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

MIS382N: Business Data Science Lab — Fall 2019

Problem Set Four

Caramanis/Dimakis Due: Thursday October 10th, 2019.

Problem 1 – Warm up. Grid Search CV.

1. Run this simple example from scikit learn, and understand what each command is doing: `https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html`

Problem 2 – Lasso, Forward Selection and Cross Validation.
Use the data generation used in the Lecture 7 notebook, where we first introduced Lasso, to generate data.

1. Manually implement forward selection. Report the order in which you add features.
2. Plot test error as a function of the size of the support. Can you use this to recover the true support?
3. Use Lasso with a manually implemented Cross validation using the metric of your choice. What is the value of the hyperparameter? (Manually implemented means that you can either do it entirely on your own, or you can use GridSearchCV, but I'm asking you not to use LassoCV, which you will use in the next problem).
4. Change the number of folds in your CV and repeat the previous step. How does the optimal value of the hyperparameter change? Try to explain any trends that you find.
5. Read about and use LassoCV from sklearn.linear_model. How does this compare with what you did in the previous step? If they agree, then explain why they agree, and if they disagree explain why. This will require you to make sure you understand what LassoCV is doing.

Problem 3
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 4: Scraping, Entropy and ICML papers.

ICML – the International Conference on Machine Learning – is a top research conference in Machine learning. Scrape all the pdfs of all ICML 2019 papers from `http://proceedings.mlr.press/v97/`.

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. (Optional) 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!

Problem 5: Logistic Regression.

The following is a logistic regression problem using a real data set, made available by the authors of the book "Applied Regression and Muiltilevel Modeling" by Gelman and Hill.

Download the data from the book, which you can find here `http://www.stat.columbia.edu/~gelman/arm/software/`. In particular, we are interested in the `arsenic` data set. The file `wells.dat` contains data on 3,020 households in Bangladesh. For each family, the natural arsenic level of each well was measured. In addition, the distance to the nearest safest well was measured. Each family is also described by a feature that relates to their community involvement, and a feature that gives the education level of the head of household. We are interested in building a model that predicts whether the family decided to switch wells or not, based on being informed of the level of arsenic in the well. Thus the "label" for this problem is the binary vector that is the first column of the dataset, labeled "switch."

- Fit a logistic regression model using only an offset term and the distance to the nearest safe well.

- Plot your answer: that is, plot the probability of switching wells as a function of the distance to the nearest safe well.

- Interpreting logistic regression coefficients: Use the "rule-of-4" discussed in class on Thursday, to interpret the solution: what can you say about the change in the probability of switching wells, for every additional 100 meters of distance?

- Now solve a logistic regression incorporating the constant term, the distance and also arsenic levels. Report the coefficients

- Next we want to answer the question of which factor is more significant, distance, or arsenic levels? This is not a well specified question, since these two features have different units. One natural choice is to ask if after normalizing by the respective standard deviations of each feature, if moving one unit in one (normalized) feature predicts a larger change in probability of switching wells, than moving one unit in the other (also normalized) feature. Use this reasoning to answer the question.

- Now consider all the features in the data set. Also consider adding interaction terms among all features that have a large main effect. Use cross validation to build the best model you can (using your training set only), and then report the test error of your best model.[1]

- (Optional) Now also play around with ℓ_1 and ℓ_2 regularization, and try to build the most accurate model you can (accuracy computed on the test data).

[1]Note that since you have essentially unlimited access to your test set, this opens the door for massive overfitting. In contrast, Kaggle competitions try to mollify this by giving you only limited access to the test set.