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

EE380L: Data Mining — Spring 2019

Problem Set Three

C. Caramanis Due: Tuesday April 9th, 11:59pm 2019.

This is a short problem set. I have also posted the midterm from the last time I offered this course (2017). The midterm will be on Saturday, and will have only one part (last time there was a part I and a part II).

Problem 1
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 2018 papers from `http://proceedings.mlr.press/v80/`.

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!