Building a Student Intervention System: An Udacity Nanodegree ML Project

Omoju Miller

May 10, 2016

Introduction

Your goal is to identify students who might need early intervention - which type of supervised machine learning problem is this, classification or regression? Why?

- This task sounds like a problem that would be best suited for a classification algorithm. The inherent task is to develop a learners that can "predicting a category." If we look at the problem from another perspective, we can consider the student data available as a "labeled" dataset. We have features that we can use to determine who has succeeded in the class versus who has not. For that insight, we could use '`passed`' column as our class label.

Models

For the student intervention challenge, three supervised learning algorithms have been selected as appropriate learners for the task. The algorithms are as follows:

1. Decision tree classifier
2. Support vector machines
3. Random forest classifier

Decision Tree Classifier

- What is the theoretical $O(n)$ time & space complexity in terms of input size?
  The theoretical time & space complexity of decision trees classifiers as implemented in scikit learn package is $O(n_{features}n^2_{samples}log(n_{samples}))$.
- What are the general applications of this model? What are its strengths and weaknesses?
  The decision tree algorithm is usually applied to classification and regression problems.
  - Advantages of decision trees:
    * Simple to understand and to interpret.
    * Requires little data preparation.

    - The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
    - Able to handle both numerical and categorical data.
    - Is an explainable algorithm—a white ox model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic.
    - Possible to validate a model using statistical tests.
    - Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
  - Disadvantages of decision trees:
    - Decision-tree learners can create over-complex trees that do not generalise the data well.
    - Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.
    - The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts.
    - There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.
    - Decision tree learners create biased trees if some classes dominate.
- Given what you know about the data so far, why did you choose this model to apply?

Table 1: Result of training with a DecisionTreeClassifier

	Training set size		
	100	200	300
Training time (secs)	0.001	0.001	0.002
Prediction time (secs)	0.000	0.000	0.000
F1 score for training set	1.000	1.000	1.000
F1 score for test set	0.683	0.703	0.758

Support Vector Machine

- What is the theoretical $O(n)$ time & space complexity in terms of input size?
- 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?

Random Forest Classifier

- What is the theoretical $O(n)$ time & space complexity in terms of input size?
- 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?

Table 2: Result of training with a Support Vector Machine

	Training set size		
	100	200	300
Training time (secs)	0.008	0.010	0.051
Prediction time (secs)	0.000	0.000	0.000
F1 score for training set	0.909	0.853	0.830
F1 score for test set	0.767	0.769	0.779

Table 3: Result of training with a Random Forest Classifier

	Training set size		
	100	200	300
Training time (secs)	0.029	0.020	0.026
Prediction time (secs)	0.000	0.000	0.000
F1 score for training set	0.984	0.989	0.997
F1 score for test set	0.722	0.774	0.694

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?

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).

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?

Conclusion

This paper has laid out some of the challenge of