Project 2: Supervised Learning

Building a student intervention system

Classification Vs Regression

This is a classification supervised learning problem. Because we are trying to predict whether students are likely to pass or not. So this is a binary classification problem.

Training and Evaluating Models

The 3 supervised learning models I choose are

1. Logistic Regression
2. Support Vector Machine
3. Gaussian Naive Bayes

Logistic regression converts the dependent variable from a classification type to a probability and then uses linear regression to predict the probability of the dependent variable. Eventually a threshold is used to convert the probability of dependent variable to a classification variable by using the logic if the probability is above threshold then true otherwise false. Logistic regression can be used to solve binary classification problem. Its strength is that it can give probability of the dependent variable, so we can have a measure of likelihood of the dependent variable. Its weakness is that given a threshold, it will not be able to determine class of some dependent variable that are equal to the threshold. Since in the current data we are trying to predict pass or no pass for a student, which is a binary classification, logistic regression fits really well.

Support Vector Machine models the features as points in space and tries to form boundary lines between different classes of variables, thus separating each class of features into different planes. Its strength is in dividing features that have a clear separation. It can quickly separate different classes of features when they distinctively apart. Its weakness is that it fails in situation when points are uniformly distributed. Since the current data can be divided into two classes, support vector machine can separate them into two planes.

Gaussian Naive Bayes gives weight to each feature depending on how it relates to the label and then uses the weighted sum of features as a model to calculate the dependent variable. Naive bayes can be simple and fast when feature don't have interdependency. But its weakness is when two features depend on each other, it would put stronger weight on those features and hence will be biased. Also naive bayes cannot optimize as well as logistic regression. Since the current data has features

independent of each other, naive byes can be used.

Below are 3 tables for model training and prediction results

Model: Logistic Regression

Training Size	Training Time(s)	Training Prediction Time(s)	Training F1	Testing Prediction Time(s)	Testing F1
100	0.004	0.000	0.910	0.000	0.708
200	0.003	0.001	0.842	0.000	0.788
300	0.007	0.000	0.843	0.000	0.782

Model: Support Vector Machine

Training Size	Training Time(s)	Training Prediction Time(s)	Training F1	Testing Prediction Time(s)	Testing F1
100	0.003	0.001	0.877	0.001	0.774
200	0.005	0.003	0.868	0.001	0.781
300	0.010	0.007	0.876	0.002	0.784

Model: Gaussian Naive Bayes

Training Size	Training Time(s)	Training Prediction Time(s)	Training F1	Testing Prediction Time(s)	Testing F1
100	0.000	0.001	0.846	0.000	0.803
200	0.001	0.000	0.840	0.000	0.724
300	0.001	0.001	0.804	0.000	0.763

Choosing the best model

Based on the results I choose support vector machine as the model for prediction. After testing logistic regression and Gaussian naive bayes, it seems support vector machine has the best F1 score with reasonably low training and prediction time compared to other algorithms.

Support vector machine maps features as points in space and tries to separate different classes of points as far apart as possible and then as a new point is given it makes prediction of which group the point belongs to based on which group would be mapped closest to the point.

Using grid search the final F1 score of the model with training data is 0.876 and with test data is 0.787