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

EE379K: Data Science Lab — Fall 2017

Lab Three

Caramanis/Dimakis Due: Monday, Sept. 25, 3:00pm 2017.

Problem 1: Linear Algebra in Python. You can use all Python functions to solve this problem.

1. Consider the linear subspace $S = span\{v_1, v_2, v_3, v_4\}$ where $v1 = [1, 2, 3, 4], v2 = [0, 1, 0, 1], v3 = [1, 4, 3, 6], v4 = [2, 11, 6, 15]$. Create a vector inside S different from v_1, v_2, v_3, v_4. Create a vector not in S. How would you check if a new vector is in S?
2. Find the dimension of the subspace S.
3. Find an orthonormal basis for the subspace S.
4. Solve the optimization problem $min_{x \in S}||x - z^*||_2$ where $z^* = [1, 0, 0, 0]$.

Problem 2: Scraping, Entropy and ICML papers.

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. (Bonus) Synthesize a random paragraph using an n-gram model on words. Synthesize a random paragraph using any model you want. Best three paragraphs win bonus (+50 Lab credit)

Problem 3: Starting in Kaggle.

1. Lets start with our first Kaggle submission in a playground regression competition. Make an account to Kaggle and find https://www.kaggle.com/c/house-prices-advanced-regression-techniques/
2. Follow the data preprocessing steps from https://www.kaggle.com/apapiu/house-prices-advanced-regression-techniques/regularized-linear-models. Then run a ridge regression using $\alpha = 0.1$. Make a submission of this prediction, what is the RMSE you get?
(Hint: remember to exponentiate np.expm1(ypred) your predictions).
3. Try to get to build the best model you can. Report the best RMSE you got on the Kaggle wall and how you got it.