Project 2: Supervised Learning

- Building a Student Intervention System –

Masashi Yamaguchi

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 problem is classification type because this target variable is binary value that "Passed" or "Not Passed". Learning problem of classification predict target's labels but regression type predicts continuous numeric value.

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 provided in the template to compute these values.

Answer:

Following is a result of code block.

```
Total number of students: 395
Number of students who passed: 265
Number of students who failed: 130
Number of features: 31
```

```
Graduation rate of the class: 67.09%
```

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.

ANSWER:

Refer to my iPython notebook student_intervation.ipynb file.

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.*

I have chosen following 3 learning models.

1. Decision Tree
2. Logistic Regression
3. Support Vector Machine

Random Forest

Random forest is one of ensemble learning model. It predicts label from votes by decision tree models. Each decision tree models are generated with some of data features chosen randomly.

- General application of this model

Image recognition (recognition of objects on image).

- Strength

Good accuracy with high dimensions data

Calculation time is short by parallel computing

Small overfiting

Easier to understand how model predict label and which feature contributes to predict.

- Weakness

The performance gets bad when redundant features are too bigger than contributed features.

- Why this is good for this data

This model is good for this data because the data has many features.

This model provides which feature is important for the result. And it is useful to make solution on each student for teacher.

Table 1 Result of RandomForest model

Training set size	Training time	Prediction Time	F1 score (training)	F1 score (test)
100	19	1	1	0.787
200	22	1	1	0.75

300	18	1	1	0.768

Logistic Regression

Logistic Regression is one of generalized linear model. It use numeric model of logistic function.

- General application of this model

Predicts incidence rate and finds cause of sickness.

- Strength

Easier to find which feature contributes to predict from regression coefficient, odds ratio.

- Weakness

Classification performance gets worth if data has many features and data size is small.

Difficult to solve non-linear classification problem

- Why this is good for this data

This model can deal target binary label as probability.

This model provides which feature is important for the result. And it is useful to make solution on each student for teacher.

Table 2 Result of LogisticRegression model

Training set size	Training time (ms)	Prediction Time (ms)	F1 score (training)	F1 score (test)
100	1	Less than 1	0.907	0.707
200	2	Less than 1	0.811	0.788
300	2	Less than 1	0.822	0.797

Support Vector Machine (SVM)

Finds the identification boundary with the biggest margin to less false recognition.

- General application of this model

Image recognition. (Recognize human face)

Scientific calculation on complex dataset

- Strength

Good accuracy and short calculation time on high dimension data

Possible to solve non-linear problem depends on kernel function

- Weakness

Difficult to understand why it classify data so for human

Long calculation time if data size is big

Poor performance if the number of features is much bigger than data size.

- Why this is good for this data

This model is good performance with high dimension data and small data size. Generally this is better performance with other models for classification.

Table 3 Result of Support Vector Machine model (kernel='rbf')

Training set size	Training time (ms)	Prediction Time (ms)	F1 score (training)	F1 score (test)
100	1	1	0.910	0.715
200	3	3	0.853	0.813
300	6	4	0.865	0.808

5. Choosing the Best Model

Based on the experiments you performed earlier, in 2-3 paragraphs explain to the board of supervisors what single model you choose as the best model. Which model has the best test F1 score and time efficiency? Which model is generally the most appropriate based on the available data, limited resources, cost, and performance? Please directly compare and contrast the numerical values recored to make your case.

Answer:

Following is the table of F1 score on test, and time on training and prediction.

Table 4 Comparison of F1 score and time on each models (Training size = 300)

Models	Training time (ms)	Prediction Time (ms)	F1 score (test)
Random Forest	18	1	0.768
Logistic Regression	2	Less than 1	0.797
SVM(rbf kernel)	6	1	0.808

I have concluded the best model is Logistic Regression.

Logistic Regression model's F1 score is good close to the SVM. The difference is only1-3 points in several tests. And the time for training and prediction is the shortest in 3 models.

The best thing about Logistic Regression is it can show the importance of each explained variables with coefficient of explained variables or odds ratio. It will be good information for teacher to choose what following up is good for students. Random forest also can show the importance, but its training time is about 10 times longer than Logistic Regression.

In 1-3 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 learn to make a prediction).

Answer:

I choose Logistic Regression as the final model. This is a one of linear model. a posterior probability of target variable is expressed with a numeric model (Eq.1) of logistic sigmoid function (Fig. 1).

$$P(1, \boldsymbol{X}) = \frac{1}{1+e^{-a}}, a \equiv ln\frac{P(1|\boldsymbol{X})}{P(0|\boldsymbol{xX})} = \boldsymbol{\omega^t} \cdot \boldsymbol{X} \quad \cdots \text{(Eq.1)}$$

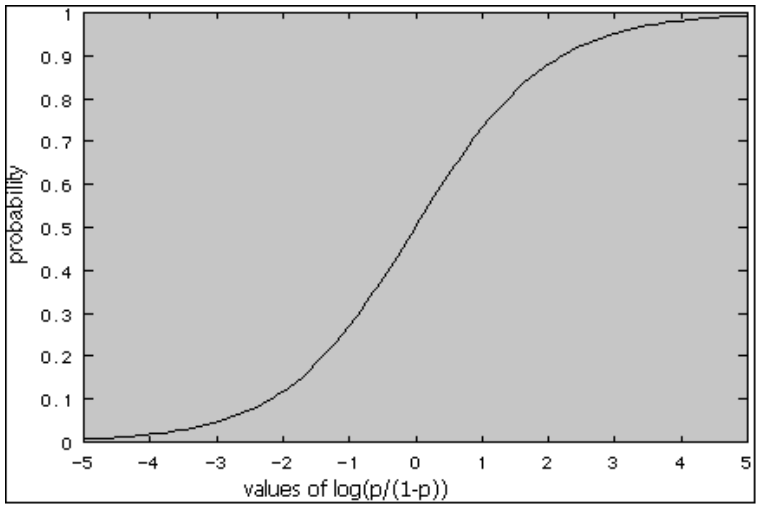

Fig. 1 Logistic Sigmoid Function

(Quoted to http://www.solver.com/logistic-regression)

Finds the parameter ω and C from the following cost function to optimize the numeric model.

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1). \quad \cdots \text{(Eq.2)}$$

For solving the above optimization problem there are several methods such as Steepest descent method, Newton-Raphson method. In scikit learn LogisticRegression supports solvers of 'newton-cg', 'lbfgs', 'liblinear', 'sag'. In this report liblinear is used for model since it is good at small datasets.

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:

Refer to the ipython notebook for the detail.

I tune the parameter of C beteween 1e-2 to 1e2 with settings (solver='liblinear', dual=True, penalty='l2', n_jobs=-1).

Following is table of F1 scores on training, testing, all data for one example.

Table 5 Final model's F1 score

	Training set	Test set	All data
F1 Score	0.80	0.837	0.806