- Predictive performance assessment
- Modeling Pipeline

Concept of a Model

- Term “model” appeared in the 1930s (econometric models).
- The concept of model has not remained static.
- Way of thinking about what a model varies across disciplines.
- Even within the same community, there may be different ideas of “model”.

Concept of a Model

- Suppose we observe a response Y
- We also observe p different predictors, $X_1, X_2, \ldots, X_p$
- We assume Y is related with $[X_1, \ldots, X_p]$
- The relationship can be written in a general form as

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

Concept of a Model

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

- $f()$ represents the systematic information—the *signal*—that the predictors provide about Y
- ϵ represents an *error* term—the *noise*—that is a catch-all for what we miss with the model

What kind of $f()$?

What kind of $f()$?

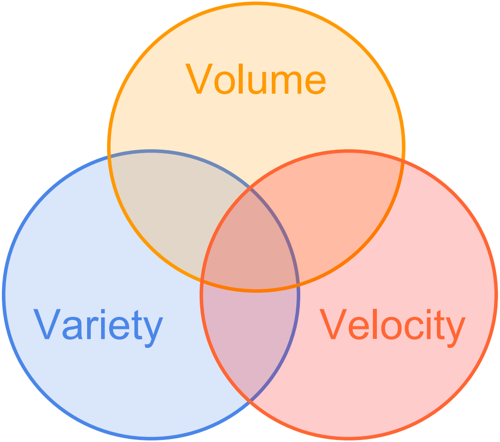
$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

- In “classic” statistics, $f()$ takes the form of a function (with parameters to be estimated)
- Within statistical learning, $f()$ is more open-ended
- It can also take the form of an algorithm
- Sometimes $f()$ is a *black box*

So where does the “modern” part come from?

Data Sets

The three V's of Data



(3Vs from conversation with Prof. David Ackerly)

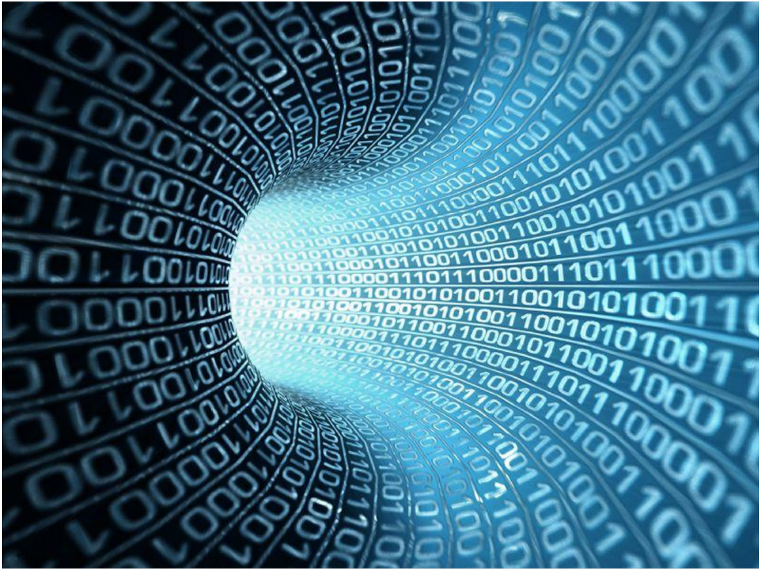
Modern Statistical Prediction?

The three V's

- **Volume**: larger data sets with reduced storage cost.
- **Velocity**: increasing rate at which data is produced/recorded.
- **Variety**: new types of data, more diverse/complex.

Volume

Velocity

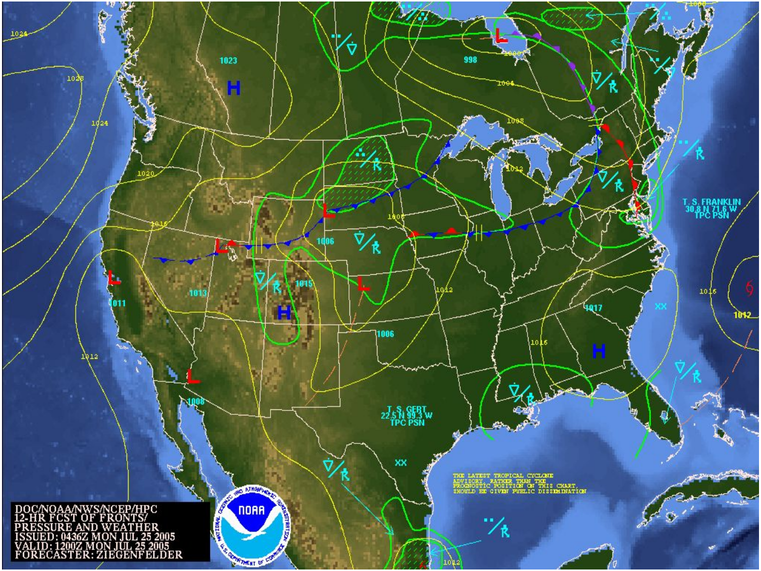


Variety

Variety

Flag Ref
0910
Abnormal
ted: 01
K/uL
M/uL
gm/dL
fL
pg
PAGE
** CONTINUED ON
GE **

Variety


DOC/NOAA/NWS/NCEP/HPC
12-HR FCST OF FRONTS/
PRESSURE AND WEATHER
ISSUED: 0436Z MON JUL 25 2005
VALID: 1200Z MON JUL 25 2005
FORECASTER: ZIEGENFELDER
NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE
T. S. FRANKLIN
30.8 N 71.6 W
TPC PSN
T. S. GERT
22.5 N 99.3 W
TPC PSN
THE LATEST TROPICAL CYCLONE
ADVISORY, RATHER THAN THE
PROGNOSTIC POSITION ON THIS CHART,
SHOULD BE GIVEN PUBLIC DISSEMINATION
H
L
1023
998
1004
1008
1012
1006
1015
1013
1011
1017
1016
1024
1020
1012
1008
XX

Variety

LinkedIn
skype
LiVEJOU
foursquare
vime
YouTube
Insta
facebook
Blogger
my
stagram
twitter
flickr
tumblr.
yelp
Lik

So where does the “modern” part come from?

Fields of Application

Fields of Application Example

For instance, consider the history of PLS Regression

(I'll talk about this with more detail when we study PLSR)

- Origins in mid-1960s with Herman Wold
- As a side-project Wold deviced a series of algorithms based on Least Squares steps
- First applications in Psychometrics and Econometrics
- Karl Joreskog (Wold's former PhD student) disruption of Structural Equation Models (1970s)
- Explosion of applications in Education, Sociology, Psychology

Fields of Application Example (cont'd)

For instance, consider the history of PLS Regression

(I'll talk about this with more detail when we study PLSR)

- Inspired by Joreskog's work, Wold's refined his framework
- Extension to multivariate regressions and systems of equations
- Herman Wold's framework poorly acknowledged (for various reasons)
- Applied to chemometrics in late 1970s
- Further adaptations by his son Svante Wold, and Harald Martens
- New regression approach via Partial Least Squares

So where does the “modern” part come from?

Mathematical/Algorithmic Tweaks

Mathematical/Algorithmic tweaks example

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

In matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Mathematical/Algorithmic tweaks example

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

In matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Predicted model

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{b}$$

Mathematical/Algorithmic tweaks example

OLS solution given by minimizing the residual sum of squares:

$$min \quad \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_j \right)^2$$

in vector-matrix notation:

$$min \quad \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

Mathematical/Algorithmic tweaks example

Assuming that $\mathbf{X}$ is of full column-rank, the OLS solution for

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{b}$$

is given by:

$$\mathbf{b} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Mathematical/Algorithmic tweaks example

Potential instability—due to multicollinearity—in the OLS solution affecting

$$(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

Mathematical/Algorithmic tweaks example

One option: Find inverse of $(\mathbf{X}^\top \mathbf{X})$ by looking for an orthogonal basis:

$$(\mathbf{X}^\top \mathbf{X})^{-1} \approx \mathbf{V}\mathbf{\Lambda}_*^{-1}\mathbf{V}^\top$$

Mathematical/Algorithmic tweaks example

One option: Find inverse of $(\mathbf{X}^\mathsf{T}\mathbf{X})$ by looking for an orthogonal basis:

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \approx \mathbf{V}\mathbf{\Lambda}_*^{-\mathbf{1}}\mathbf{V}^\mathsf{T}$$

Another option: Modify $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-\mathbf{1}}$ by adding a small constant k to the diagonal entries of $\mathbf{X}^\mathsf{T}\mathbf{X}$ before taking the inverse:

$$\mathbf{X}^\mathsf{T}\mathbf{X} + k\mathbf{I}$$

So where does the “modern” part come from?

Concept of “Predictive Modeling”

Modeling Goals

A statistical model typically aims to

Provide a certain comprehension of the data and the mechanism that generated them through a parsimonious representation of a random phenomenon.

Sometimes also, a statistical model seeks to

Predict new observations with “good” accuracy.

Modeling for what?

Goal Tradeoff

Understanding -vs- Prediction

Introduction

Understanding?

Understand could mean a model of a distribution for a random vector but it could also mean a regression model.

From a classic point of view, a model should be simple, and its parameters should be interpretable in terms of its domain of application (e.g. elasticity, odds-ratio, etc).

Paradoxes

Paradox 1

A “good” statistical model does not necessarily gives accurate predictions (at an individual level). E.g. risk factors in epidemiology.

Paradoxes

Paradox 1

A "good" statistical model does not necessarily gives accurate predictions (at an individual level). E.g. risk factors in epidemiology.

Paradox 2

We can predict without understanding

- no need for a theory of consumer to predict marketing target
- a model may be just simply an algorithm

Inference

Classic Inferential Statistics

Methodology for extracting information from data and expressing the amount of uncertainty in decisions we make.

- Assume distributions for the data
- Inferential aspects
- More theory-based
- More focused on testing hypotheses

So where does the “modern” part come from?

Assessing Predictive Performance

Model Performance

How do we define what a “good” model is?

- A model that fits the data well?
  (e.g. minimize resubstitution error)
- A model with optimal parameters?
  (e.g. most likely coefficients)
- A model that adequately predicts new (unseen) observations?
  (e.g. minimize generalization error)

Predictive Modeling

The Process of developing a mathematical tool or model that generates an accurate prediction.

Kuhn and Johnson, 2013

Predictive Modeling

The art of building and using models that make predictions based on patterns extracted from historical data.

Kelleher et al, 2015

Model Performance

- From the predictive modeling standpoint, a “good” model is one which gives accurate predictions.
- By *predictions* we mean predictions of new data.
- Therefore we focus on the generalization ability of the model to predict unobserved data
- This involves finding measure(s) of accuracy for predictions.

So where does the “modern” part come from?

Modeling Pipeline

Cycle of DAP and Predictive Modeling

- Data collection
- Data preparation (cleansing, formatting, transformations)
  - Feature selection
  - Feature extraction
- Model Building
  - Select modeling techniques
  - Select validation approach
  - Find optimal model
- Evaluation
- Deployment (decision making)

Predictive Modeling Process

Main Considerations

1. What data do you have?
2. What do you want to predict about the data?
3. What predictive methods/techniques should you use?
4. How accurate predictions look like?
5. What is the predictive performance?
6. Is there overfitting?

Predictive Modeling

We think that good data analysis depends not only on clear thinking but also on substantive knowledge. Mere numerology will not do, nor is there a good cookbook.

David Freedman, 1987

Terminology (Lebart, 1995)

Statistics	Machine Learning
Variables	Attributes (fields)
Individuals (objects, observations)	Instances (records, samples)
Predictors (independent)	Input
Response (dependent)	Output (target)
Model	Machine
Coefficients	Weights
Fit Criteria	Cost function
Estimation	Learning
Prediction	Supervised
Structure	Unsupervised

Bibliography

- **Modern Multivariate Statistical Techniques** by Izenman (2008). Springer.
- **Applied Predictive Modeling** by Kuhn and Johnson (2013).
- **Fundamentals for Machine Learning for Predictive Data Analytics** by Kelleher et al (2015).