Scikit-Learn Basics

1. Load Data Sets

a) Built-in datsets

```
from sklearn import datasets
```

dir(datasets) – will list the available data sets

dset = datasets.load_diabetes() – this is a dictionary which contains actual data, target variable, feature names, and description

b). Other datasets can be loaded in as Numpy arrays or data frames

```
pd.read_csv(), pd.read_table(), pd.read_excel()
```

2. Data Preparation

Train/test split

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y)
```

Standardization

```
from sklearn.preprocessing import StandardScaler
scaler  = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

Polynomial Features

```
from sklearn.preprocessing import PolynomialFeatures
poly =  PolynomialFeatures()
X_train_poly = poly.fit_transform(X_train)
X_test_poly = poly.fit_transform(X_test)
```

3. Training a model

Few ML algorithms

```
from sklearn.cluster import Kmeans
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
```

Fitting the model & making predictions (Ex: Clustering)

```
kmc = Kmeans()
kmc.fit(X)
kmc.labels_
```

Ex: Linear regression

```
lr = LinearRegression()
lr.fit(X_train, y_train)
```

y_pred = lr.predict(X_test) – for predictions on test data

y_prob = lr.predict_proba(X_test) – for probabilities on test data

4. Evaluation

a) Regression

```
from sklearn.metrics import r2_score, mean_squared_error
r2_score(y_test,y_pred), mean_squared_error(y_test,y_pred)
```

b) Classification

```
from sklearn.metrics import accuracy_score, recall_score,
        precision_score, confusion_matrix, roc_auc_score
confusion_matrix(y_test,y_pred), accuracy_score(y_test,y_pred)
recall_score(y_test,y_pred), precision_score(y_test,y_pred)
roc_auc_score(y_test,y_prob)
```

Dr.Palle, spalle@emory.edu , Big Data Analytics Program, ECE, Emory University, Atlanta, GA