

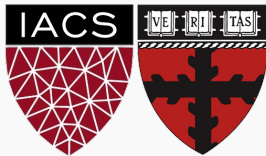
# Lecture #15: A Brief Review

Data Science 1

CS 109A, STAT 121A, AC 209A, E-109A

Pavlos Protopapas    Kevin Rader

Margo Levine    Rahul Dave



## CS 109A: What have we learned?

# Modules

---

The semester has been organized into 3 major modules so far:

# Modules

---

The semester has been organized into 3 major modules so far:

- ▶ Module 0: Intro to Data and Data Science (and Python)
- ▶ Module 1: Transportation Data (and Regression)
- ▶ Module 2: Medical Data (and Classification)

We have learned various approaches to perform both predictions and inferences within each of these frameworks.

# Regression Methods

---

When is it appropriate to perform a regression method?  
What regression models have we learned?

# Regression Methods

---

When is it appropriate to perform a regression method?  
What regression models have we learned?

1. Linear Regression (simple, multiple, polynomial, interactions, model selection, Ridge & Lasso, etc...)
2.  $k$ -NN

What is the main difference between these two types of modules?

# Classification Methods

---

When is it appropriate to perform a classification method? What classification models have we learned?

# Classification Methods

---

When is it appropriate to perform a classification method? What classification models have we learned?

1. Logistic Regression: same details as linear regression apply
2.  $k$ -NN
3. Discriminant Analysis: LDA/QDA
4. Classification Trees

What is the main difference between these two types of models (advantages and disadvantages)? When should you use each method?



## Choosing between Models

---

How can we choose between our various methods/models to answer a question at hand? What approaches/measures can we use to make this determination?

# Choosing between Models

---

How can we choose between our various methods/models to answer a question at hand? What approaches/measures can we use to make this determination?

1. In-sample: AIC, BIC
2. Out-of-sample: Cross-Validation

What measure(s) should we use when we perform cross-validation?

# Dealing with Messy Data

---

What issues have arisen when dealing with real data?  
How have we handled them?

# Dealing with Messy Data

---

What issues have arisen when dealing with real data?  
How have we handled them?

1. Categorical Predictors: might make sense to one-hot encode
2. Missing Data: might make sense to impute
3. High Dimensionality: might make sense to use a data reduction technique.
4. Too many observations: due preliminary analysis on a subset

How are predictions affected? How are inferences affected?

# Dealing with High Dimensionality

---

What does 'high dimensionality' mean? What issues arise when this happens? How can we handle it?

# Dealing with High Dimensionality

---

What does 'high dimensionality' mean? What issues arise when this happens? How can we handle it?

1. Model Selection: subset variable selection
2. Regularization: LASSO and Ridge like approaches (penalize the loss function)
3. PCA: create new predictor variables that encapsulate the 'essence' of all your predictor data with a minimal number of variables.

How can we compare methods to determine which approach is best?

## Other things we've learned

---

- ▶ Scraping, Data Gathering, Data Wrangling
- ▶ EDA: Visualization and Summary Statistics
- ▶ t-tests and  $p$ -values: probabilistic/ approaches to perform inferences
- ▶ Bootstrapping: empirical approach to perform inferences
- ▶ Misclassification Rates, Types of Errors, Confusion Matrices/Tables, and ROC Curves
- ▶ Train vs. Test vs. Classification
- ▶ Standardization vs. Normalization. When should we do it?
- ▶ Anything else?

# The Data Science Process

Don't forget what everything is all about:

