BST 219
Core Principles of Data Science

Lecture 30: Machine Learning continued, Data Science next steps
December 19, 2024

Recipe of the Day!

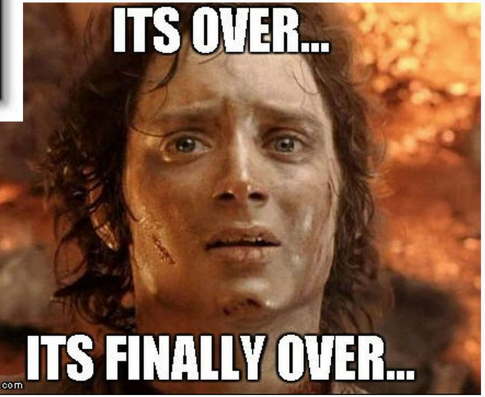
Holiday Cocktails

Holiday Mocktails

Agenda

- Announcements
  - No lab this week!
  - No office hour on Thursday
  - Please complete the course evaluation - we value your feedback!
- Continue Machine Learning module
  - Regularization
    - Nice LASSO tutorial video
  - Cross-validation explanation
- Next steps in Data Science

HOW YOU WALK OUT OF YOUR LAST CLASS
AT THE END OF THE SEMESTER
memegenerator.net



Course Highlight Reel

What did we learn?

Course Roadmap

Importing (loading) the data

Processing (cleaning, wrangling) the data

Visualizing and summarizing the data

Building models (statistical and ML)

Interpretation and communication of results

Topics we covered

Module 1 R Basics	Module 2 Git and GitHub	Module 3 Data Visualization	Module 4 Advanced Data Wrangling	Module 5 Machine Learning
-R Markdown -Data types -Vectors -Sorting -Vector arithmetic -Indexing -Data wrangling -Importing data -Functions -For loops -If else statements	-Cloning repositories via RStudio -Committing and pushing assignments via RStudio -Pulling course notes via RStudio	-ggplot2 -Data visualization principles -Faceting -Fixing scales -Time series plots -Transformations -Ordering by a value -Making comparisons -Maps	-Tidy data format -Importing data -Reshaping data -Join functions and combining tables -Dates and times -TableOne	-Fundamentals of ML including train/test split and the process -Logistic regression, Naive Bayes, kNN, QDA, LDA, Decision trees, Random Forests, LASSO, Ridge regression, PCA -ROC and AUROC -Confusion matrix -Performance metrics

Data Science Next Steps

Spring Courses

BST 263 - Statistical Learning (Full Spring)	BST 261 - Data Science II (Deep Learning, Spring 2)	BST 221 - Applied Data Structures and Algorithms (Full Spring)
Language: R or Python ● Material: An Introduction to Statistical Learning ● Additional Reading: The Elements of Statistical Learning Probability Basics Assessing Model Accuracy Linear and Logistic Regression Classification Cross-validation and bootstrap Subset selection Penalty-based methods Ridge regression and LASSO Dimension Reduction Polynomial Regression Step functions and Basis functions Generalized Additive Models (GAMs) Classification and Regression Trees (CART) Bagging and Random Forests Ensembles Support Vector Machines Unsupervised Learning Bayesian Methods	Language: Python ● Material: Deep Learning, Deep Learning with Python Brief review of Python 3 Brief review of Linear Algebra and Probability Brief review of Machine Learning Feedforward Neural Networks Convolutional Neural Networks Recurrent Neural Networks Generative Adversarial Networks (GANs) Reinforcement Learning Transfer Learning Hyperparameter Tuning Model Selection	Language: Python Overview of numerical analysis material Concepts of Algorithms, Complexity and Sorting Algorithms Data Structures and Heapsort Parallel Programming Greedy Algorithms and Dynamic Programming Numerical Stability (Pseudo) Random Number Generation Efficient Algorithms for Linear Algebra Least-Squares Problem, Eigenvalue Decompositions Algorithms for Numerical Integration, MC-Integration, Importance Sampling Graphs and Network Algorithms

Other Courses

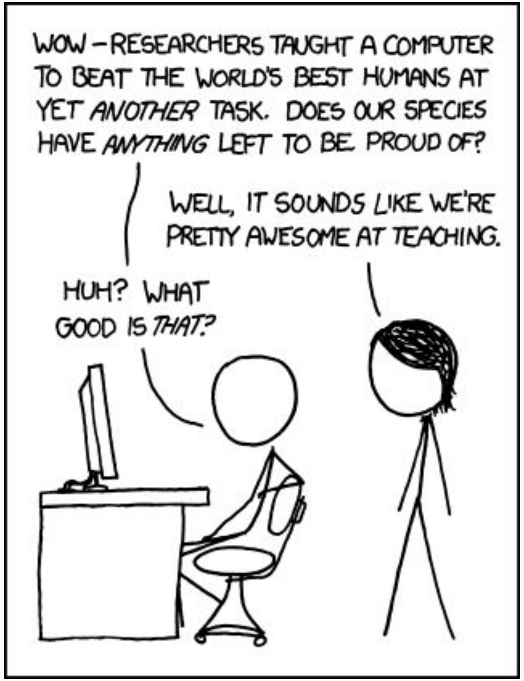
- BST 262: Computing for Big Data (only offered in Fall 2)
- BST 267: Introduction to Social and Biological Networks (only offered in Fall 2)
- APMTH 120: Applied Linear Algebra and Big Data (offered in Spring, main campus)
- Other machine learning and NLP courses at Harvard and MIT
    - Caution: these tend to be VERY difficult and a ton of work

Skills Development

- Keep using R for research or other projects!
- Learn Linear and Matrix Algebra
- Diversify your Programming Languages (Python, SQL, C++, Java, etc.)
- Kaggle competitions
- Online courses on Coursera, edX, Data Camp, etc.
- Data Science Meetups in Boston
- Data science podcasts and social media
- Data science blogs
- Data science tutorials and videos

Become a TF!

We are always in need of teaching fellows for our Biostats courses, including this one!


WOW – RESEARCHERS TAUGHT A COMPUTER TO BEAT THE WORLD'S BEST HUMANS AT YET *ANOTHER* TASK. DOES OUR SPECIES HAVE *ANYTHING* LEFT TO BE PROUD OF?
WELL, IT SOUNDS LIKE WE'RE PRETTY AWESOME AT TEACHING.
HUH? WHAT GOOD IS *THAT?*