P2: BUILD A STUDENT INTERVENTION SYSTEM

Dalong Li

March 2016

1 Classification vs Regression

Your goal is to identify students who might need early intervention - which type of supervised machine learning problem is this, classification or regression? Why?
Answer: This is a classification problem. What matters to determine if something is classification or regression is whether the output is from a discrete small set or some continuous quantity.

2 Exploring the Data

Can you find out the following facts about the dataset?

- Total number of students
- Number of students who passed
- Number of students who failed
- Graduation rate of the class(%)
- Number of features(excluding the label/target column)

Use the code block below to compute these values.

3 Preparing the Data

Execute the following steps to prepare the data for modeling, training and testing:

- Identify feature and target columns
- Preprocess feature columns
- Split data into training and test sets

Starter code snippets for these steps have been provided in the template.

4 Training and Evaluating Models

Choose 3 supervised learning models that are available in scikit-learn, and appropriate for this problem. For each model:

- What are the general applications of this model? What are its strengths and weaknesses?
- Given what you know about the data so far, why did you choose this model to apply?
- Fit this model to the training data, try to predict labels (for both training and test sets), and measure the F1 score. Repeat this process with different training set sizes (100, 200, 300), keeping test set constant.
- Produce a table showing training time, prediction time, F1 score on training set and F1 score on test set, for each training set size.

Note: You need to produce 3 such tables - one for each model.
Answer:

- Decision Tree classifier:

  1. Decision trees are a non-parametric supervised learning method, it's goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It's advantages are automatic feature selections, little data prep effort, handle data non-linearity, easy to interpret, but it's disadvantage is that they're prone to over fitting, especially if the data has lots of features and create a complicated decision tree.
  2. Decision trees handles well on predicting a categorical value, and don't need any preparation on data.
  3.

	Training set size		
	100	200	300
Training time(secs)	0.001	0.002	0.003
Prediction time(secs)	0.000	0.000	0.000
F1 score for training set	1.0	1.0	1.0
F1 score for test set	0.644	0.761	0.7

- SVM classifier:

  1. Support vector machines(SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. The advantages of SVMs are: effective in high dimensional spaces, memory efficient, different kernel functions can be specified. It's disadvantages are bad time performance on big data, prone to over fitting to some of the noise data.
  2. The given student-data has 30 features. SVMs can effective address high dimensional issue, and it's memory effictive.

3.

	Training set size		
	100	200	300
Training time(secs)	0.003	0.007	0.0015
Prediction time(secs)	0.002	0.002	0.003
F1 score for training set	0.859	0.858	0.858
F1 score for test set	0.833	0.841	0.846

- Gaussian Naive Bayes classifier:

  1. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictions. Naive Bayes is super simple and it'll converge quick if independence assumption holds. Its main disadvantage is that it can't learn interactions between features.
  2. Naive Bayesian model is fast and easy , besides it needs less training data.
  3.

	Training set size		
	100	200	300
Training time(secs)	0.002	0.001	0.001
Prediction time(secs)	0.001	0.000	0.000
F1 score for training set	0.835	0.788	0.792
F1 score for test set	0.740	0.645	0.672

5 Choosing the Best Model

- Based on the experiments you performed earlier, in 1-2 paragraphs explain to the board of supervisors what single model you chose as the best model. Which model is generally the most appropriate based on the available data, limited resources, cost, and performance?
  Answer: I advise to use SVMs model. First, based on the experiment we can see SVMs gives the best accuracy(F1 test DT=0.77, SVM=0.85 , GaussianNB=0.67); second, the data have lots of features, it's suitable for SVMs; finally, SVMs is memory efficient, so we can save some pay for memory.

- In 1-2 paragraphs explain to the board of supervisors in layman's terms how the final model chosen is supposed to work (for example if you chose a Decision Tree or Support Vector Machine, how does it make a prediction).
  Answer: In training phase, we random pick some students data that's already known whether student pass or failure to learn our prediction model. For simple, we assume that the students data only have two features, studytime and absences, we want to find a line to divide those picked students data into two parts as well as possible. One part is students who pass, another part is students who failure. Besides, we hope the separated line has biggest distance between the closest student data point in each of the pass group and failure group. In fact, the students data has 30 features, we can't just draw a line to address the problem. But the SVMs

algorithm can use a tricky to make students data linear separable, finally we can find a hyperplane to separate the data.
In predictive phase, depending on where a new student data point lands on either side of the hyperplane, that's what class we can classify the new data as. That's say, if the new student data lands on pass side of the hyperplane, we predict the student will pass, otherwise, we predict the student will failure.

- Fine-tune the model. Use Gridsearch with at least one important parameter tuned and with at least 3 settings. Use the entire training set for this.

- What is the model's final F1 score?
  Answer: 0.846