IEOR 242: Applications in Data Analysis, Fall 2021

Homework Assignment #3

October 7, 2021

Problem 0:

a) Please self-grade Problems 1 and 2 from Homework 2. Please use the following rubric to assign yourself a score of $0 - 2$ for each part of the two problems:

   (a) 0 points = Did not attempt or very wrong

   (b) 1 point = Showed something non-trivial, but some mistakes or some missing justification

   (c) 2 points = 100% correct

   The GSIs will make an announcement concerning how to submit your self-grade scores on Gradescope.

b) Please submit your project proposal on Gradescope. (This is not part of the assignment, but simply a reminder.)

Problem 1: (30 points)

Consider the algorithm for building a CART model in the case of regression. Following and expanding on the notation from class, suppose that our current tree, denoted by T_{old}, has $|T_{\text{old}}| = M$ terminal nodes/buckets. For each bucket $m = 1, \ldots, M$, let:

1. N_m denote the number of observations in bucket m ,
2. $Q_m(T_{\text{old}})$ denote the value of the impurity function at bucket m , and
3. R_m denote the region in the feature space corresponding to bucket m .

Also let N be the overall total number of observations. Recall that, in the case of regression we have that:

$$Q_m(T_{\text{old}}) = \frac{1}{N_m} \sum_{i : x_i \in R_m} (y_i - \hat{y}_m)^2 \ ,$$

where $\hat{y}_m = \frac{1}{N_m} \sum_{i : x_i \in R_m} y_i$ is the mean response in bucket m.

Then the total impurity cost of the tree T_{old} is defined as:

$$C_{\text{imp}}(T_{\text{old}}) = \sum_{m=1}^{M} N_m Q_m(T_{\text{old}}) \ .$$

Consider a potential split at the final bucket M (we're using M just for ease of notation), which results in a new tree T_{new}. This new tree has $|T_{\text{new}}| = M + 1$ terminal nodes/buckets, and for this new tree we let

1. $\tilde{N}_m$ denote the number of observations in bucket m ,
2. $\tilde{Q}_m(T_{\text{new}})$ denote the value of the impurity function at bucket m , and
3. $\tilde{R}_m$ denote the region in the feature space corresponding to bucket m .

The total impurity cost of the tree T_{new} is defined analogously as:

$$C_{\text{imp}}(T_{\text{new}}) = \sum_{m=1}^{M+1} \tilde{N}_m \tilde{Q}_m(T_{\text{new}}) \ .$$

Please answer the following:

a) (10 points) Let $\Delta = C_{\text{imp}}(T_{\text{old}}) - C_{\text{imp}}(T_{\text{new}})$ be the absolute decrease in total impurity resulting from the split. Derive a formula for Δ that can be computed locally at the bucket M, in other words it should only depend on the data points that fall in region R_M in the original tree T_{old}. (Hint: we've discussed this concept in class, this question is asking for a more formal argument. You may assume that the two new buckets in T_{new} resulting from the split are labeled as buckets M and $M+1$ in T_{new}.)

b) (10 points) Show that $\Delta \geq 0$, hence splitting always reduces the total impurity cost. (Hint: you can use the fact that, given a sequence of real numbers $z_1, z_2, \ldots, z_n$, the mean $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ is the minimizer of the function $\text{RSS}(z) = \sum_{i=1}^{n}(z_i - z)^2$)

c) (10 points) Let R^2_{old} be the training set R^2 value for the model defined by T_{old}, and likewise let R^2_{new} be the training set R^2 value for the model defined by T_{new}. Let $\text{SST} = \sum_{i=1}^{N}(y_i - \bar{y})^2$ be the total sum of squared errors, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the overall mean. For a given value of the complexity parameter (cp) $\alpha \geq 0$, recall the modified cost function that is relevant in the pruning step:

$$C_\alpha(T) = C_{\text{imp}}(T) \ + \ \alpha \cdot \text{SST} \cdot |T|$$

Show that $C_\alpha(T_{\text{new}}) \leq C_\alpha(T_{\text{old}})$ if and only if $R^2_{\text{new}} - R^2_{\text{old}} \geq \alpha$. (Hence the choice of retaining a split if the increase in R^2 is at least α is equivalent to retaining a split if the modified cost function is smaller after the split.)

Problem 2: Predicting Yelp Ratings (70 points)

Yelp is a widely popular platform that publishes information and reviews of local businesses such as restaurants, plumbers, hair salons, and others. Any user of Yelp is able to write a review, and each review includes a star rating between 1 and 5 in addition to written comments. In this problem, you will build models for predicting the star ratings of restaurants in Las Vegas, Nevada based on attributes contained in their Yelp profiles. Such a model may be useful for businesses to understand which factors are most important in attaining a high star rating and in gaining popularity more generally.

The data for this problem is contained in the files `yelp242_train.csv` and `yelp242_test.csv`, and was retrieved from the larger Yelp Dataset[1] provided by Yelp. In particular, we will focus our analysis on a subset of the Yelp data concerning restaurants in the Las Vegas, Nevada area. We have performed a random 70/30 split, resulting in 6,272 observations in the training set and 2,688 observations in the test set. Each observation contains the average star rating, number of reviews, and a list of attributes collected from the Yelp page of a particular restaurant in the Las Vegas area. These attributes are described in Table 1. Note that **variable selection is not required** for this problem.

NOTE: Whenever a question asks you to perform some coding task, such as building a model, please include an explanation of what the code does (e.g., "Below is the code for building a linear regression model") as well as the supporting code.

a) (5 points) There are many missing entries in this dataset, denoted by `(Missing)` in the data files. In particular, all of the attribute features contain missing values and Table 1 reports the percentage of observations where each attribute is missing. In general, there are several approaches for dealing with missing values in supervised learning. Each attribute with missing values in our dataset is a categorical feature and, in the subsequent models that you will build, you should treat `(Missing)` as an explicit category. Do you think this modeling approach is reasonable or not? Explain your answer.

b) (15 points) Let us start by building regression models for predicting `stars` based on all of the provided features listed in Table 1. All of your models should, of course, be built only using the training data provided in the `yelp242_train.csv` file.

i) First build a linear regression model. Remember to use all of the provided independent variables, and you do not have to do variable selection in this problem. For each of the categorical variables, you should use `(Missing)` as the reference level to be incorporated into the intercept term. This does not affect the predictive performance of the model, but it does lead to a cleaner interpretation. This can be achieved in statsmodels with a slight modification to the R-style formulas. For example, you could use code like

```
"stars ~ review_count + C(GoodForKids, Treatment(reference='(Missing)'))"
```

for a model regressing `stars` on `review_count` and `GoodForKids`.[2]

ii) Now build a regression tree model (using an implementation of the CART algorithm). Select the complexity parameter (i.e., `ccp_alpha` in sklearn) value for the tree through

[1] `https://www.yelp.com/dataset`

[2] More details may be found at
`https://stackoverflow.com/questions/22431503/specifying-which-category-to-treat-as-the-base-with-statsmodels`

Table 1: Variables in the dataset `yelp242`.

Variable	**Description**	**Levels**	**Missing rate**
`stars`	The average star rating of the business (from 1 to 5).		0.0%
`review_count`	The number of reviews received by the business.		0.0%
`GoodForKids`	Whether this business is good for kids.	T, F, (Missing)	30.19%
`Alcohol`	The kind of alcohol provided at this business.	Beer_and_wine, full_bar, none, (Missing)	34.43%
`BusinessAcceptsCreditCards`	Whether the business accepts credit cards.	T, F, (Missing)	6.15%
`WiFi`	Whether the business provides WiFi.	free, no, paid, (Missing)	32.82%
`BikeParking`	Whether bike parking is available at the business.	T, F, (Missing)	29.39%
`ByAppointmentOnly`	Whether the business is by appointment only.	T, F, (Missing)	87.86%
`WheelechairAccessible`	Whether the business is wheelechair accessible.	T, F, (Missing)	74.43%
`OutdoorSeating`	Whether the business provides outdoor seating.	T, F, (Missing)	27.69%
`RestaurantsReservations`	Whether the business takes any reservation.	T, F, (Missing)	31.60%
`DogsAllowed`	Whether the business allows dogs.	T, F, (Missing)	72.65%
`Caters`	Whether the business provides catering.	T, F, (Missing)	34.12%

cross-validation, and explain how you did the cross-validation and how you selected the complexity parameter value.

iii) Using the test set data provided in the `yelp242_test.csv` file, compute the OSR^2 values of your linear regression and regression tree models. Also, compute the MAE (mean absolute error) values of both models. How do you judge the performance of the two models?

c) (5 points) Regression may not be the most appropriate modeling technique for this data. In particular, it is plausible that restaurants may be mostly concerned with ensuring that their star rating is high enough and not particularly concerned with precisely predicting the value of this rating. Therefore, let us instead consider a classification problem where the goal is to predict if the star rating is greater than or equal to 4 or not.

i) Construct a new variable in your training and test datasets called **`fourOrAbove`**. This variable should be equal to 1 if **`stars`** is greater than or equal to 4 and equal to 0 otherwise.

d) (30 points) Let us now work on building classification models for predicting **`fourOrAbove`** based on all of the provided features listed in Table 1. All of your models should, of course, be built only using the training data provided in the **`yelp242_train.csv`** file.

i) In this problem, we will weigh false positives and false negatives equally and therefore focus on accuracy (equivalently, error rate) as the primary performance metric. Do you think this modeling choice is reasonable or not? Explain your answer.

ii) A simple approach here is to use the previously built linear regression and regression tree models to address the classification task by thresholding their predictions at the value of 4. Write code for implementing this thresholding procedure for converting the predictions of your previously built regression models to predictions of **`fourOrAbove`**.

iii) Now build a logistic regression model. Remember to use all of the provided independent variables, and you do not have to do variable selection in this problem. Please again use **`(Missing)`** as the reference level to be incorporated into the intercept term.

iv) Now build a classification tree model (using an implementation of the CART algorithm). Select the complexity parameter (i.e., **`ccp_alpha`** in sklearn) value for the tree through cross-validation, and explain how you did the cross-validation and how you selected the complexity parameter value.

v) Produce a table for evaluating the performance of the models that you have built. The table should consider the following models: a baseline model that predicts the most frequent outcome for **`fourOrAbove`**, the two regression models with thresholding, and all of the classification models built in this part of the problem. Additionally, the table should report the following performance metrics: accuracy (our primary performance metric), TPR, and FPR. All of these performance metrics should be computed using the test set data provided in the **`yelp242_test.csv`** file. How do you judge the performance of the models? Do the results seem reasonable to you? Which model would you recommend for this problem and why?

e) (15 points) Suppose that you are a data scientist working for Yelp and you have been tasked with producing a “how to guide” for Las Vegas restaurants, which is supposed to include tips for actions restaurants may take to achieve a high star rating. Use the provided data to construct a list of three such tips. Each of your tips should be justified by the data, and each justification should be understandable by a restaurant owner who may not know much about machine learning. This question is purposefully open ended, and you may consider using modeling techniques and/or visualizations to come up with and justify your three tips.