WEEK 3

CLASS 5: LINEAR REGRESSION

HOMEWORK:

- Watch Rahul Patwari's videos on probability (5 minutes) and odds (8 minutes) if you're not familiar with either of those terms.
- Read these excellent articles from BetterExplained:
  - An Intuitive Guide To Exponential Functions & e
  - Demystifying the Natural Logarithm (ln).

RESOURCES:

- Setosa has an excellent interactive visualization of linear regression.
- To go much more in-depth on linear regression, read Chapter 3 of An Introduction to Statistical Learning, from which this lesson was adapted. Alternatively, watch the related videos .
- To learn more about Statsmodels and how to interpret the output, DataRobot has some decent posts on simple linear regression and multiple linear regression.
- This introduction to linear regression is much more detailed and mathematically thorough, and includes lots of good advice.
- This is a relatively quick post on the assumptions of linear regression.
- John Rauser's talk on Statistics without the Agonizing Pain (12 minutes) gives a great explanation of how the null hypothesis is rejected.
- A major scientific journal recently banned the use of p-values: Scientific American has a nice summary of the ban.
- This response to the ban in Nature argues that "decisions that are made earlier in data analysis have a much greater impact on results".
- Andrew Gelman has a readable paper in which he argues that "it's easy to find a p < .05 comparison even if nothing is going on, if you look hard enough".

CLASS 5: LOGISTIC REGRESSION

HOMEWORK:

- If you aren't yet comfortable with all of the confusion matrix terminology, watch Rahul Patwari's videos on Intuitive sensitivity and specificity (9 minutes) and The tradeoff between sensitivity and specificity (13 minutes).
- Exercise with Titanic data

RESOURCES:

- To go deeper into logistic regression, read the first three sections of Chapter 4 of An Introduction to Statistical Learning, or watch the first three videos (30 minutes) from that chapter.
- For a math-ier explanation of logistic regression, watch the first seven videos (71 minutes) from week 3 of Andrew Ng's machine learning course, or read the related lecture notes compiled by a student.
- For more on interpreting logistic regression coefficients, read this excellent guide by UCLA's IDRE and these lecture notes from the University of New Mexico.
- The scikit-learn documentation has a nice explanation of what it means for a predicted probability to be calibrated.
- Supervised learning superstitions cheat sheet is a very nice comparison of four classifiers we cover in the course (logistic regression, decision trees, KNN, Naive Bayes) and one classifier we do not cover (Support Vector Machines).

CLASS 6: ADVANCED MODEL EVALUATION

HOMEWORK:

- Yelp reviews

RESOURCES:

- Rahul Patwari has a great video on ROC Curves (12 minutes).
- An introduction to ROC analysis is a very readable paper on the topic.
- These lesson notes from a course at the University of Georgia include some simple, real-world examples of the use of ROC curves.
- ROC curves can be used across a wide variety of applications, such as comparing different feature sets for detecting fraudulent Skype users, and comparing different classifiers on a number of popular datasets.
- This blog post about Amazon Machine Learning contains a neat graphic showing how classification threshold affects different evaluation metrics.
- scikit-learn has extensive documentation on model evaluation.
- Section 3.3.1 of An Introduction to Statistical Learning (4 pages) has a great explanation of dummy encoding for categorical features.
- Azure ML Machine Learning Algorithm Choice
- Choosing a Machine Learning Classifier
- Machine learning done wrong
- Practical Machine Learning tricks from the KDD 2011
- Evaluating Machine Learning Models (Alice Zheng, Download O'Reilly EBOOK)
- Scikit-learn Machine learning Map

CLASS 6: WEB SCRAPING

HOMEWORK:

- Read Jeff Leek's guide to creating a reproducible analysis, and watch this related Colbert Report video (8 minutes).

RESOURCES:

- The Beautiful Soup documentation is incredibly thorough, but is hard to use as a reference guide. However, the section on specifying a parser may be helpful if Beautiful Soup appears to be parsing a page incorrectly.
- For more Beautiful Soup examples and tutorials,
  - See Web Scraping 101 with Python,
  - Alex's well-commented notebook on scraping Craigslist
  - This notebook from Stanford's Text As Data course
  - This notebook and associated video from Harvard's Data Science course.
- For a much longer web scraping tutorial covering Beautiful Soup, lxml, XPath, and Selenium, watch Web Scraping with Python (3 hours 23 minutes) from PyCon 2014. The slides and code are also available.
- For more complex web scraping projects, Scrapy is a popular application framework that works with Python. It has excellent documentation, and here's a tutorial with detailed slides and code.
- robotstxt.org has a concise explanation of how to write (and read) the robots.txt file.
- import.io and Kimono claim to allow you to scrape websites without writing any code.
- How a Math Genius Hacked OkCupid to Find True Love and How Netflix Reverse Engineered Hollywood are two fun examples of how web scraping has been used to build interesting datasets.