Udacity P2: Student Intervention Project Answers

John Williams

April 16, 2015

Contents

1. Classification vs Regression

There are two types of supervised machine learning problems, determoned by the tupe of response variable(s): 1) regression ML problems have numerical (continuous or near continuous) response variables; and 2) classification ML problems have categorical response variables. The Studens intervention data set calls for a classification solution as the response variable is categorical - labeled “passed” and taking one of two values: “yes” or “no”, i.e., a binary categorical response variable.

2. Exploring the Data

Total number of students: 395 Number of students who passed: 265 Number of students who failed: 130 Number of features: 30 Graduation rate of the class: 0.67%

3. Preparing the Data

The output of this section does not present well in this manually created pdf document - please see the IPN file for this output.

4. Training and Evaluating Models

Naive Bayes, Support Vector Machine, and Gradient Tree Boosting were chosen as candidates models for student intervention.

Naive Bayes (NB)

NB classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. Strengths of NB: requires a small amount of training data to estimate necessary parameters, and NB learners and classifiers can be extremely fast compared to more sophisticated methods. The major weakness of Naive Bayes classification is it is known to be a bad estimator. I chose NB for its track record in real-world applications, such as student intervention, and its speed.

Gradient Tree Boosting (GBT)

Gradient Tree Boosting or Gradient Boosted Regression Trees (GBRT) is a generalization of boosting to arbitrary differentiable loss functions. GBRT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems. Gradient Tree Boosting models are used in a variety of areas including Web search ranking and ecology. Strengths of GTB are natural handling of heterogeneous (mixed type) features, good predictive power, and robustness to outliers. Weaknesses of GTB are scalability, which due to the sequential nature of boosting can not be parallelized, and generally slower execution time than other classification methods, possibly very slow depending on the data. I choose GBT for its predictive power, and its handling of heterogenous features.

Support Vector Machine (SVM)

SVMs are a set of supervised learning methods used for classification, regression and outlier detection. Strengths of SVMs: effective in high dimensional spaces, and in cases where the number of dimensions is greater than the number of samples. SVMs use a subset of training points in the decision function (called support vectors), so it is also memory efficient. SVM is versatile - different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. SVM's can Work well with both dense and sparse data. Weaknesses of SVMs: If the number of features is much greater than the number of samples, the method is likely to yield poor performance. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. SVM's can be extremely fast or extremely slow, depending on the data and chosen parameters. I chose SVM for its versatility, effectiveness in high dimensional space, as the student intervention data has 48 features (after preprocessing).

5. Choosing the Best Model

Model Selection

The Naive Bayes model is inconsistant with this data set - modeling results vary significantly with training set size. The problem with Gradient Tree Boosting is time inefficiency, and while not as extreme as Naive Bayes, results vary with training set size. The SVM(rbf) model does not vary with training set size and is efficient in time and space. The optimal constant test was found for each model, and SVM had the highest and consistant prediction F1 scores for each training set size. We therefore recommend SVM(rbf) to model student intervention data.

Model Recommendation to Board of Supervisors

Based on analysis of available data, limited resources, cost and performance, support vector machine (SVM) classification is the most appropriate model for predicting student intervention candidates. SVM consistantly performs at a high level (i.e., prediction results & speed of execution) across varying amounts of student data. SVM's speed of execution and small memory footprint minimize processing cost, conforming to budget constraints. Through the use of scoring metrics, SVM will allow the board of supervisors to match the most needy intervention candidates based on available intervention resources and intervention strategy.

Model Explanation to Board of Supervisors

Support Vector Machine (SVM) classification works by defining classification (i.e., passed = yes or no) boundaries amoung the predictive features of student intervention data. SVM considers all feasible boundaries and chooses the boundary with the largest "margin", i.e., distance between the boundary and the closest student data point. The SVM implementing this boundary is known as the maximum margin classifier and provides the best predictive capabilities among SVMs with alternate boundaries. SVM recommends students for intervention by looking where the student's features lie relative to the boundary - the side of the boundary that student features lie upon determines the predicted classification (passed = yes or no) for that student. Students predicted as passed = no are recommended for intervention. The model can be adjusted so that only the most needy students are recommended for intervention, matching the resources available for intervention. I.e., if intervention resources are only available for x number of students, the SVM model can be configured to recommend the x students most in need of intervention. Comversely, if the board of supervisors seeks to also intervene with low performing students predicted to pass, the SVM model can be adjusted to reconmend such students for intervention as well.

Tuned Model F1 Score

The best parameters are {'kernel': 'rbf', 'C': 31622.776601683792, 'gamma': 9.9999999999999995e-08} with a F1 score of 0.80

Comparison of FInal Model to Default Model

The performance of the tuned model, as measured by its F1 score of 0.80, performs as well or better that the default model, with its F1 score - also 0.80.