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

460J: Data Science Lab — Fall 2020

Lab Three

Caramanis/Dimakis Due: Friday Sept. 25th, 2020.

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
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: Starting in Kaggle.
Soon you will be participating in the in-class Kaggle competition made for this class. In that one, you will be participating on your own. This is a warmup- the more effort and research you put into this assignment the easier it will be to compete into the real Kaggle competition that you will need to do soon. We expect you to spend 10 times more effort on this problem compared to the others.

1. Let's 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. Compare a ridge regression and a lasso regression model. Optimize the alphas using cross validation. What is the best score you can get from a single ridge regression model and from a single lasso model?

4. Plot the l_0 norm (number of nonzeros) of the coefficients that lasso produces as you vary the strength of regularization parameter alpha.

5. Add the outputs of your models as features and train a ridge regression on all the features plus the model outputs (This is called Ensembling and Stacking). Be careful not to overfit. What score can you get? (We will be discussing ensembling more, later in the class, but you can start playing with it now).

6. Install XGBoost (Gradient Boosting) and train a gradient boosting regression. What score can you get just from a single XGB? (you will need to optimize over its parameters). We will discuss boosting and gradient boosting in more detail later. XGB is a great friend to all good Kagglers!

7. Do your best to get the more accurate model. Try feature engineering and stacking many models. You are allowed to use any public tool in python. No non-python tools allowed.

8. (Optional) Read the Kaggle forums, tutorials and Kernels in this competition. This is an excellent way to learn. Include in your report if you find something in the forums you like, or if you made your own post or code post, especially if other Kagglers liked or used it afterwards.

9. Be sure to read and learn the rules of Kaggle! No sharing of code or data outside the Kaggle forums. Every student should have their own individual Kaggle account and teams can be formed in the Kaggle submissions with partners. This is more important for live competitions of course.

10. As in the real in-class Kaggle competition (which will be next), you will be graded based on your public score (include that in your report) and also on the creativity of your solution. In your report (**that you will submit as a pdf file**), explain what worked and what did not work. Many creative things will not work, but you will get partial credit for developing them. We will invite teams with interesting solutions to present them in class.