UD120: Intro to Machine Learning Fall 2016

Project: Free Response Questions

Kevin Mather

1 Overview

A critical part of machine learning is making sense of your analysis process and communicating it to others. The questions below will help us understand your decision-making process and allow us to give feedback on your project. Please answer each question; your answers should be about 1-2 paragraphs per question. If you find yourself writing much more than that, take a step back and see if you can simplify your response!

When your evaluator looks at your responses, he or she will use a specific list of rubric items to assess your answers. Here is the link to that rubric: Link to the rubric Each question has one or more specific rubric items associated with it, so before you submit an answer, take a look at that part of the rubric. If your response does not meet expectations for all rubric points, you will be asked to revise and resubmit your project. Make sure that your responses are detailed enough that the evaluator will be able to understand the steps you took and your thought processes as you went through the data analysis.

Once youve submitted your responses, your coach will take a look and may ask a few more focused follow-up questions on one or more of your answers.

We cant wait to see what youve put together for this project!

2 Questions

Question 1. *Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those? [relevant rubric items: data exploration, outlier investigation]*

The goal of this project is to be able to identify people of interest in the fraud that happened at Enron between 2000-2002. We can accomplish this by using a machine learning to build a classifier to identify the people that might have been involved in the fraud that accrued. The data that will be used for the classifier is from the CMU Enron data set which contains a large collection of emails from senior management of Enron. Yes there was a major outlier called “TOTAL” which was removed after we first inspected the data and made some plots.

Question 2. *What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in* *the dataset – explain what feature you tried to make, and the rationale behind it. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importances of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values. [relevant rubric items: create new features, properly scale features, intelligently select feature]*

We selected the following features

$$[salary,\ bonus,\ total\_stock\_value,\ expenses,\ exercised\_stock\_options]$$

as well as the two custom features we made

$$\left[\frac{bonus}{\log(salary)},\ \frac{shared\_receipt\_with\_poi + 3 * from\_poi\_to\_this\_person}{to\_messages}\right]$$

We picked these features after see what possible feature had enough available data, greater then 55%, for each person. Then we pick just decided to pick the above six features from the available eleven feature that were left after our cutoff since we feel like someone what would commit fraud would want to line there pockets and expenses the company as much as possible before the company collapsed. We also decided to add the feature $\frac{bonus}{\log(salary)}$ since if someone knew that the company was going to collapse then they might try to hide the fact that they knew what was happening by taking a large bribe in the form of a large bonus that is a order of magnitude grater then or equal to there normal salary. We also would suspect that someone was a person of interest if a large proposition of email is from other know POI.

We did not need to do any feature scaling since the decision trees are not affected from scaling hence AdaBoost is not affected from feature scaling.

Question 3. *What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: pick an algorithm]*

In the end we ended up using the AdaBoost ensemble algorithm. We also tried Naive Bayes, a single decision tree, multiple decision trees in the form of random forest as well as SVM before finding that AdaBoost was a better performer. We found that SVM was very poor compared to the others. We also found that Naive Bayes took less time to train as did fairly well. Random forest also did quite well compared to AdaBoost.

Question 4. *What does it mean to tune the parameters of an algorithm, and what can happen if you dont do this well? How did you tune the parameters of your particular algorithm? (Some algorithms do not have parameters that you need to tune – if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric item: tune the algorithm]*

To tune a parameter of an algorithm means that there is a free variable θ that we need to pick for the algorithm and we need to figure out what the best or a very good value for θ is for the algorithm can deliver better results. We first tuned our parameters using sklearns GridSearchCV to help find a good starting point then we further tuned parameters by hand to get the best results

Question 5. *What is validation, and whats a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric item: validation strategy]*

Validation is a form a sanity check to make sure that you are not over fitting your model to your specific training set and that your model preforms well with new unseen data. If you preform validation wrong you can trick yourself into thinking that you have a good model when in fact you actually have a very poor model. We used 15-fold cross validation to test our model. For validation we would evaluate our model by looking at the percentage of elements correctly classified as well as the precision, recall and the F1 score.

Question 6. *Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm's performance. [relevant rubric item: usage of evaluation metrics]*

AdaBoost	
Accuracy	0.86050
Precision	0.51502
Recall	0.40300
F1	0.45217

Accuracy is the frequency of how many examples are classified correctly as either POI or non-POI. Precision is how often a POI is correctly classified as a POI.