The University of Texas at Austin
Department of Electrical and Computer Engineering

460J: Data Science Lab — Fall 2021

Lab Three

Caramanis/Dimakis Due: Sunday Sept. 19th Midnight, 2021.

Problem 0: Optional.
The ISL book ('An Introduction to Statistical Learning' by G. James et al.) is a good place to read what we have covered and additional material. The book is available online http://faculty.marshall.usc.edu/gareth-james/ISL/ Study chapters 3 and 4.
(Unfortunately the examples are written in R but we use python in our course)

Problem 1 (A bit of Information Theory)
Read Shannon's 1948 paper 'A Mathematical Theory of Communication'. Focus on pages 1-19 (up to Part II), the remaining part is more relevant for communication.
http://math.harvard.edu/ ctm/home/text/others/shannon/entropy/entropy.pdf
Summarize what you learned briefly (e.g. half a page).

Problem 2: Scraping, Entropy and ICML papers.

ICML is a top research conference in Machine learning. Scrape all the pdfs of all ICML 2017 papers from http://proceedings.mlr.press/v70/.

1. What are the top 10 common words in the ICML papers?

2. Let Z be a randomly selected word in a randomly selected ICML paper. Estimate the entropy of Z.

3. Synthesize a random paragraph using the marginal distribution over words.

4. (Extra credit) Synthesize a random paragraph using an n-gram model on words. Synthesize a random paragraph using any model you want. Top five synthesized text paragraphs win bonus (+30 points).

Problem 3: More on Kaggle Advanced Regression.
Continue building your toolbox on Kaggle. Work on submissions for the same competition https://www.kaggle.com/c/house-prices-advanced-regression-techniques/

1. What is the best Kaggle forum post that you found? Briefly describe what you learned from it.

2. What is the best public leader board (LB) score you can achieve? Describe your approach.

3. Submit a model that is definitely overfitting and a model that is definitely underfitting. Overfitting means that your training error is much smaller compared to your test error (and

LB score). Underfitting means that your model is too simple and even the training error is very large (and so will the test error). You can experiment with depth of decision trees in random forests or XGBoost classifiers as the metric of complexity for your models, or any other family of models you want.