Building a Student Intervention System

Supervised Learning Project

Classification vs Regression

This is a classification problem. The reason is that the student can graduate or not is not a continue variable which is suitable for regression. There are only two states for student's graduation: pass or fail. So it is classification problem.

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: 67.00%

Preparing the Data

Data has been identified and pre-processed.

Training and Evaluating Models

I choose the following four models to analyze the dataset:

- Support Vector Machine
- K-Nearest Neighbor
- Random Forest

Support Vector Machine

Description

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Time and Memory Complexity

- Space Complexity: O(n^2)
- Training time of O(n^3)

Strengths

- Support Vector Machines are very good for data that has a large number of features;
- Memory efficient
- Easy for nonlinear problems with the kernel method

Weaknesses

- low speed time, especially for nonlinear kernel function

Reasons for Selection

Since we only have 395 data samples, but 30 features, support vector machine may be a good method to handle so many features.

Summaries of time and score

Training set size	100	200	300
Training time (secs)	0.001	0.004	0.006
Prediction time (secs)	0.001	0.003	0.005
F1 score for training set	0.877	0.847	0.869
F1 score for test set	0.794	0.802	0.757

K-Nearest Neighbor

Description

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

Time and Memory Complexity

- Space Complexity: O(n)
- Training time of O(log(n + k))

Strengths

- Very Fast for both training and prediction, Memory and computation time efficient
- Easy algorithm

Weaknesses

- Since k-NN is an instance-based learning, so it need to store all data-points with the model. So It may be a drawback for large amount of data-points.

Reasons for Selection

Common sense tells me that the students who cannot graduate may have the same family background or other features, such as number of class failures. So k-NN is a good choice from its algorithm. Also in this data-set, the number of data-points is not very large. This is another reason I select it.

Summaries of time and score

Training set size	100	200	300

Training time (secs)	0.001	0.001	0.000
Prediction time (secs)	0.001	0.002	0.006
F1 score for training set	0.875	0.860	0.860
F1 score for test set	0.681	0.776	0.759

Random Forest

Description

Random forests is an ensemble learning method for classification, regression method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set.

Time and Memory Complexity

- Space Complexity: O(d^0.5 * n * log(n))
- Training time of O(M * d^0.5 * log(n + k))
- d means the number of features, and M means number of trees.

Strengths

- Relative fast
- Less over-fitting than decision trees

Weaknesses

- Random Forest cannot be used out of the range in the training data.

Reasons for Selection

Random forests is chosen because it is an extension of decision trees. By analyzing the data features, we find out there are a lot category variables, such like, mother's job, parent's cohabitation status, and also a lot binary variables, for example Internet access at home, family educational support. Those kinds of features are very suitable for decision trees and random forest.

Summaries of time and score

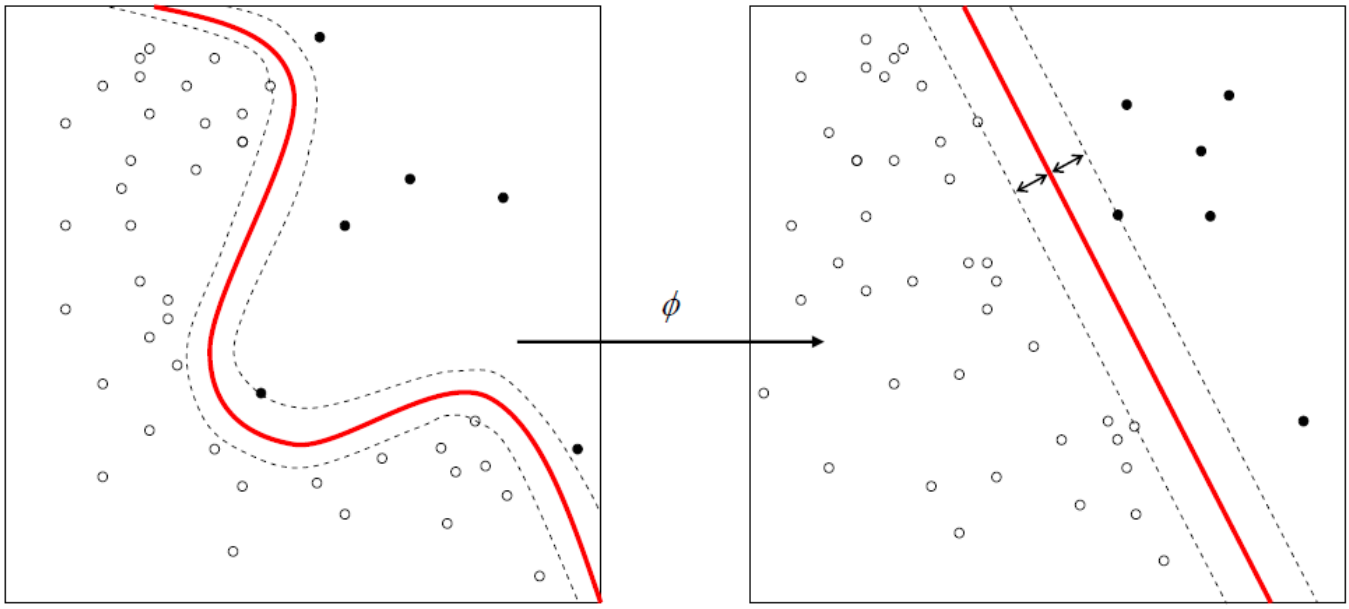
Training set size	100	200	300
Training time (secs)	0.005	0.007	0.007
Prediction time (secs)	0.001	0.001	0.001
F1 score for training set	1.000	1.000	1.000
F1 score for test set	0.716	0.649	0.715

Choosing the Best Model

Based on the testing score from Section 4, Support Vector Machine (SVM) has the best performance score. The negative part of SVM is large training time comparing with other two models. However, the testing time of SVM is still acceptable, which is 0.005s for 300 data size, similar to the k-NN method. Based the fact that computation power is increasing rapidly in recent days, the model with better performance (metric score) is my first choice.

Support Vector Machines means we need a decision lines that can septate different categories. In the ideal situation, this line should go through the middle between two different categories data-point, so that make the gap between line and data-points as wide as possible, just like right below figure.

Furthermore, SVMs can efficiently perform a non-linear classification using what is called the kernel trick. The kernel can be treated as a transformation function. Through this function, the nonlinear boundary line can be transformed to a linear boundary line, shown in the below figure.



copied from Wikipedia

Three parameters of SVM model has been tuned, which are "gamma", "C" and "tolerance". The final model and its F1 score are shown below:

```
SVC(C=160, cache_size=200, class_weight=None, coef0=0.0, degree=3,

    gamma=0.0001, kernel='rbf', max_iter=-1, probability=False,

    random_state=None, shrinking=True, tol=0.01, verbose=False)

Predicting labels using SVC...

Done!

Prediction time (secs): 0.004

F1 score for training set: 0.836820083682

Predicting labels using SVC...

Done!

Prediction time (secs): 0.002

F1 score for test set: 0.794520547945
```

Template code

Open the template iPython notebook student_intervention.ipynb and follow along.

Dataset

Attributes for student-data.csv:

- school - student's school (binary: "GP" or "MS")
- sex - student's sex (binary: "F" - female or "M" - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: "U" - urban or "R" - rural)
- famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian (nominal: "mother", "father" or "other")
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)

- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)
- passed - did the student pass the final exam (binary: yes or no)