STT481 Capstone in Statistics

Lecture 0: Introduction

Chih-Li Sung

09/02/2020

Statistics in the news

- Quote of the Day, New York Times, August 5, 2009

  "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."

– Hal Varian, chief economist at Google

- Quote of Forbes, Dec 11, 2017

  "LinkedIn's Fastest-Growing Jobs Today Are In Data Science And Machine Learning"

– Louis Columbus

Are you ready to be a statistician/data scientist

- Data scientist skill-set (source)

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

What we have learned so far

- Basic Statistics
- Probability
- Hypothesis Testing
- Linear Regression (maybe?)
- . . .

What this course is about

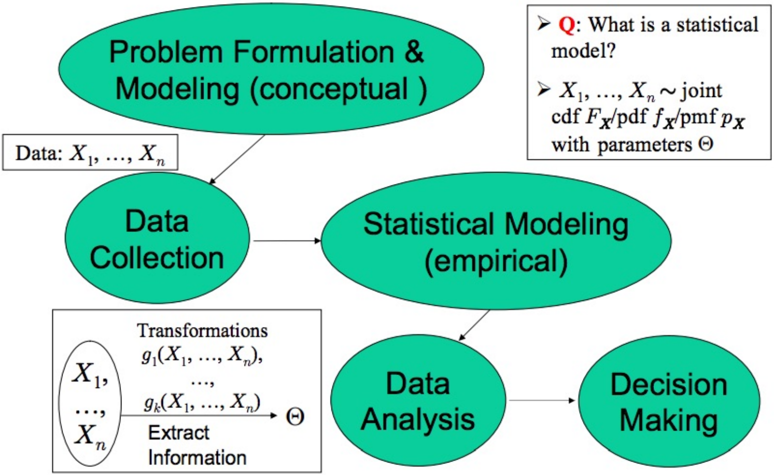
- Statistical capstone experiences are essential for statisticians/data scientists to perform an in-depth analysis of real-world data.
- Capstone experiences can develop statistical thinking by engaging in a consulting-like experience that requires skills outside the scope of traditional courses:
  - defining a complex problem,
  - analyzing data,
  - building a strong team,
  - programming techniques,
  - and communicating effectively.

What you should expect to learn in this course

- Problem formulation
- Data collection
- Advanced statistical modeling, preliminary data analysis, and machine learning
- Statistical software (R)
- Thorough and elaborate statistical analyses of data
- Presentation and data visualization

A standard procedure of statistical analysis

- Statistics divides the study of data into *five* steps:



Problem Formulation & Modeling (conceptual)

Data: $X_1, \ldots, X_n$

Data Collection

Statistical Modeling (empirical)

Data Analysis

Decision Making

- **Q**: What is a statistical model?
- $X_1, \ldots, X_n \sim$ joint cdf $F_{\mathbf{X}}$/pdf $f_{\mathbf{X}}$/pmf $p_{\mathbf{X}}$ with parameters Θ

Transformations $g_1(X_1, \ldots, X_n)$, …, $g_k(X_1, \ldots, X_n)$

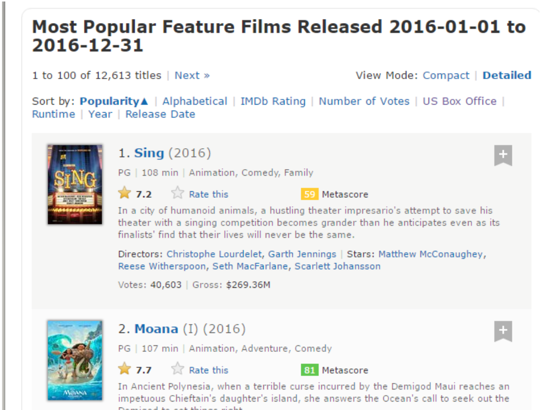
$X_1, \ldots, X_n \xrightarrow{\text{Extract Information}} \Theta$

Figure 2:

Data collection

“In God we trust; all others bring data.”" – Edwards Deming

- Design of experiment (not covered)
- Survey sampling (not covered)
- Web scraping
- Text mining (not covered)



Most Popular Feature Films Released 2016-01-01 to 2016-12-31

1 to 100 of 12,613 titles | Next »

View Mode: Compact | **Detailed**

Sort by: **Popularity▲** | Alphabetical | IMDb Rating | Number of Votes | US Box Office | Runtime | Year | Release Date

1. **Sing** (2016)

PG | 108 min | Animation, Comedy, Family

★ 7.2 ☆ Rate this 59 Metascore

In a city of humanoid animals, a hustling theater impresario's attempt to save his theater with a singing competition becomes grander than he anticipates even as its finalists' find that their lives will never be the same.

Directors: Christophe Lourdelet, Garth Jennings | Stars: Matthew McConaughey, Reese Witherspoon, Seth MacFarlane, Scarlett Johansson

Votes: 40,603 | Gross: $269.36M

2. **Moana** (I) (2016)

PG | 107 min | Animation, Adventure, Comedy

★ 7.7 ☆ Rate this 81 Metascore

In Ancient Polynesia, when a terrible curse incurred by the Demigod Maui reaches an impetuous Chieftain's daughter's island, she answers the Ocean's call to seek out the Demigod to set things right

Statistical modeling

> *"Essentially, all models are wrong, but some are useful." – George Box*

- Supervised learning
  - Linear regression model
  - Logistic Regression
  - Linear Discriminant Analysis
  - KNN
  - Linear Model Selection and Regularization
  - Nonparametric Regression
  - Tree-Based Methods
  - Support Vector Machines
- Unsupervised learning
  - Principal Components Analysis
  - Clustering Methods

Decision making/visualization

- ggplot2
- plotly
- shiny (myapp)
- D3.js

Grading

- Final grades will be based on
  - one course project (20%)
  - five homework assignments ($50\% = 5 \times 10\%$)
  - two midterms ($30\% = 10\% + 20\%$)
- The final grade would be based on your total grade percentage and will be determined roughly as:

```
## Warning: package 'knitr' was built under R version 3.5.
```

```
##
## %      90-100 85-89 80-84 75-79 70-74 65-69 60-64 0-59
## Grade       4   3.5     3   2.5     2   1.5     1    0
```

Homework

- Homework assignments include conceptual and applied exercises. Typesetting your reports/solutions in *Latex* or *R markdown* (link1,link2) is strongly encouraged. Unreadable handwriting is subject to zero credit. Unreadable handwriting is subject to zero credit.
- Homework Policy: link

Project

- Kaggle is a open platform for predictive modeling and analytics competitions, where companies and researchers can post their data and problems for users to solve.
- In this project, you are given the house dataset on Kaggle and the goal is to predict the final price of each home in Ames, Iowa. See link.
- Grading:
    - Accurate prediction (10%) -

  min{15%, 0.12/(your prediction score)×10%}

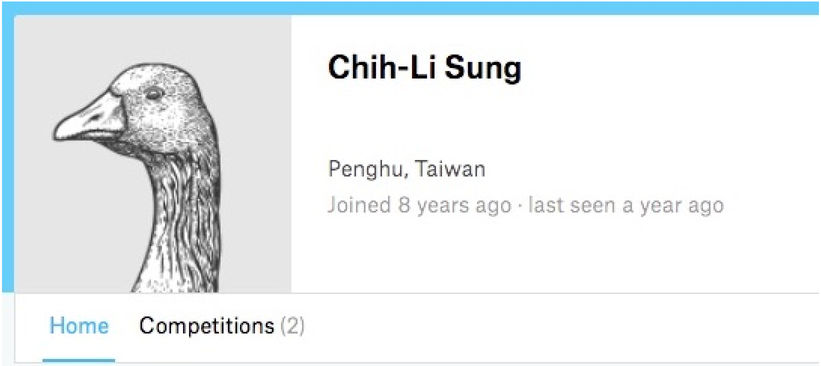
    - Final Report (10%)
- Final Report: See link.

Important Dates and Policy

See syllabus

Any questions or feedback?

Homework 1, Q1: Create an account on Kaggle (take a screenshot like the image below). (Homework questions can be found on D2L/Assignments)



Chih-Li Sung

Penghu, Taiwan

Joined 8 years ago · last seen a year ago

Figure 4:

Introduce yourself

- Your name
- Major
- Senior or Junior
- What time is it now in your place