ECE 562 Machine Learning, Fall 2019
Instructor: Zafer Aydın
Computer Homework 6

Introduction

Download the template codes for Udacity's Intro to Machine Learning course by running the following command in a terminal

git clone https://github.com/udacity/ud120-projects.git

The starter code can be found in k_means/k_means_cluster.py, which reads in the email + financial (E+F) dataset and gets us ready for clustering. You'll start with performing k-means based on just two financial features

Submit your scripts as q1.py and/or a text document that includes your answers to the questions below through Canvas.

Assignment

1. Take a look at the code. Which features the code uses for clustering? Include the names of these features into your report.

2. Run the code and include the scatter plot into your report.

3. Implement and deploy k-means clustering on the financial_features data, with 2 clusters and number of iterations set to 100. Store your cluster predictions to a list called pred, so that the Draw() command at the bottom of the script works properly. Include the scatterplot obtained by the Draw() command into your report.

The cluster corresponding to blue points include those people who either have high salary or high exercised_stock_options.

4. Add a third feature to features_list, "total_payments". Now rerun clustering (number of clusters set to 2 and number of iterations to 100), using 3 input features instead of 2 (obviously we can still only visualize the original 2 dimensions). Include and compare the plot with the clusterings to the one you obtained with 2 input features. Do any points switch clusters? How many? This new clustering, using 3 features, couldn't have been guessed by eye--it was the k-means algorithm that identified it.

(You'll need to change the code that makes the scatterplot to accommodate 3 features instead of 2, see the comments in the starter code for instructions on how to do this.)

5. In the next lesson, we'll talk about feature scaling. It's a type of feature preprocessing that you should perform before some classification and regression tasks. Here's a sneak preview that should call your attention to the general outline of what feature scaling does.

What are the maximum and minimum values taken by the "exercised_stock_options" feature used in this example?

(NB: if you look at finance_features, there are some "NaN" values that have been cleaned away and replaced with zeroes--so while those might look like the minima, it's a bit deceptive

because they're more like points for which we don't have information, and just have to put in a number. So for this question, go back to data_dict and look for the maximum and minimum numbers that show up there, ignoring all the "NaN" entries.)

6. Repeat question 5 for “salary” feature.

7. Apply feature scaling to your k-means clustering code from the last lesson, on the “salary” and “exercised_stock_options” features (use only these two features). What would be the rescaled value of a "salary" feature that had an original value of $200,000, and an "exercised_stock_options" feature of $1 million? (Be sure to represent these numbers as floats, not integers!)