Assignment 6, Machine Learning Fall 2018

Cristian Mitroi, dmn470

8 January 2019

Contents

Task 1: Visualization of input data

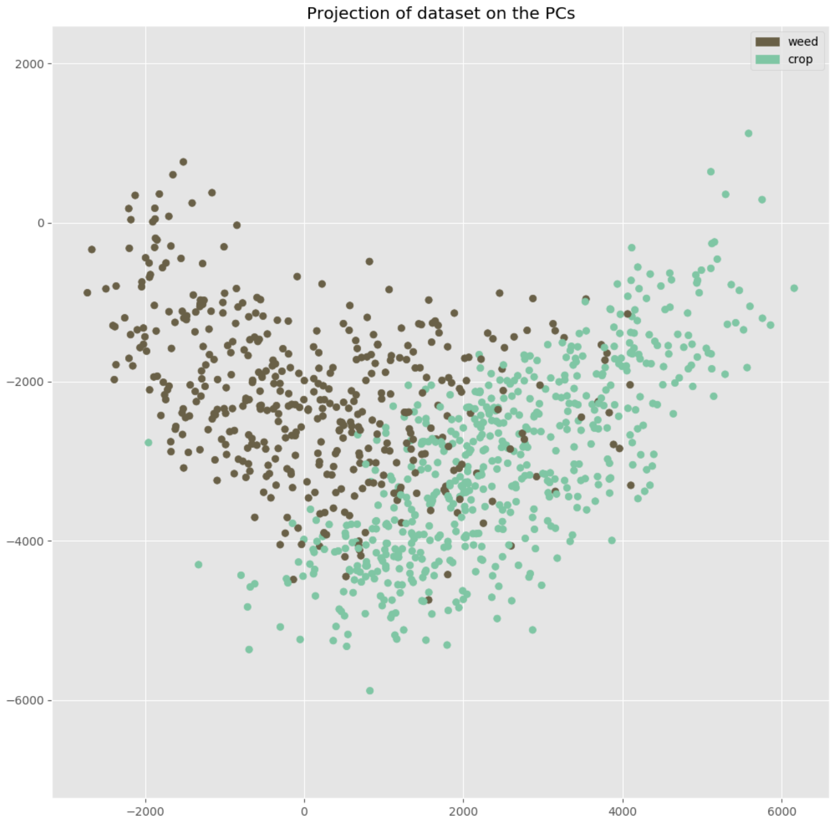

Figure 1: task 1

Task 2: Clustering

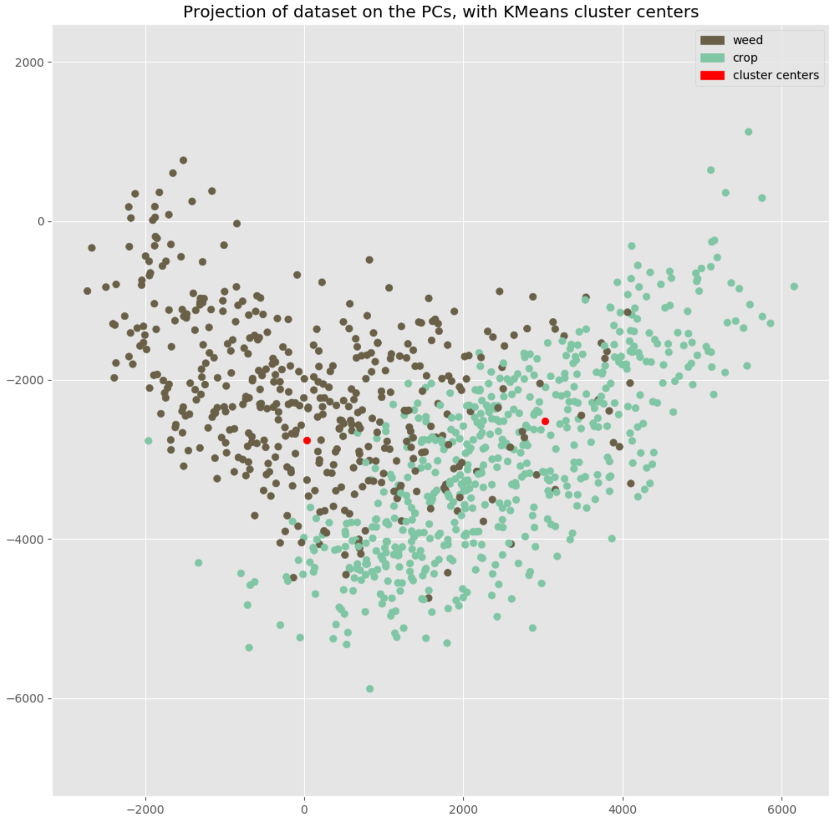

Figure 2: task 2

In the figure “task 2” we can see the two cluster centers plotted in red. They represent the mean of the two respective classes, “weed” and “crop”. The data in this plot has been reduced to two dimensions from the original 13 using PCA (Principal Component Analysis).

The cluster centers themselves are obtained using the *KMeans* algorithm, as implemented by the Python `sklearn` library (Pedregosa et al. 2011). "This algorithm clusters data by minimizing within-cluster coherency - sum of squares criterion. It starts with assigning n samples from the dataset X as being the centroids of our n clusters. It then assigns data points to their nearest centroids. It then computes the new centroids of the clusters from the mean and re-assigns the data points to their new closest centroids. This is repeated until the difference between the old and new centroids is less than a chosen threshold.

[. . .]

For this exercise we are also selecting the first centroids ourselves, even though this is generally not recommended. We set the number of clusters n = 2 and select the first two data points in our set as the initial centroids."[1]

[1] I have had a similar question in an assignment in a previous course: “Introduction to Data Science” Fall 2017, assignment 3. Thus I have put this between quotation marks.

References

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.