student_intervention

May 15, 2016

1 Machine Learning Engineer Nanodegree

1.1 Supervised Learning

1.2 Project 2: Building a Student Intervention System

Welcome to the second project of the Machine Learning Engineer Nanodegree! In this notebook, some template code has already been provided for you, and it will be your job to implement the additional functionality necessary to successfully complete this project. Sections that begin with **'Implementation'** in the header indicate that the following block of code will require additional functionality which you must provide. Instructions will be provided for each section and the specifics of the implementation are marked in the code block with a `'TODO'` statement. Please be sure to read the instructions carefully!

In addition to implementing code, there will be questions that you must answer which relate to the project and your implementation. Each section where you will answer a question is preceded by a **'Question X'** header. Carefully read each question and provide thorough answers in the following text boxes that begin with **'Answer:'**. Your project submission will be evaluated based on your answers to each of the questions and the implementation you provide.

> **Note:** Code and Markdown cells can be executed using the **Shift + Enter** keyboard shortcut. In addition, Markdown cells can be edited by typically double-clicking the cell to enter edit mode.

1.2.1 Question 1 - Classification vs. Regression

Your goal for this project is to identify students who might need early intervention before they fail to graduate. Which type of supervised learning problem is this, classification or regression? Why?

Answer: Classification problem. Because the result to predict is two discrete values, needing early intervention or not.

1.3 Exploring the Data

Run the code cell below to load necessary Python libraries and load the student data. Note that the last column from this dataset, `'passed'`, will be our target label (whether the student graduated or didn't graduate). All other columns are features about each student.

```
In [49]: # Import libraries
         import numpy as np
         import pandas as pd
```

```
from time import time
from sklearn.metrics import f1_score

# Read student data
student_data = pd.read_csv("student-data.csv")
print "Student data read successfully!"
```

```
Student data read successfully!
```

1.3.1 Implementation: Data Exploration

Let's begin by investigating the dataset to determine how many students we have information on, and learn about the graduation rate among these students. In the code cell below, you will need to compute the following: - The total number of students, `n_students`. - The total number of features for each student, `n_features`. - The number of those students who passed, `n_passed`. - The number of those students who failed, `n_failed`. - The graduation rate of the class, `grad_rate`, in percent (%).

```
In [2]: # TODO: Calculate number of students
        n_students = student_data.shape[0]

        # TODO: Calculate number of features
        n_features = student_data.shape[1]

        # TODO: Calculate passing students
        n_passed = sum(student_data[student_data.columns[-1]] == 'yes')

        # TODO: Calculate failing students
        n_failed = sum(student_data[student_data.columns[-1]] == 'no')

        # TODO: Calculate graduation rate
        grad_rate = n_passed / float(n_passed + n_failed)

        # Print the results
        print "Total number of students: {}".format(n_students)
        print "Number of features: {}".format(n_features)
        print "Number of students who passed: {}".format(n_passed)
        print "Number of students who failed: {}".format(n_failed)
        print "Graduation rate of the class: {:.2f}%".format(grad_rate)
```

```
Total number of students: 395
Number of features: 31
Number of students who passed: 265
Number of students who failed: 130
Graduation rate of the class: 0.67%
```

1.4 Preparing the Data

In this section, we will prepare the data for modeling, training and testing.

1.4.1 Identify feature and target columns

It is often the case that the data you obtain contains non-numeric features. This can be a problem, as most machine learning algorithms expect numeric data to perform computations with.

Run the code cell below to separate the student data into feature and target columns to see if any features are non-numeric.

```
In [3]: # Extract feature columns
        feature_cols = list(student_data.columns[:-1])

        # Extract target column 'passed'
        target_col = student_data.columns[-1]

        # Show the list of columns
        print "Feature columns:\n{}".format(feature_cols)
        print "\nTarget column: {}".format(target_col)

        # Separate the data into feature data and target data (X_all and y_all, res
        X_all = student_data[feature_cols]
        y_all = student_data[target_col]

        # Show the feature information by printing the first five rows
        print "\nFeature values:"
        print X_all.head()
```

```
Feature columns:
['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu', 'Mjob', '

Target column: passed

Feature values:
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher
1     GP   F   17       U     GT3       T     1     1  at_home     other
2     GP   F   15       U     LE3       T     1     1  at_home     other
3     GP   F   15       U     GT3       T     4     2   health  services
4     GP   F   16       U     GT3       T     3     3    other     other

   reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
0  course   mother           2          2         0       yes     no   no
1  course   father           1          2         0        no    yes   no
2   other   mother           1          2         3       yes     no  yes
3    home   mother           1          3         0        no    yes  yes
4    home   father           1          2         0        no    yes  yes
```

```
  activities nursery
0         no     yes ...
1         no      no ...
2         no     yes ...
3        yes     yes ...
4         no     yes ...

[5 rows x 30 columns]
```

1.4.2 Preprocess Feature Columns

As you can see, there are several non-numeric columns that need to be converted! Many of them are simply `yes`/`no`, e.g. `internet`. These can be reasonably converted into `1`/`0` (binary) values.

Other columns, like `Mjob` and `Fjob`, have more than two values, and are known as *categorical variables*. The recommended way to handle such a column is to create as many columns as possible values (e.g. `Fjob_teacher`, `Fjob_other`, `Fjob_services`, etc.), and assign a `1` to one of them and `0` to all others.

These generated columns are sometimes called *dummy variables*, and we will use the `pandas.get_dummies()` function to perform this transformation. Run the code cell below to perform the preprocessing routine discussed in this section.

```
In [4]: def preprocess_features(X):
            ''' Preprocesses the student data and converts non-numeric binary varia
                binary (0/1) variables. Converts categorical variables into dummy v

            # Initialize new output DataFrame
            output = pd.DataFrame(index = X.index)

            # Investigate each feature column for the data
            for col, col_data in X.iteritems():

                # If data type is non-numeric, replace all yes/no values with 1/0
                if col_data.dtype == object:
                    col_data = col_data.replace(['yes', 'no'], [1, 0])

                # If data type is categorical, convert to dummy variables
                if col_data.dtype == object:
                    # Example: 'school' => 'school_GP' and 'school_MS'
                    col_data = pd.get_dummies(col_data, prefix = col)

                # Collect the revised columns
                output = output.join(col_data)

            return output

        X_all = preprocess_features(X_all)
        print "Processed feature columns ({} total features):\n{}".format(len(X_all
```

```
Processed feature columns (48 total features):
['school_GP', 'school_MS', 'sex_F', 'sex_M', 'age', 'address_R', 'address_U', 'fams
```

1.4.3 Implementation: Training and Testing Data Split

So far, we have converted all *categorical* features into numeric values. For the next step, we split the data (both features and corresponding labels) into training and test sets. In the following code cell below, you will need to implement the following: - Randomly shuffle and split the data (`X_all`, `y_all`) into training and testing subsets. - Use 300 training points (approximately 75%) and 95 testing points (approximately 25%). - Set a `random_state` for the function(s) you use, if provided. - Store the results in `X_train`, `X_test`, `y_train`, and `y_test`.

```
In [5]: # TODO: Import any additional functionality you may need here
        from sklearn.utils import shuffle
        from sklearn.cross_validation import train_test_split

        # TODO: Set the number of training points
        num_train = 300

        # Set the number of testing points
        num_test = X_all.shape[0] - num_train

        # TODO: Shuffle and split the dataset into the number of training and testi
        X_shuffle, y_shuffle = shuffle(X_all, y_all, random_state=13)
        X_train, X_test, y_train, y_test = train_test_split(X_shuffle, y_shuffle, t

        # Show the results of the split
        print "Training set has {} samples.".format(X_train.shape[0])
        print "Testing set has {} samples.".format(X_test.shape[0])
```

```
Training set has 300 samples.
Testing set has 95 samples.
```

1.5 Training and Evaluating Models

In this section, you will choose 3 supervised learning models that are appropriate for this problem and available in `scikit-learn`. You will first discuss the reasoning behind choosing these three models by considering what you know about the data and each model's strengths and weaknesses. You will then fit the model to varying sizes of training data (100 data points, 200 data points, and 300 data points) and measure the F1 score. You will need to produce three tables (one for each model) that shows the training set size, training time, prediction time, F1 score on the training set, and F1 score on the testing set.

1.5.1 Question 2 - Model Application

List three supervised learning models that are appropriate for this problem. What are the general applications of each model? What are their strengths and weaknesses? Given what you know about the data, why did

you choose these models to be applied?

Answer: Logistic Regression, SVM and Random Forest. All of these are use for classification problem. LR is easy to implement, computational efficient and space saving, but easy to over-fitting, hard to deal with multi-classification problem and lack of accuracy. SVM has excellent generalization performance, and easy adaptation to multi-classification problem, but there is no general method for its kernel function selection. Random Forest also has excellent generalization performance without the burden of feature selection, and easy adaptation to multi-classification problem, but training data with larger noise may make it overfitting. Random Forest also need large training data.

1.5.2 Setup

Run the code cell below to initialize three helper functions which you can use for training and testing the three supervised learning models you've chosen above. The functions are as follows: - `train_classifier` - takes as input a classifier and training data and fits the classifier to the data. - `predict_labels` - takes as input a fit classifier, features, and a target labeling and makes predictions using the F1 score. - `train_predict` - takes as input a classifier, and the training and testing data, and performs `train_clasifier` and `predict_labels`. - This function will report the F1 score for both the training and testing data separately.

```
In [6]: def train_classifier(clf, X_train, y_train):
            ''' Fits a classifier to the training data. '''

            # Start the clock, train the classifier, then stop the clock
            start = time()
            clf.fit(X_train, y_train)
            end = time()

            # Print the results
            print "Trained model in {:.4f} seconds".format(end - start)


        def predict_labels(clf, features, target):
            ''' Makes predictions using a fit classifier based on F1 score. '''

            # Start the clock, make predictions, then stop the clock
            start = time()
            y_pred = clf.predict(features)
            end = time()

            # Print and return results
            print "Made predictions in {:.4f} seconds.".format(end - start)
            return f1_score(target, y_pred, pos_label='yes')


        def train_predict(clf, X_train, y_train, X_test, y_test):
            ''' Train and predict using a classifer based on F1 score. '''
```

```
    # Indicate the classifier and the training set size
    print "Training a {} using a training set size of {}. . .".format(clf._

    # Train the classifier
    train_classifier(clf, X_train, y_train)

    # Print the results of prediction for both training and testing
    print "F1 score for training set: {:.4f}.".format(predict_labels(clf, X
    print "F1 score for test set: {:.4f}.".format(predict_labels(clf, X_tes
```

1.5.3 Implementation: Model Performance Metrics

With the predefined functions above, you will now import the three supervised learning models of your choice and run the `train_predict` function for each one. Remember that you will need to train and predict on each classifier for three different training set sizes: 100, 200, and 300. Hence, you should expect to have 9 different outputs below — 3 for each model using the varying training set sizes. In the following code cell, you will need to implement the following: - Import the three supervised learning models you've discussed in the previous section. - Initialize the three models and store them in `clf_A`, `clf_B`, and `clf_C`. - Use a `random_state` for each model you use, if provided. - **Note:** Use the default settings for each model — you will tune one specific model in a later section. - Create the different training set sizes to be used to train each model. - *Do not reshuffle and resplit the data! The new training points should be drawn from `X_train` and `y_train`.* - Fit each model with each training set size and make predictions on the test set (9 in total).
Note: Three tables are provided after the following code cell which can be used to store your results.

In [52]:
```
# TODO: Import the three supervised learning models from sklearn
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

# TODO: Initialize the three models
clf_A = LogisticRegression(random_state=13)
clf_B = SVC(random_state=13)
clf_C = RandomForestClassifier(random_state=13)

# TODO: Set up the training set sizes
X_train_100 = X_train[:100]
y_train_100 = y_train[:100]

X_train_200 = X_train[:200]
y_train_200 = y_train[:200]

X_train_300 = X_train[:300]
y_train_300 = y_train[:300]

# TODO: Execute the 'train_predict' function for each classifier and each
# train_predict(clf, X_train, y_train, X_test, y_test)
```

```
# LogisticRegression:
print("LogisticRegression:")
train_predict(clf_A, X_train_100, y_train_100, X_test, y_test)
train_predict(clf_A, X_train_200, y_train_200, X_test, y_test)
train_predict(clf_A, X_train_300, y_train_300, X_test, y_test)
# SVC
print("SVC:")
train_predict(clf_B, X_train_100, y_train_100, X_test, y_test)
train_predict(clf_B, X_train_200, y_train_200, X_test, y_test)
train_predict(clf_B, X_train_300, y_train_300, X_test, y_test)
# RandomForestClassifier
print("RandomForestClassifier:")
train_predict(clf_C, X_train_100, y_train_100, X_test, y_test)
train_predict(clf_C, X_train_200, y_train_200, X_test, y_test)
train_predict(clf_C, X_train_300, y_train_300, X_test, y_test)
```

```
LogisticRegression:
Training a LogisticRegression using a training set size of 100. . .
Trained model in 0.0016 seconds
Made predictions in 0.0001 seconds.
F1 score for training set: 0.8904.
Made predictions in 0.0001 seconds.
F1 score for test set: 0.7692.
Training a LogisticRegression using a training set size of 200. . .
Trained model in 0.0029 seconds
Made predictions in 0.0001 seconds.
F1 score for training set: 0.8215.
Made predictions in 0.0001 seconds.
F1 score for test set: 0.7910.
Training a LogisticRegression using a training set size of 300. . .
Trained model in 0.0031 seconds
Made predictions in 0.0002 seconds.
F1 score for training set: 0.8270.
Made predictions in 0.0001 seconds.
F1 score for test set: 0.8116.
SVC:
Training a SVC using a training set size of 100. . .
Trained model in 0.0018 seconds
Made predictions in 0.0014 seconds.
F1 score for training set: 0.8434.
Made predictions in 0.0013 seconds.
F1 score for test set: 0.7821.
Training a SVC using a training set size of 200. . .
Trained model in 0.0041 seconds
Made predictions in 0.0030 seconds.
F1 score for training set: 0.8395.
Made predictions in 0.0016 seconds.
F1 score for test set: 0.8026.
```

```
Training a SVC using a training set size of 300. . .
Trained model in 0.0086 seconds
Made predictions in 0.0055 seconds.
F1 score for training set: 0.8718.
Made predictions in 0.0017 seconds.
F1 score for test set: 0.8000.
RandomForestClassifier:
Training a RandomForestClassifier using a training set size of 100. . .
Trained model in 0.0066 seconds
Made predictions in 0.0007 seconds.
F1 score for training set: 0.9928.
Made predictions in 0.0007 seconds.
F1 score for test set: 0.7391.
Training a RandomForestClassifier using a training set size of 200. . .
Trained model in 0.0077 seconds
Made predictions in 0.0008 seconds.
F1 score for training set: 0.9926.
Made predictions in 0.0005 seconds.
F1 score for test set: 0.7442.
Training a RandomForestClassifier using a training set size of 300. . .
Trained model in 0.0080 seconds
Made predictions in 0.0010 seconds.
F1 score for training set: 0.9976.
Made predictions in 0.0005 seconds.
F1 score for test set: 0.7299.
```

1.5.4 Tabular Results

Edit the cell below to see how a table can be designed in Markdown. You can record your results from above in the tables provided.

** Classifer 1 - LogisticRegression**

Training Set Size	Prediction Time (train)	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0001	0.0001	0.8904	0.7692
200	0.0001	0.0001	0.8215	0.7910
300	0.0002	0.0001	0.8270	0.8116

** Classifer 2 - SVC**

Training Set Size	Prediction Time (train)	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0014	0.0013	0.8434	0.7821
200	0.0030	0.0016	0.8395	0.8026
300	0.0055	0.0017	0.8718	0.8000

** Classifer 3 - RandomForestClassifier**

Training Set Size	Prediction Time (train)	Prediction Time (test)	F1 Score (train)	F1 Score (test)
100	0.0007	0.0007	0.9928	0.7391
200	0.0008	0.0005	0.9926	0.7442
300	0.0010	0.0005	0.9976	0.7299

1.6 Choosing the Best Model

In this final section, you will choose from the three supervised learning models the *best* model to use on the student data. You will then perform a grid search optimization for the model over the entire training set (`X_train` and `y_train`) by tuning at least one parameter to improve upon the untuned model's F1 score.

1.6.1 Question 3 - Chosing the Best Model

Based on the experiments you performed earlier, in one to two 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?

Answer: I choose SVC as the best model. Although the Logistic Regression has the best performance on testing set, its F1 score on the training set is just 0.8270 even without normalization. For RandomForestClassifier, there is highly overfitting on it. I don't think there will be good performance even after model tuning, because the dataset is too small for it. I perfer SVC to have a good performance after normalization and model tuning on it due to its not larger overfitting and its highly enough F1 score on training set although it's time and resources consuming.

1.6.2 Question 4 - Model in Layman's Terms

In one to two paragraphs, explain to the board of directors in layman's terms how the final model chosen is supposed to work. For example if you've chosen to use a decision tree or a support vector machine, how does the model go about making a prediction?

Answer: I choose the support vector machine. It first maps the dataset on a higher dimension with the kenerl technique, then separates it into two classes with a hyperplane, normalization is also needed during the process. Grid search and cross validation is used to tune the hyperparameters and select the best model.

1.6.3 Implementation: Model Tuning

Fine tune the chosen model. Use grid search (`GridSearchCV`) with at least one important parameter tuned with at least 3 different values. You will need to use the entire training set for this. In the code cell below, you will need to implement the following: - Import `sklearn.grid_search.gridSearchCV` and `sklearn.metrics.make_scorer`. - Create a dictionary of parameters you wish to tune for the chosen model. - Example: `parameters = {'parameter' : [list of values]}`. - Initialize the classifier you've chosen and store it in `clf`. - Create the F1 scoring function using `make_scorer` and store it in `f1_scorer`. - Set the `pos_label` parameter to the correct value! - Perform grid search on the classifier `clf` using

`f1_scorer` as the scoring method, and store it in `grid_obj`. - Fit the grid search object to the training data (`X_train`, `y_train`), and store it in `grid_obj`.

```
In [50]: # TODO: Import 'gridSearchCV' and 'make_scorer'
         from sklearn.grid_search import GridSearchCV
         from sklearn.metrics import make_scorer
         from sklearn.metrics import f1_score

         # TODO: Create the parameters list you wish to tune
         parameters = [{'C': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000],
                        'kernel': ['linear']},
                       {'C': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000],
                        'kernel': ['rbf'],
                        'gamma': [0.0001, 0.001, 0.01, 1, 10, 100, 1000]}]

         # TODO: Initialize the classifier
         clf = SVC(random_state=13)

         # TODO: Make an f1 scoring function using 'make_scorer'
         f1_scorer = make_scorer(f1_score, pos_label='yes')

         # TODO: Perform grid search on the classifier using the f1_scorer as the s
         grid_obj = GridSearchCV(clf, param_grid=parameters, scoring=f1_scorer, cv=

         # TODO: Fit the grid search object to the training data and find the optim
         grid_obj = grid_obj.fit(X_train, y_train)

         # Get the estimator
         clf = grid_obj.best_estimator_

         # Report the final F1 score for training and testing after parameter tunin
         print "Tuned model has a training F1 score of {:.4f}.".format(predict_labe
         print "Tuned model has a testing F1 score of {:.4f}.".format(predict_label
         print(clf)
```

```
Made predictions in 0.0050 seconds.
Tuned model has a training F1 score of 0.8323.
Made predictions in 0.0021 seconds.
Tuned model has a testing F1 score of 0.8299.
SVC(C=10, cache_size=200, class_weight=None, coef0=0.0, degree=3, gamma=0.001,
  kernel='rbf', max_iter=-1, probability=False, random_state=13,
  shrinking=True, tol=0.001, verbose=False)
```

1.6.4 Question 5 - Final F1 Score

What is the final model's F1 score for training and testing? How does that score compare to the untuned model?

Answer: The training and testing F1 score is 0.8323 and 0.8299. It nearly eliminates the overfitting problem on the untuned model. It should be noted that the final model uses a rbf kernel which is useful for the nonlinear separatable dataset while the untuned model's kernel is linear. Its testing F1 score(0.8299) is even better the training score(0.8270) on the untuned LogisticRegression model. So LogisticReggression model has no change to exceed SVC.

> **Note**: Once you have completed all of the code implementations and successfully answered each question above, you may finalize your work by exporting the iPython Notebook as an HTML document. You can do this by using the menu above and navigating to
> **File -> Download as -> HTML (.html)**. Include the finished document along with this notebook as your submission.