Model2C19-Appendix

February 25, 2015

0.0.1 Train/Test split already done

```
In [2]: #from sklearn.cross_validation import train_test_split

        # create 80%-20% train-test split
        #X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=5555)

In [3]: twoC19_test = pd.read_csv("data/test2c19.csv", index_col=’SID’)
        twoC19_train = pd.read_csv("data/training2c19.csv", index_col=’SID’)

In [6]: # Isolate response variable
        ActivityScore = twoC19_train[’ActivityScore’]
        y_train = np.where(ActivityScore >= 40,1,0)

        ActivityScore2 = twoC19_test[’ActivityScore’]
        y_test = np.where(ActivityScore2 >= 40,1,0)

In [7]: # looks right sized
        y_train.shape, y_test.shape

Out[7]: ((9462,), (2366,))

In [8]: y_test

Out[8]: array([1, 1, 0, ..., 1, 0, 0])

In [9]: # We don’t need this column anymore
        to_drop = [’ActivityScore’]
        inhib_feat_space = twoC19_train.drop(to_drop,axis=1)
        inhib_feat_space_test = twoC19_test.drop(to_drop,axis=1)

In [10]: # Pull out features for future use
         features = inhib_feat_space.columns
         features_test = inhib_feat_space_test.columns

In [11]: X_train = inhib_feat_space.as_matrix().astype(np.float)
         X_test = inhib_feat_space_test.as_matrix().astype(np.float)

In [12]: X_train.shape, X_test.shape

Out[12]: ((9462, 186), (2366, 186))

In [13]: n_pos1 = y_test.sum()
         n_pos1

Out[13]: 1193
```

```
In [14]: n_pos2 = y_train.sum()
         n_pos2
```

```
Out[14]: 4721
```

```
In [15]: print(’Feature space holds ’+repr(X_train.shape[0])+’ observations and ’+repr(X_test.shape[1])
         print(’Unique target labels: ’+repr(np.unique(y_train)))

         print(’Feature space holds ’+repr(X_test.shape[0])+’ observations and ’+repr(X_test.shape[1])+
         print(’Unique target labels: ’+repr(np.unique(y_test)))
```

```
Feature space holds 9462 observations and 186 features
Unique target labels: array([0, 1])
Feature space holds 2366 observations and 186 features
Unique target labels: array([0, 1])
```

```
In [16]: X_test.shape[1]
```

```
Out[16]: 186
```

0.1 Scale the features before training model

```
In [17]: from sklearn.preprocessing import StandardScaler
         scaler = StandardScaler()
         X_train = scaler.fit_transform(X_train)
         X_test = scaler.fit_transform(X_test)
```

```
In [18]: from sklearn.cross_validation import KFold

         def run_cv(X,y,clf_class,**kwargs):
             # Construct a kfolds object
             kf = KFold(len(y),n_folds=5,shuffle=True)
             y_pred = y.copy()

             # Iterate through folds
             for train_index, test_index in kf:
                 X_train, X_test = X[train_index], X[test_index]
                 y_train = y[train_index]
                 # Initialize a classifier with key word arguments
                 clf = clf_class(**kwargs)
                 clf.fit(X_train,y_train)
                 y_pred[test_index] = clf.predict(X_test)
             return y_pred
```

```
In [19]: from sklearn.svm import SVC
         from sklearn.ensemble import RandomForestClassifier as RF
         from sklearn.neighbors import KNeighborsClassifier as KNN

         def accuracy(y_true,y_pred):
             # NumPy interpretes True and False as 1. and 0.
             return np.mean(y_true == y_pred)

         print("K-nearest-neighbors (training set):")
         print("%.3f" % accuracy(y_train, run_cv(X_train,y_train,KNN)))
         print("K-nearest-neighbors (test set):")
         print("%.3f" % accuracy(y_test, run_cv(X_test,y_test,KNN)))
```

```
          print('Support vector machines (training set):')
          print("%.3f" % accuracy(y_train, run_cv(X_train,y_train,SVC)))
          print('Support vector machines (test set):')
          print("%.3f" % accuracy(y_test, run_cv(X_test,y_test,SVC)))
          print("Random forest (training set):")
          print("%.3f" % accuracy(y_train, run_cv(X_train,y_train,RF)))
          print("Random forest (test set):")
          print("%.3f" % accuracy(y_test, run_cv(X_test,y_test,RF)))
```

```
K-nearest-neighbors (training set):
0.730
K-nearest-neighbors (test set):
0.720

/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)

Support vector machines (training set):
0.767

/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)

Support vector machines (test set):
0.756
Random forest (training set):
0.736
Random forest (test set):
0.721
```

```
In [21]: from sklearn.metrics import confusion_matrix

          y_train = np.array(y_train)
          class_names = np.unique(y_train)

          confusion_matrices_training = [
              ( "K-Nearest-Neighbors training", confusion_matrix(y_train,run_cv(X_train,y_train,KNN)) ),
              ( "Support Vector Machines training", confusion_matrix(y_train,run_cv(X_train,y_train,SVC)
              ( "Random Forest taining", confusion_matrix(y_train,run_cv(X_train,y_train,RF)) ),
          ]

          y_test = np.array(y_test)
          class_names = np.unique(y_test)

          confusion_matrices_test = [
              ( "K-Nearest-Neighbors test", confusion_matrix(y_test,run_cv(X_test,y_test,KNN)) ),
              ( "Support Vector Machines test", confusion_matrix(y_test,run_cv(X_test,y_test,SVC)) ),
              ( "Random Forest test", confusion_matrix(y_test,run_cv(X_test,y_test,RF)) ),
          ]

          #draw_confusion_matrices(confusion_matrices,class_names)
          confusion_matrices_training, confusion_matrices_test
```

```
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
/home/ubuntu/miniconda3/lib/python3.3/site-packages/sklearn/svm/base.py:233: DeprecationWarning: using a
  max_iter=self.max_iter, random_seed=random_seed)
```

```
Out[21]: ([('K-Nearest-Neighbors training', array([[3322, 1419],
                   [1113, 3608]])),
           ('Support Vector Machines training', array([[3451, 1290],
                   [ 934, 3787]])),
           ('Random Forest taining', array([[3623, 1118],
                   [1407, 3314]]))],
          [('K-Nearest-Neighbors test', array([[823, 350],
                   [301, 892]])), ('Support Vector Machines test', array([[832, 341],
                   [233, 960]])), ('Random Forest test', array([[890, 283],
                   [387, 806]]))])
```