Machine Learning
Assignment 6

Mao Xiqing (tls868)

January 12, 2021

Contents

1 Unsupervised Learning - Clustering

Source code for this question can be found in the notebook `Kmeans.ipynb`

1.1 Task 1: Principal component analysis and visualisation

Used libraries: `numpy, matplotlib and PCA from sklearn.decomposition.`

The plot of the eigenspectrum is shown on Figure 1.

From `pca.explained_variance_ratio_` I got the percentage of variance explained by each of the selected components:

```
[6.52324185e-01 2.64445180e-01 4.64823465e-02 2.26590394e-02
 9.00833001e-03 2.96225104e-03 1.06610321e-03 6.10103544e-04
 2.37590050e-04 1.93309798e-04 6.52133523e-06 5.04007844e-06
 2.17665260e-10],
```

which means two components are necessary to explain 90% of the variance, because the sum of first two has exceeded 90% ($0.6523 + 0.2644 = 0.9167$).

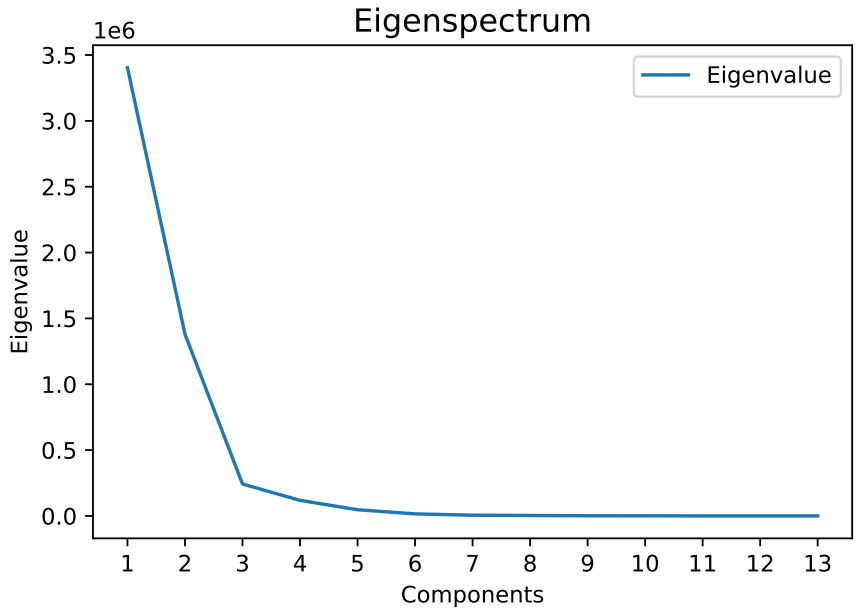

Figure 1: This figure shows eigenvalue vs. rank of corresponding eigenvector after sorting by eigenvalue.

Then I use `pca.components_` to get sorted eigenvectors matrix $\mathbf{E}$. The first two rows of it was used to make new orthogonal projection matrix $\mathbf{P}$. The original 13-dimensional

points was mapped to two dimensions by following equation:

$$\mathbf{x}' = \mathbf{P}\mathbf{x}. \tag{1}$$

The plot for first two principal components is shown on Figure 2.

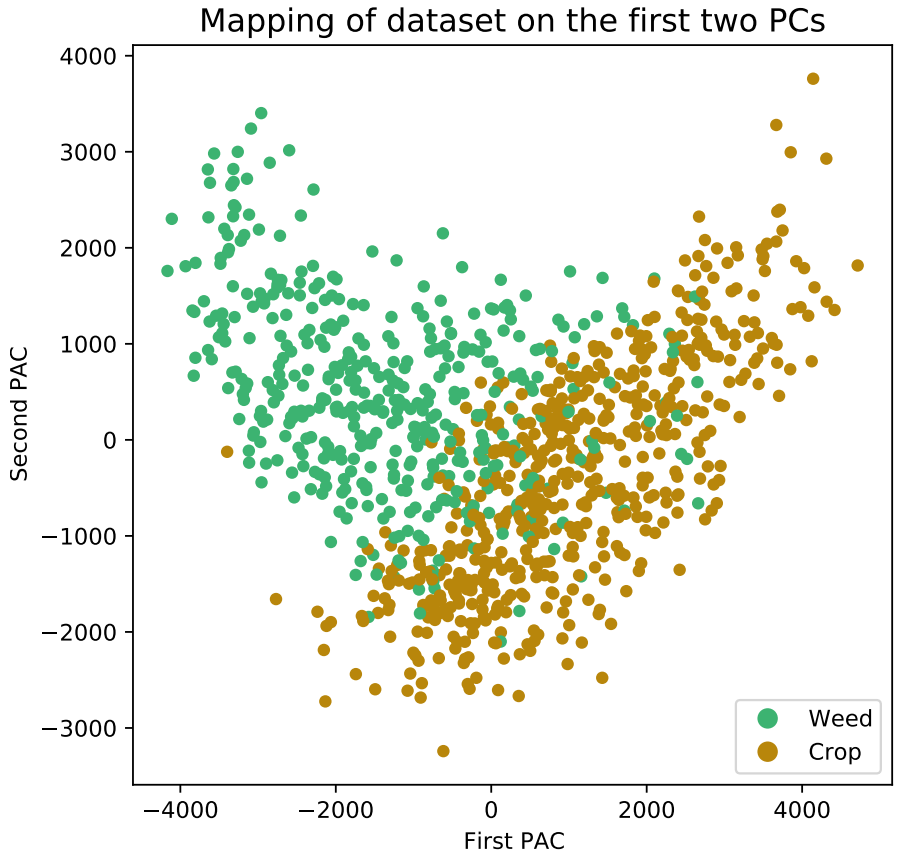

Figure 2: This figure shows first two principal components with green and brown, which indicates weed and crop, respectively.

1.2 Task 2: Clustering

Used libraries: `numpy`, `matplotlib`, `PCA from sklearn.decomposition and KMeans from sklearn.cluster`.

I use the following code to implement clustering and calculate cluster centers:

```
cluster = KMeans(n_clusters = 2, init = X[0:2, 0:13]).fit(X)
center = np.dot(cluster.cluster_centers_, pca_1_2.T)
```

The plot of cluster centers are shown on Figure 3.

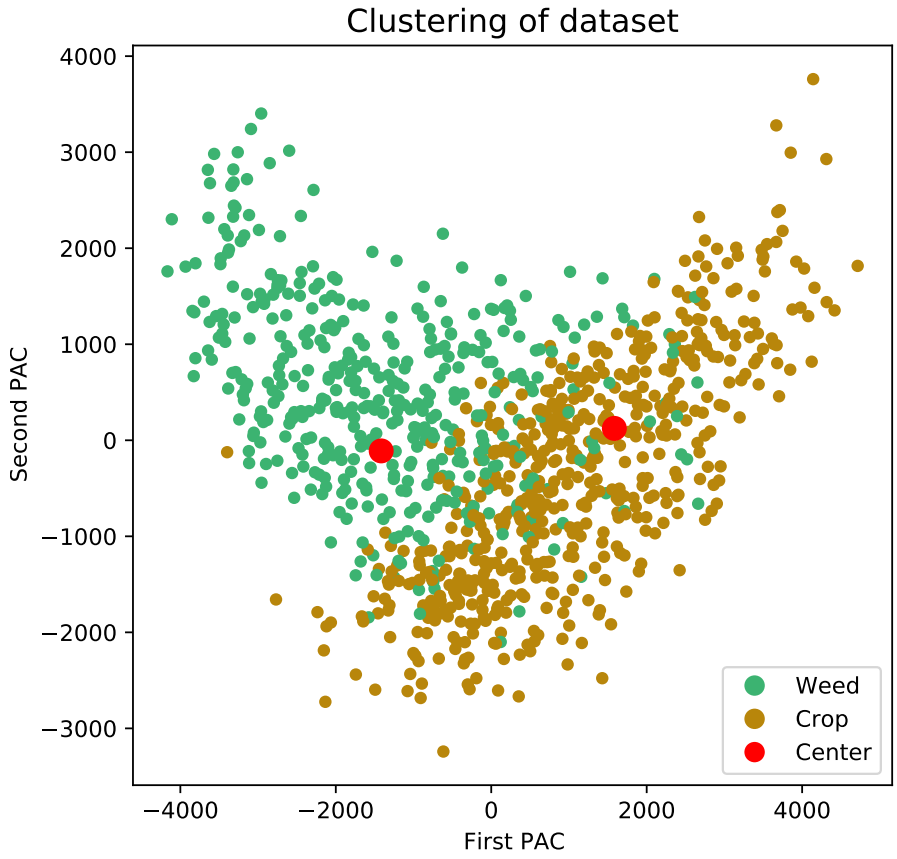

Figure 3: This figure shows cluster centers (red dots).

It seems that KMeans works very well. But, actually, if I re-fill the color of data points using labels generated by KMeans, then around 30% points are wrongly classified, because our data are partly inseparable. So I do not think it is a meaningful clusters. See Figure 4.

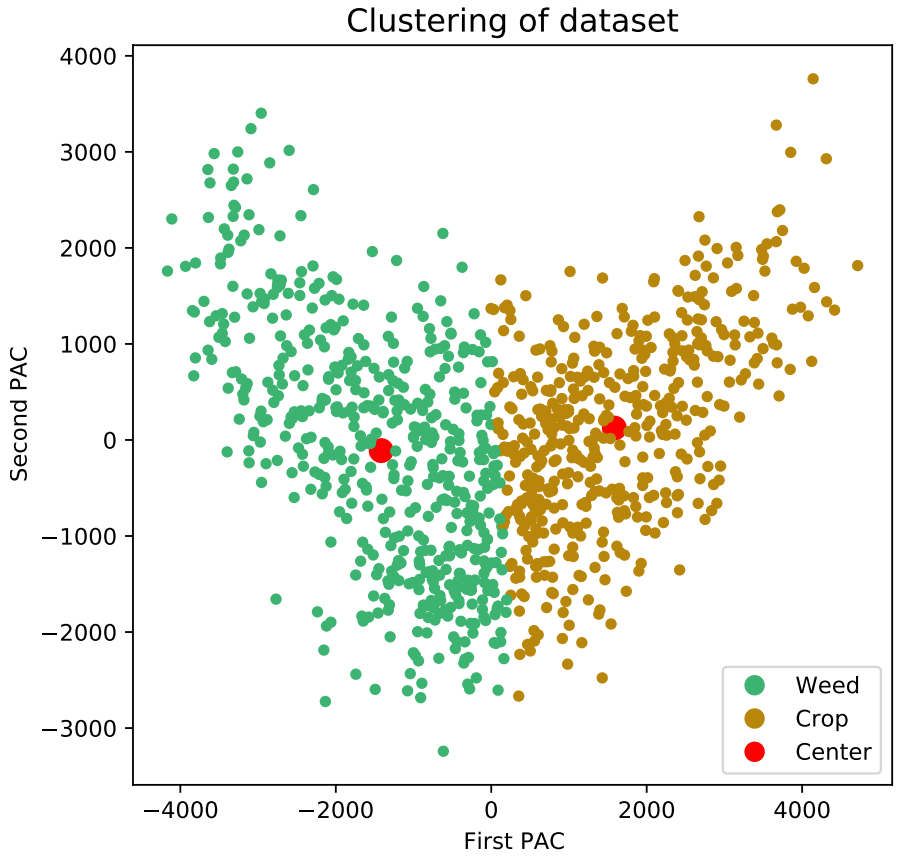

Figure 4: This figure shows cluster centers (red dots) and cluster results with green and brown, which indicates weed and crop, respectively.