Physics Dataset Practice Problems

Week of July 20, 2020

2 Breast Cancer

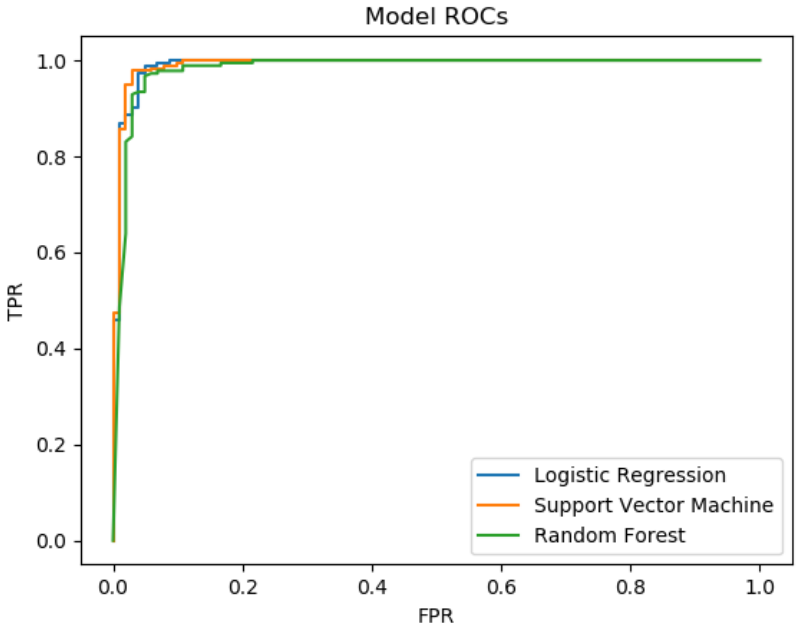
2.1 Visualize the Dataset

```
import pandas as pd
cancer = pd.read_csv('BreastCancer.csv')
#print(cancer.columns)

import matplotlib.pyplot as plt
from seaborn import pairplot
pairplot(cancer, hue='target', vars=cancer.columns[0:5])
plt.show()
```

2.2 Build a Model

```
from sklearn.model_selection import train_test_split
```

```
X = cancer.iloc[:, :-1].values
y = cancer.iloc[:, -1].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5)

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

from sklearn.metrics import roc_curve, roc_auc_score

from sklearn.linear_model import LogisticRegression
logistic_classifier = LogisticRegression()
logistic_classifier.fit(X_train, y_train)
logistic_pred = logistic_classifier.predict_proba(X_test)[:,1]
logistic_roc = roc_curve(y_test, logistic_pred)
logistic_auroc = roc_auc_score(y_test, logistic_pred)

from sklearn.svm import SVC
SV_classifier = SVC(kernel='rbf', probability=True)
SV_classifier.fit(X_train, y_train)
SV_pred = SV_classifier.predict_proba(X_test)[:,1]
SVC_roc = roc_curve(y_test, SV_pred)
SVC_auroc = roc_auc_score(y_test, SV_pred)

from sklearn.ensemble import RandomForestClassifier
RF_classifier = RandomForestClassifier(n_estimators=100)
RF_classifier.fit(X_train, y_train)
RF_pred = RF_classifier.predict_proba(X_test)[:,1]
RF_roc = roc_curve(y_test, RF_pred)
RF_auroc = roc_auc_score(y_test, RF_pred)

models = ['Logistic Regression', 'Support Vector Machine', 'Random Forest']
rocs = [logistic_roc, SVC_roc, RF_roc]
aurocs = [logistic_auroc, SVC_auroc, RF_auroc]

for i in range(3):
    plt.plot(rocs[i][0], rocs[i][1], label=models[i])
plt.legend()
plt.title('Model ROCs')
plt.xlabel('FPR')
plt.ylabel('TPR')
plt.show()
```




```
print('\nAUROCs')
for i in range(3):
    print(models[i] + ': ', aurocs[i])
```

OUTPUTS:

AUROCs
Logistic Regression: 0.9907318118504232
Support Vector Machine: 0.9917497053466194
Random Forest: 0.983044037287046

All three models including linear logistic regression are able to classify malignancy given these features with near 100% accuracy.