Problem Sets 5

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: There are two deadlines for this problem set. The usual deadline for the document is November 26 at 1:00 pm. The deadline for predictions is November 25 at 8:00 pm.

The repo link to create your submission is
https://classroom.github.com/a/MVelV6xy

1 Theory Exercises: Boosting and Text as Data

1. AdaBoost. In one of the last steps AdaBoost updates the following error function:

$$L = \sum_{i=1}^{N} w_i^{(m)} exp\left(-\frac{1}{2}\alpha_m y_i G_m(x_i)\right) \tag{1}$$

show that if we differentiate that function with respect to α_m, then ADABoost are updating using $\alpha_m = log(\frac{1-err_m}{err_m})$.

2. Consider a model where we count the number of words in a document. These words come from a fixed dictionary. We use a multinomial distribution to model this process

$$f(X|\theta) = N!\Pi_{k=1}^{K}\frac{\theta_k^{N_k}}{N_k!} \tag{2}$$

where θ is the probability of $k-th$ word, N_k is the number of times that word appears, K is the number of words in that dictionary, and $N = \sum_k N_k$, the sum of occurrences of all words.

   (a) Adding the restriction that all probabilities must add to one ($\sum_k \theta_k = 1$), derive the maximum likelihood estimator of θ

   (b) Let's assume that the prior of θ follows a Dirichlet distribution ($\theta \sim Dir(\theta|\alpha_1, \ldots, \alpha_K)$) where:

$$Dir(\theta|\alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)}\Pi_{k=1}^{K}\theta_k^{\alpha_k - 1} \tag{3}$$

where $B(\alpha_1, \ldots, \alpha_K) = \frac{\pi_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$. Find the posterior distribution. (Hint: Dirichlet distribution is a conjugate prior for the Multinomial distribution)

(c) Find the posterior mode by maximizing the posterior likelihood function with respect to θ_k, imposing the constrain $\sum_k \theta_k = 1$

(d) Show that the above result under a uniform prior ($\alpha_k = 1$), we get back the MLE estimator.

(e) Intuitively, what does it mean to impose a Uniform prior?

3. PCA. In class, I showed that the solution for the first component is $\delta_1' \Sigma \delta_1 = \lambda_1$. Now you have to find the second component, in a way that we sill want to further minimize the variance, but it has the following constrains $\delta_2' \delta_2 = 1$ and $\delta_2' \delta_1 = 0$.

(a) Intuitively explain the maximization problem and what each constrain means.

(b) Show that the value of δ_2 that minimizes the above problem is given by the eigenvector of Σ with the largest eigenvalue.

(c) Now let $\mathcal{L}(\delta_2, f_2) = \frac{1}{n} \sum_{i=1}^{N} (x_i - f_{i1}\delta_1 - f_{i2}\delta_2)'(x_i - f_{i1}\delta_1 - f_{i2}\delta_2)$. Show that $\frac{\partial \mathcal{L}}{\partial \delta_2} = 0$ implies that $f_{i2} = \delta_2' x_i$

2 Empirical Problem

The main objective of this section is to apply the concepts we learned using "real" world data. With these, I also expect that you sharpen your data collection and wrangling skills. Finally, you should pay attention to your writing.

I encourage you to turn the following section of the problem set in a way that resembles a paper. As such, I expect graphs, tables, and writing to be as neat as possible. You can write it in Spanish or English, and either language is acceptable. For students in the Ph.D., it would be a good practice to do it in English.

Don't forget to upload everything to your repository and follow the template repository.

2.1 *"A rose by any other name would smell as sweet"* Juliet Capulet

There is an adage that says, *"choose your words carefully."* Words themselves may reveal far more than what we're trying to say. There's mounting evidence that our written words show who we are.

The objective today is to predict to whom each tweet belongs. The training dataset contains around 7,000 tweets of four prominent Colombian politicians' accounts: Claudia Lopez, Gustavo Petro, Alvaro Uribe, y Alejandro Gaviria. The test dataset contains 500 unlabeled tweets. We want to predict which account posted the tweets in the test set. All the relevant data sets are available in the data_tweets folder, available here.

The expected output of this section are a document and a `.csv` file of predictions.

1. The document should:
   (a) Describe how you prepossessed the data.
   (b) Present a descriptive/explanatory analysis of the data.
   (c) Describe in detail the model you used for your final prediction, for example, explain the hyper-parameters that you chose, how you got there, etc. You can use any model that you want. The only restriction is that you have to use at least one model from the following classes: (1) bag-of-words representation, (2) language modelling (word2vec, BERT, etc.). You can find a Spanish corpus to use with word2vec here, and the Spanish version of BERT, BETO, here.
   (d) Show the performance of that model relative to other models. Don't forget to mention what is your performance metric and argue why it was the model that you selected.
   (e) Besides writing up your results, you should submit a `.csv`. The submission file format is the variable `id`, and a `name` variable that indicates whether the tweet belongs to: Lopez, Gaviria, Petro, or Uribe.
2. Your grade will depend on your relative prediction accuracy. The best model will receive top marks, the others will be scaled accordingly.