Lecture 24: Wrap-Up

CS109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader

Course Review (and a brief 109B preview)

HW 9 Winners!

Student	Reserve test set Accuracy
Jake Seaton	98.9
Zeeshan Ali	98.8
David Gibson	98.4
Edgardo Hernandez	98.3
Joe Davison	98.3
Georgina Gibson	98.3
Jinsoo Kim	98.1
Pavlos	98.0

Modules

The semester has been organized into 4 major ‘modules’:

- Module 0: Intro to Data and Data Science (and Python)
- Module 1: Regression
- Module 2: Classification
- Module 3: Ensemble Methods

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

Module 1: 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
3. Regression Trees

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

Module 2: 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?

Module 3: Ensemble Methods

What does it mean for a model to be an ensemble method?

1. Bagging Trees
2. Random Forests
3. Boosting Models
4. Stacking Models
5. Neural Networks

What approach does each model take to improve prediction accuracy?

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 Data Issues

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: do 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?

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
- Bias-Variance Trade-off
- Train vs. Test vs. Validation
- Standardization vs. Normalization. When should we do it?

Anything lingering questions or thoughts?

Other things we haven’t discussed

There are lots of topics we have not covered in one semester…some are covered in 109B in the Spring:

- Support Vector Machines (SVMs)
- Unsupervised Classification/Clustering
- Smoothers
- Bayesian Data Analysis
- Reinforcement Learning
- Other versions of Neural Networks (and ‘Deep Learning’)
- Interactive Visualizations
- Database Management
- And much, much more…

CS109B SPRING 2019 SCHEDULE

Date	Lecture #	Topics	Instructor	Lab #	Date	Lab Topic	Assignment	A-sec #	advanced sections topics
1/28	1	Intro + Review of 109A Preview of 109B	PP						
1/30	2	Smoothing and Additive 1/3	MG	**1**	1/31	Setting up enviroment			
2/4	3	Smoothing and Additive 2/3	MG						
2/6	4	Smoothing and GAM 3/3	MG	**2**	2/7	Smoothing/GAM	1 Smoothing (maps HW1 2018)		
2/11	5	Feed Forward + Reg + Review from NN fall	PP						
2/13	6	Optimization of NN (Solvers)	PP	**3**	2/14	**Optimization**	2 **Neural Net 1 (maps to HW5 2018)**	**1**	Optimization/EMD
2/18	7	AWS scalable systems	RD						
2/20	8	SQL	RD	**4**	2/21	**Setting UP AWS**		**2**	Dropout +
2/25	9	CNNs-1	PP						
2/27	10	CNNs-2	PP	**5**	2/28	**CNNs**	3 **CNNs (maps to HW6 Q1+Q2)**	**3**	ConvNets: LeNet, AlexNet, VGG-15, ResNet and Inception +
3/4	11	RNN 1	PP						
3/6	12	RNN 2	PP	**6**	3/7	**RNNS**	4 **RNNs (maps to HW6 Q3)**	**4**	LSTN, GRU in NLP +
3/11	13	Unsupervised learning/clustering 1	MG						
3/13	14	Unsupervised learning/clustering 2	MG	**7**	3/14	Clusterig	5 Unsup + AE (maps to HW2 2018)	**5**	Neural style transfer learning +
3/25	15	Autoencoders	PP						
3/27	16	Bayesian 1/3	MG	**8**	3/28	Bayes 1		**6**	Deep RL +
4/1	17	Bayesian 2/3	MG						
4/3	18	Bayesian 3/3	MG	**9**	4/4	Bayes 2	6 LDA and Bayes (maps to HW3 2018)		
4/8	19	Generative Models Varational Autoenders 1	PP						
4/10	20	Generative Models Varational Autoenders 2 and GANS	PP	**10**	4/11	**VAE**	7 **VAE+GANS**	**7**	Variational Inference +
4/15	21	GANS 1	PP						
4/17	22	GANS 2	PP	**11**	4/18	**GANS**		**8**	GANS
4/22	23	MODULE: LECTURE DOMAIN	MI						
4/24	24	MODULE: PROBLEM BACKGROUND	MI						
4/29	25	MODULE: PROBLEM BACKGROUND	MI						
5/1	26	PROJECT WORK							
5/6		PROJECT WORK							
5/8		PROJECT WORK							
5/13									
5/16		FINAL PRESENTATIONS AND SHOWCASE							

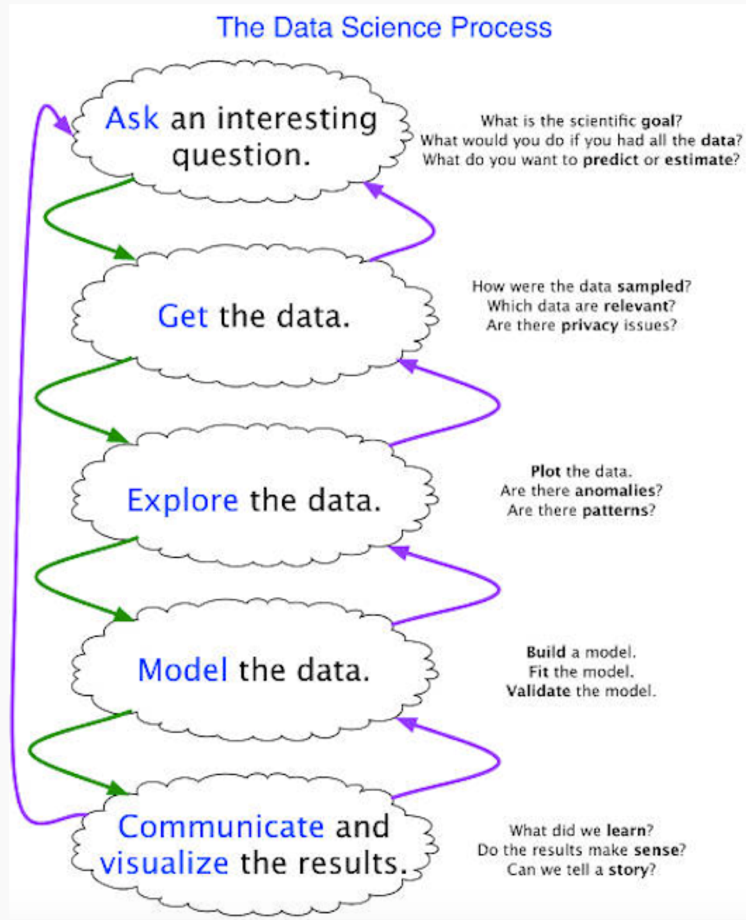
Courses Related to Data Science

- CS 109B: Advanced Topics in Data Science
- CS 171: Visualizations
- CS 181: Machine Learning
- CS 182: Artificial Intelligence (AI)
- CS 205: Distributive Computing
- Stat 110: Probability Theory
- Stat 111: Statistical Inference
- Stat 139: Linear Models
- Stat 149: Generalized Linear Models
- Stat 195: Intro to Statistical Machine Learning

This list is not exhaustive!

The Data Science Process

Don't forget what everything is all about:



Thanks for all your hard work!

It's been a long semester for everyone involved. Thank you for your patience, your hard work, and your commitment to data science!

It's sad to see you go...