Unsupervised Learning

April 26, 2021

---

*You are currently looking at **version 1.1** 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: Unsupervised Learning

1.1 Preamble and Datasets

```
In [13]: %matplotlib notebook
         import numpy as np
         import pandas as pd
         import seaborn as sn
         import matplotlib.pyplot as plt
         from sklearn.datasets import load_breast_cancer

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

         # Our sample fruits dataset
         fruits = pd.read_table('readonly/fruit_data_with_colors.txt')
         X_fruits = fruits[['mass','width','height', 'color_score']]
         y_fruits = fruits[['fruit_label']] - 1
```

1.2 Dimensionality Reduction and Manifold Learning

1.2.1 Principal Components Analysis (PCA)

Using PCA to find the first two principal components of the breast cancer dataset

```
In [14]: from sklearn.preprocessing import StandardScaler
         from sklearn.decomposition import PCA
         from sklearn.datasets import load_breast_cancer

         cancer = load_breast_cancer()
```

```
        (X_cancer, y_cancer) = load_breast_cancer(return_X_y = True)

        # Before applying PCA, each feature should be centered (zero mean) and wit
        X_normalized = StandardScaler().fit(X_cancer).transform(X_cancer)

        pca = PCA(n_components = 2).fit(X_normalized)

        X_pca = pca.transform(X_normalized)
        print(X_cancer.shape, X_pca.shape)
```

```
(569, 30) (569, 2)
```

Plotting the PCA-transformed version of the breast cancer dataset

```
In [15]: from adspy_shared_utilities import plot_labelled_scatter
         plot_labelled_scatter(X_pca, y_cancer, ['malignant', 'benign'])

         plt.xlabel('First principal component')
         plt.ylabel('Second principal component')
         plt.title('Breast Cancer Dataset PCA (n_components = 2)');
```

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

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

Plotting the magnitude of each feature value for the first two principal components

```
In [16]: fig = plt.figure(figsize=(8, 4))
         plt.imshow(pca.components_, interpolation = 'none', cmap = 'plasma')
         feature_names = list(cancer.feature_names)

         plt.gca().set_xticks(np.arange(-.5, len(feature_names)));
         plt.gca().set_yticks(np.arange(0.5, 2));
         plt.gca().set_xticklabels(feature_names, rotation=90, ha='left', fontsize=
         plt.gca().set_yticklabels(['First PC', 'Second PC'], va='bottom', fontsize

         plt.colorbar(orientation='horizontal', ticks=[pca.components_.min(), 0,
                                                       pca.components_.max()], pad=
```

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

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

PCA on the fruit dataset (for comparison)

```
In [17]: from sklearn.preprocessing import StandardScaler
         from sklearn.decomposition import PCA

         # each feature should be centered (zero mean) and with unit variance
         X_normalized = StandardScaler().fit(X_fruits).transform(X_fruits)

         pca = PCA(n_components = 2).fit(X_normalized)
         X_pca = pca.transform(X_normalized)

         from adspy_shared_utilities import plot_labelled_scatter
         plot_labelled_scatter(X_pca, y_fruits, ['apple','mandarin','orange','lemon

         plt.xlabel('First principal component')
         plt.ylabel('Second principal component')
         plt.title('Fruits Dataset PCA (n_components = 2)');
```

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

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

1.2.2 Manifold learning methods

Multidimensional scaling (MDS) on the fruit dataset

```
In [18]: from adspy_shared_utilities import plot_labelled_scatter
         from sklearn.preprocessing import StandardScaler
         from sklearn.manifold import MDS

         # each feature should be centered (zero mean) and with unit variance
         X_fruits_normalized = StandardScaler().fit(X_fruits).transform(X_fruits)

         mds = MDS(n_components = 2)

         X_fruits_mds = mds.fit_transform(X_fruits_normalized)

         plot_labelled_scatter(X_fruits_mds, y_fruits, ['apple', 'mandarin', 'orang
         plt.xlabel('First MDS feature')
         plt.ylabel('Second MDS feature')
         plt.title('Fruit sample dataset MDS');
```

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

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

Multidimensional scaling (MDS) on the breast cancer dataset (This example is not covered in the lecture video, but is included here so you can compare it to the results from PCA.)

In [ ]:
```
from sklearn.preprocessing import StandardScaler
from sklearn.manifold import MDS
from sklearn.datasets import load_breast_cancer

cancer = load_breast_cancer()
(X_cancer, y_cancer) = load_breast_cancer(return_X_y = True)

# each feature should be centered (zero mean) and with unit variance
X_normalized = StandardScaler().fit(X_cancer).transform(X_cancer)

mds = MDS(n_components = 2)

X_mds = mds.fit_transform(X_normalized)

from adspy_shared_utilities import plot_labelled_scatter
plot_labelled_scatter(X_mds, y_cancer, ['malignant', 'benign'])

plt.xlabel('First MDS dimension')
plt.ylabel('Second MDS dimension')
plt.title('Breast Cancer Dataset MDS (n_components = 2)');
```

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

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

t-SNE on the fruit dataset (This example from the lecture video is included so that you can see how some dimensionality reduction methods may be less successful on some datasets. Here, it doesn't work as well at finding structure in the small fruits dataset, compared to other methods like MDS.)

In [ ]:
```
from sklearn.manifold import TSNE

tsne = TSNE(random_state = 0)

X_tsne = tsne.fit_transform(X_fruits_normalized)

plot_labelled_scatter(X_tsne, y_fruits,
    ['apple', 'mandarin', 'orange', 'lemon'])
plt.xlabel('First t-SNE feature')
plt.ylabel('Second t-SNE feature')
plt.title('Fruits dataset t-SNE');
```

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

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

t-SNE on the breast cancer dataset Although not shown in the lecture video, this example is included for comparison, showing the results of running t-SNE on the breast cancer dataset. See the reading "How to Use t-SNE effectively" for further details on how the visualizations from t-SNE are affected by specific parameter settings.

```
In [ ]: tsne = TSNE(random_state = 0)

        X_tsne = tsne.fit_transform(X_normalized)

        plot_labelled_scatter(X_tsne, y_cancer,
            ['malignant', 'benign'])
        plt.xlabel('First t-SNE feature')
        plt.ylabel('Second t-SNE feature')
        plt.title('Breast cancer dataset t-SNE');
```

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

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

1.3 Clustering

1.3.1 K-means

This example from the lecture video creates an artificial dataset with make_blobs, then applies k-means to find 3 clusters, and plots the points in each cluster identified by a corresponding color.

```
In [ ]: from sklearn.datasets import make_blobs
        from sklearn.cluster import KMeans
        from adspy_shared_utilities import plot_labelled_scatter

        X, y = make_blobs(random_state = 10)

        kmeans = KMeans(n_clusters = 3)
        kmeans.fit(X)

        plot_labelled_scatter(X, kmeans.labels_, ['Cluster 1', 'Cluster 2', 'Cluste
```

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

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

Example showing k-means used to find 4 clusters in the fruits dataset. Note that in general, it's important to scale the individual features before applying k-means clustering.

```
In [ ]: from sklearn.datasets import make_blobs
        from sklearn.cluster import KMeans
        from adspy_shared_utilities import plot_labelled_scatter
        from sklearn.preprocessing import MinMaxScaler

        fruits = pd.read_table('readonly/fruit_data_with_colors.txt')
        X_fruits = fruits[['mass','width','height', 'color_score']].as_matrix()
        y_fruits = fruits[['fruit_label']] - 1

        X_fruits_normalized = MinMaxScaler().fit(X_fruits).transform(X_fruits)

        kmeans = KMeans(n_clusters = 4, random_state = 0)
        kmeans.fit(X_fruits_normalized)

        plot_labelled_scatter(X_fruits_normalized, kmeans.labels_,
                              ['Cluster 1', 'Cluster 2', 'Cluster 3', 'Cluster 4'])
```

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

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

1.3.2 Agglomerative clustering

```
In [ ]: from sklearn.datasets import make_blobs
        from sklearn.cluster import AgglomerativeClustering
        from adspy_shared_utilities import plot_labelled_scatter

        X, y = make_blobs(random_state = 10)

        cls = AgglomerativeClustering(n_clusters = 3)
        cls_assignment = cls.fit_predict(X)

        plot_labelled_scatter(X, cls_assignment,
                ['Cluster 1', 'Cluster 2', 'Cluster 3'])
```

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

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

Creating a dendrogram (using scipy) This dendrogram plot is based on the dataset created in the previous step with make_blobs, but for clarity, only 10 samples have been selected for this example, as plotted here:

```
In [ ]: X, y = make_blobs(random_state = 10, n_samples = 10)
        plot_labelled_scatter(X, y,
```

```
                ['Cluster 1', 'Cluster 2', 'Cluster 3'])
        print(X)
```

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

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

```
[[  5.69192445  -9.47641249]
 [  1.70789903   6.00435173]
 [  0.23621041  -3.11909976]
 [  2.90159483   5.42121526]
 [  5.85943906  -8.38192364]
 [  6.04774884 -10.30504657]
 [ -2.00758803  -7.24743939]
 [  1.45467725  -6.58387198]
 [  1.53636249   5.11121453]
 [  5.4307043   -9.75956122]]
```

And here's the dendrogram corresponding to agglomerative clustering of the 10 points above using Ward's method. The index 0..9 of the points corresponds to the index of the points in the X array above. For example, point 0 (5.69, -9.47) and point 9 (5.43, -9.76) are the closest two points and are clustered first.

```
In [ ]: from scipy.cluster.hierarchy import ward, dendrogram
        plt.figure()
        dendrogram(ward(X))
        plt.show()
```

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

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

1.3.3 DBSCAN clustering

```
In [ ]: from sklearn.cluster import DBSCAN
        from sklearn.datasets import make_blobs

        X, y = make_blobs(random_state = 9, n_samples = 25)

        dbscan = DBSCAN(eps = 2, min_samples = 2)

        cls = dbscan.fit_predict(X)
        print("Cluster membership values:\n{}".format(cls))

        plot_labelled_scatter(X, cls + 1,
                ['Noise', 'Cluster 0', 'Cluster 1', 'Cluster 2'])
```

```
Cluster membership values:
[ 0  1  0  2  0  0  0  2  2 -1  1  2  0  0 -1  0  0  1 -1  1  1  2  2  2  1]


<IPython.core.display.Javascript object>


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