STT481 Capstone in Statistics Lecture 0: Introduction

Chih-Li Sung

08/28/2019

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 <u>Forbes</u>, 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 (<u>source</u>)

MODERN DATA SCIENTIST

Data Sointist, the sexiest job of 2th 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 neguly hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTIC

- ☆ Machine learning
- 🖈 Statistical modelin
- 🖨 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

- Passionale about the business
- ✿ Curious about data
- ✿ Influence without authority
- ✿ Hacker mindse
- 🖈 Problem solve
- Stategic, proactive, creative innovative and collaborative

PROGRAMMING & DATABASE

- 🗢 Computer science fundamenta
- 🖈 Scripting language e.g. Pythor
- ☆ Statistical computing package e.g.
- 🖈 Databases SQL and NoSQL
- 🖈 Relational algebr
- Parallel databases and parallel que processing
- ☆ MapReduce conce
- 🖈 Hadoop and Hive/Pi
- 🖈 Custom reduce
- ☆ Experience with xaaS like AW

COMMUNICATION & VISUALIZATION

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

MarketingDistillenceme is a group of prachitimens in the area of a common marketing. Our fields of coperties includes marketing state by and aprimization construent backing and on oble analytics are predictive analytics and econometics: data worthousing and the data systems marketing channels in only this Pradicationet. SUS, Social CPM and Nami-



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:



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 Po 2016-12	pular Feature Films R 2-31	eleased 2016-01-01 to				
1 to 100 of 12	,613 titles Next »	View Mode: Compact Detailed				
Sort by: Popu Runtime Yea	llarity▲ Alphabetical IMDb Rating r Release Date	Number of Votes US Box Office				
SING	1. Sing (2016) PG 108 min Animation, Comedy, Famil ★ 7.2 ★ Rate this ★ 1.2 ★ Rate this the singing competition becom- finalists' find that their lives will never be Directors: Christophe Lourdelet, Garth Je Reese Witherspoon, Seth MacFarlane, Sc Votes: 40,603 Gross: \$269.36M	y 2 Metascore the thetar impresario's attempt to save his es grander than he anticipates even as its the same. mings Stars: Natthew McConaughey, ariett Johansson				
Ŕ	2. Moana (I) (2016) PG 107 min Animation, Adventure, Co ☆ 7.7 ☆ Rate this In Ancient Polynesia, when a terrible curs impetuous Chieftain's daughter's island, s	medy Metascore in incurred by the Demigod Maui reaches an he answers the Occean's call to seek out the				

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



Grading

Final grades will be based on

- one course project (20%)
- five homework assignments $(40\% = 5 \times 8\%)$
- two midterms ($40\% = 2 \times 20\%$)
- The final grade would be based on your total grade percentage and will be determined roughly as:

##

##	%	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

Class Participation:

- Your class participation will be used to determine whether your grade can be lifted in case you are right on the edge of two grades.
- Participation means attending classes, participating in class discussions, asking relevant questions, volunteering to provide answers to questions, and providing constructive criticism and creative suggestions that improve the course.

Homework

Homework assignments include conceptual and applied exercises. Typesetting your reports/solutions in *Latex* or *R* markdown (link1,link2) is strongly encouraged (you will receive 5 extra credit points). Unreadable handwriting is subject to zero credit. Unreadable handwriting is subject to zero credit.

Homework Policy: <u>link</u>

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, lowa. See <u>link</u>.
- You may form groups of up to three students.
- Grading:
 - Accurate prediction 0.12/(your prediction score)×10% on 12/11 7:45am (10%)
 - Model Interpretation & Final Presentation on 12/11 7:45am -9:45am (10%)
- Final Presentation: See link.

Important Dates and Policy

See syllabus

Any questions or feedback?

Homework 1, Q1: Find your teammates (up to 3 students in a group) and create an account on Kaggle (take a screenshot as below).



Figure 4: