Module 4

April 25, 2021

---

*You are currently looking at **version 1.0** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the Jupyter Notebook FAQ course resource.*

---

1 Applied Machine Learning: Module 4 (Supervised Learning, Part II)

1.1 Preamble and Datasets

In [4]:
```
%matplotlib notebook
import numpy as np
import pandas as pd
import seaborn as sn
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.datasets import make_classification, make_blobs
from matplotlib.colors import ListedColormap
from sklearn.datasets import load_breast_cancer
from adspy_shared_utilities import load_crime_dataset


cmap_bold = ListedColormap(['#FFFF00', '#00FF00', '#0000FF','#000000'])

# fruits dataset
fruits = pd.read_table('readonly/fruit_data_with_colors.txt')

feature_names_fruits = ['height', 'width', 'mass', 'color_score']
X_fruits = fruits[feature_names_fruits]
y_fruits = fruits['fruit_label']
target_names_fruits = ['apple', 'mandarin', 'orange', 'lemon']

X_fruits_2d = fruits[['height', 'width']]
y_fruits_2d = fruits['fruit_label']

# synthetic dataset for simple regression
```

```
from sklearn.datasets import make_regression
plt.figure()
plt.title('Sample regression problem with one input variable')
X_R1, y_R1 = make_regression(n_samples = 100, n_features=1,
                             n_informative=1, bias = 150.0,
                             noise = 30, random_state=0)
plt.scatter(X_R1, y_R1, marker= 'o', s=50)
plt.show()

# synthetic dataset for more complex regression
from sklearn.datasets import make_friedman1
plt.figure()
plt.title('Complex regression problem with one input variable')
X_F1, y_F1 = make_friedman1(n_samples = 100, n_features = 7,
                            random_state=0)

plt.scatter(X_F1[:, 2], y_F1, marker= 'o', s=50)
plt.show()

# synthetic dataset for classification (binary)
plt.figure()
plt.title('Sample binary classification problem with two informative featur
X_C2, y_C2 = make_classification(n_samples = 100, n_features=2,
                                 n_redundant=0, n_informative=2,
                                 n_clusters_per_class=1, flip_y = 0.1,
                                 class_sep = 0.5, random_state=0)
plt.scatter(X_C2[:, 0], X_C2[:, 1], marker= 'o',
           c=y_C2, s=50, cmap=cmap_bold)
plt.show()

# more difficult synthetic dataset for classification (binary)
# with classes that are not linearly separable
X_D2, y_D2 = make_blobs(n_samples = 100, n_features = 2,
                        centers = 8, cluster_std = 1.3,
                        random_state = 4)
y_D2 = y_D2 % 2
plt.figure()
plt.title('Sample binary classification problem with non-linearly separable
plt.scatter(X_D2[:,0], X_D2[:,1], c=y_D2,
           marker= 'o', s=50, cmap=cmap_bold)
plt.show()

# Breast cancer dataset for classification
cancer = load_breast_cancer()
(X_cancer, y_cancer) = load_breast_cancer(return_X_y = True)

# Communities and Crime dataset
(X_crime, y_crime) = load_crime_dataset()
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

```
        ---------------------------------------------------------------------------------

        FileNotFoundError                         Traceback (most recent call last)

        <ipython-input-4-9671647cd5a5> in <module>()
         73
         74 # Communities and Crime dataset
    ---> 75 (X_crime, y_crime) = load_crime_dataset()


        /home/jovyan/work/adspy_shared_utilities.py in load_crime_dataset()
         17     # https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unn
         18
    ---> 19     crime = pd.read_table('readonly/CommViolPredUnnormalizedData.txt',
         20     # remove features with poor coverage or lower relevance, and keep V
         21     columns_to_keep = [5, 6] + list(range(11,26)) + list(range(32, 103)


        /opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in parser_f(fil
        644                     skip_blank_lines=skip_blank_lines)
        645
    --> 646         return _read(filepath_or_buffer, kwds)
        647
        648     parser_f.__name__ = name
```

```
/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in _read(filepa
387
388     # Create the parser.
--> 389     parser = TextFileReader(filepath_or_buffer, **kwds)
390
391     if (nrows is not None) and (chunksize is not None):


/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in __init__(sel
728             self.options['has_index_names'] = kwds['has_index_names']
729
--> 730         self._make_engine(self.engine)
731
732     def close(self):


/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in _make_engine
921     def _make_engine(self, engine='c'):
922         if engine == 'c':
--> 923             self._engine = CParserWrapper(self.f, **self.options)
924         else:
925             if engine == 'python':


/opt/conda/lib/python3.6/site-packages/pandas/io/parsers.py in __init__(sel
1388         kwds['allow_leading_cols'] = self.index_col is not False
1389
-> 1390         self._reader = _parser.TextReader(src, **kwds)
1391
1392         # XXX


pandas/parser.pyx in pandas.parser.TextReader.__cinit__ (pandas/parser.c:41


pandas/parser.pyx in pandas.parser.TextReader._setup_parser_source (pandas/


FileNotFoundError: File b'CommViolPredUnnormalizedData.txt' does not exist
```

1.2 Naive Bayes classifiers

In [ ]:
```python
from sklearn.naive_bayes import GaussianNB
from adspy_shared_utilities import plot_class_regions_for_classifier
```

```
X_train, X_test, y_train, y_test = train_test_split(X_C2, y_C2, random_stat

nbclf = GaussianNB().fit(X_train, y_train)
plot_class_regions_for_classifier(nbclf, X_train, y_train, X_test, y_test,
                                  'Gaussian Naive Bayes classifier: Dataset
```

```
In [ ]: X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2,
                                                            random_state=0)

        nbclf = GaussianNB().fit(X_train, y_train)
        plot_class_regions_for_classifier(nbclf, X_train, y_train, X_test, y_test,
                                          'Gaussian Naive Bayes classifier: Dataset
```

1.2.1 Application to a real-world dataset

```
In [ ]: X_train, X_test, y_train, y_test = train_test_split(X_cancer, y_cancer, ran

        nbclf = GaussianNB().fit(X_train, y_train)
        print('Breast cancer dataset')
        print('Accuracy of GaussianNB classifier on training set: {:.2f}'
             .format(nbclf.score(X_train, y_train)))
        print('Accuracy of GaussianNB classifier on test set: {:.2f}'
             .format(nbclf.score(X_test, y_test)))
```

1.3 Ensembles of Decision Trees

1.3.1 Random forests

```
In [ ]: from sklearn.ensemble import RandomForestClassifier
        from sklearn.model_selection import train_test_split
        from adspy_shared_utilities import plot_class_regions_for_classifier_subplo

        X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2,
                                                            random_state = 0)
        fig, subaxes = plt.subplots(1, 1, figsize=(6, 6))

        clf = RandomForestClassifier().fit(X_train, y_train)
        title = 'Random Forest Classifier, complex binary dataset, default settings
        plot_class_regions_for_classifier_subplot(clf, X_train, y_train, X_test,
                                                  y_test, title, subaxes)

        plt.show()
```

1.3.2 Random forest: Fruit dataset

```
In [ ]: from sklearn.ensemble import RandomForestClassifier
        from sklearn.model_selection import train_test_split
        from adspy_shared_utilities import plot_class_regions_for_classifier_subplo
```

```
X_train, X_test, y_train, y_test = train_test_split(X_fruits.as_matrix(),
                                                    y_fruits.as_matrix(),
                                                    random_state = 0)
fig, subaxes = plt.subplots(6, 1, figsize=(6, 32))

title = 'Random Forest, fruits dataset, default settings'
pair_list = [[0,1], [0,2], [0,3], [1,2], [1,3], [2,3]]

for pair, axis in zip(pair_list, subaxes):
    X = X_train[:, pair]
    y = y_train

    clf = RandomForestClassifier().fit(X, y)
    plot_class_regions_for_classifier_subplot(clf, X, y, None,
                                              None, title, axis,
                                              target_names_fruits)

    axis.set_xlabel(feature_names_fruits[pair[0]])
    axis.set_ylabel(feature_names_fruits[pair[1]])

plt.tight_layout()
plt.show()

clf = RandomForestClassifier(n_estimators = 10,
                            random_state=0).fit(X_train, y_train)

print('Random Forest, Fruit dataset, default settings')
print('Accuracy of RF classifier on training set: {:.2f}'
     .format(clf.score(X_train, y_train)))
print('Accuracy of RF classifier on test set: {:.2f}'
     .format(clf.score(X_test, y_test)))
```

Random Forests on a real-world dataset

```
In [ ]: from sklearn.ensemble import RandomForestClassifier

        X_train, X_test, y_train, y_test = train_test_split(X_cancer, y_cancer, ran

        clf = RandomForestClassifier(max_features = 8, random_state = 0)
        clf.fit(X_train, y_train)

        print('Breast cancer dataset')
        print('Accuracy of RF classifier on training set: {:.2f}'
             .format(clf.score(X_train, y_train)))
        print('Accuracy of RF classifier on test set: {:.2f}'
             .format(clf.score(X_test, y_test)))
```

1.3.3 Gradient-boosted decision trees

In [ ]:
```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from adspy_shared_utilities import plot_class_regions_for_classifier_subplo

X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2, random_stat
fig, subaxes = plt.subplots(1, 1, figsize=(6, 6))

clf = GradientBoostingClassifier().fit(X_train, y_train)
title = 'GBDT, complex binary dataset, default settings'
plot_class_regions_for_classifier_subplot(clf, X_train, y_train, X_test,
                                          y_test, title, subaxes)

plt.show()
```

Gradient boosted decision trees on the fruit dataset

In [ ]:
```
X_train, X_test, y_train, y_test = train_test_split(X_fruits.as_matrix(),
                                                    y_fruits.as_matrix(),
                                                    random_state = 0)
fig, subaxes = plt.subplots(6, 1, figsize=(6, 32))

pair_list = [[0,1], [0,2], [0,3], [1,2], [1,3], [2,3]]

for pair, axis in zip(pair_list, subaxes):
    X = X_train[:, pair]
    y = y_train

    clf = GradientBoostingClassifier().fit(X, y)
    plot_class_regions_for_classifier_subplot(clf, X, y, None,
                                              None, title, axis,
                                              target_names_fruits)

    axis.set_xlabel(feature_names_fruits[pair[0]])
    axis.set_ylabel(feature_names_fruits[pair[1]])

plt.tight_layout()
plt.show()
clf = GradientBoostingClassifier().fit(X_train, y_train)

print('GBDT, Fruit dataset, default settings')
print('Accuracy of GBDT classifier on training set: {:.2f}'
     .format(clf.score(X_train, y_train)))
print('Accuracy of GBDT classifier on test set: {:.2f}'
     .format(clf.score(X_test, y_test)))
```

Gradient-boosted decision trees on a real-world dataset

```
In [ ]: from sklearn.ensemble import GradientBoostingClassifier

        X_train, X_test, y_train, y_test = train_test_split(X_cancer, y_cancer, ran

        clf = GradientBoostingClassifier(random_state = 0)
        clf.fit(X_train, y_train)

        print('Breast cancer dataset (learning_rate=0.1, max_depth=3)')
        print('Accuracy of GBDT classifier on training set: {:.2f}'
             .format(clf.score(X_train, y_train)))
        print('Accuracy of GBDT classifier on test set: {:.2f}\n'
             .format(clf.score(X_test, y_test)))

        clf = GradientBoostingClassifier(learning_rate = 0.01, max_depth = 2, rando
        clf.fit(X_train, y_train)

        print('Breast cancer dataset (learning_rate=0.01, max_depth=2)')
        print('Accuracy of GBDT classifier on training set: {:.2f}'
             .format(clf.score(X_train, y_train)))
        print('Accuracy of GBDT classifier on test set: {:.2f}'
             .format(clf.score(X_test, y_test)))
```

1.4 Neural networks

Activation functions

```
In [ ]: xrange = np.linspace(-2, 2, 200)

        plt.figure(figsize=(7,6))

        plt.plot(xrange, np.maximum(xrange, 0), label = 'relu')
        plt.plot(xrange, np.tanh(xrange), label = 'tanh')
        plt.plot(xrange, 1 / (1 + np.exp(-xrange)), label = 'logistic')
        plt.legend()
        plt.title('Neural network activation functions')
        plt.xlabel('Input value (x)')
        plt.ylabel('Activation function output')

        plt.show()
```

1.4.1 Neural networks: Classification

Synthetic dataset 1: single hidden layer

```
In [ ]: from sklearn.neural_network import MLPClassifier
        from adspy_shared_utilities import plot_class_regions_for_classifier_subplo

        X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2, random_stat
```

```
fig, subaxes = plt.subplots(3, 1, figsize=(6,18))

for units, axis in zip([1, 10, 100], subaxes):
    nnclf = MLPClassifier(hidden_layer_sizes = [units], solver='lbfgs',
                          random_state = 0).fit(X_train, y_train)

    title = 'Dataset 1: Neural net classifier, 1 layer, {} units'.format(ur

    plot_class_regions_for_classifier_subplot(nnclf, X_train, y_train,
                                              X_test, y_test, title, axis)
    plt.tight_layout()
```

Synthetic dataset 1: two hidden layers

```
In [ ]: from adspy_shared_utilities import plot_class_regions_for_classifier

        X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2, random_stat

        nnclf = MLPClassifier(hidden_layer_sizes = [10, 10], solver='lbfgs',
                              random_state = 0).fit(X_train, y_train)

        plot_class_regions_for_classifier(nnclf, X_train, y_train, X_test, y_test,
                                          'Dataset 1: Neural net classifier, 2 layer
```

Regularization parameter: alpha

```
In [ ]: X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2, random_stat

        fig, subaxes = plt.subplots(4, 1, figsize=(6, 23))

        for this_alpha, axis in zip([0.01, 0.1, 1.0, 5.0], subaxes):
            nnclf = MLPClassifier(solver='lbfgs', activation = 'tanh',
                                  alpha = this_alpha,
                                  hidden_layer_sizes = [100, 100],
                                  random_state = 0).fit(X_train, y_train)

            title = 'Dataset 2: NN classifier, alpha = {:.3f} '.format(this_alpha)

            plot_class_regions_for_classifier_subplot(nnclf, X_train, y_train,
                                                      X_test, y_test, title, axis)
            plt.tight_layout()
```

The effect of different choices of activation function

```
In [ ]: X_train, X_test, y_train, y_test = train_test_split(X_D2, y_D2, random_stat

        fig, subaxes = plt.subplots(3, 1, figsize=(6,18))
```

```
for this_activation, axis in zip(['logistic', 'tanh', 'relu'], subaxes):
    nnclf = MLPClassifier(solver='lbfgs', activation = this_activation,
                         alpha = 0.1, hidden_layer_sizes = [10, 10],
                         random_state = 0).fit(X_train, y_train)

    title = 'Dataset 2: NN classifier, 2 layers 10/10, {} \
activation function'.format(this_activation)

    plot_class_regions_for_classifier_subplot(nnclf, X_train, y_train,
                                             X_test, y_test, title, axis)
    plt.tight_layout()
```

1.4.2 Neural networks: Regression

```
In [ ]: from sklearn.neural_network import MLPRegressor

        fig, subaxes = plt.subplots(2, 3, figsize=(11,8), dpi=70)

        X_predict_input = np.linspace(-3, 3, 50).reshape(-1,1)

        X_train, X_test, y_train, y_test = train_test_split(X_R1[0::5], y_R1[0::5],

        for thisaxisrow, thisactivation in zip(subaxes, ['tanh', 'relu']):
            for thisalpha, thisaxis in zip([0.0001, 1.0, 100], thisaxisrow):
                mlpreg = MLPRegressor(hidden_layer_sizes = [100,100],
                                     activation = thisactivation,
                                     alpha = thisalpha,
                                     solver = 'lbfgs').fit(X_train, y_train)
                y_predict_output = mlpreg.predict(X_predict_input)
                thisaxis.set_xlim([-2.5, 0.75])
                thisaxis.plot(X_predict_input, y_predict_output,
                             '^', markersize = 10)
                thisaxis.plot(X_train, y_train, 'o')
                thisaxis.set_xlabel('Input feature')
                thisaxis.set_ylabel('Target value')
                thisaxis.set_title('MLP regression\nalpha={}, activation={})'
                                  .format(thisalpha, thisactivation))
                plt.tight_layout()
```

Application to real-world dataset for classification

```
In [ ]: from sklearn.neural_network import MLPClassifier
        from sklearn.preprocessing import MinMaxScaler


        scaler = MinMaxScaler()
```

```
X_train, X_test, y_train, y_test = train_test_split(X_cancer, y_cancer, ran
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

clf = MLPClassifier(hidden_layer_sizes = [100, 100], alpha = 5.0,
                   random_state = 0, solver='lbfgs').fit(X_train_scaled, y_

print('Breast cancer dataset')
print('Accuracy of NN classifier on training set: {:.2f}'
     .format(clf.score(X_train_scaled, y_train)))
print('Accuracy of NN classifier on test set: {:.2f}'
     .format(clf.score(X_test_scaled, y_test)))
```