# STA141C: Homework 3

Wangqian Miao

June 7, 2018

## Environment

- 8GB RAM, Intel i5-6200U laptop.
- Python 3.6 on Windows.

## 1 Problem 1. K-means clustering

### 1.1 Results and Analyse

The dataset is `data_dense.pl`.

| Iteration | Time consuming(s) | Objective function |
|---|---|---|
| 10 | 89.41 s | 55737.841321 |
| 20 | 181.32 s | 55685.979060 |
| 30 | 262.24 s | 55685.808653 |
| 40 | 332.67 s | 55685.808653 |

Analyze the results from the exprement.

1. The k-means algorithm has converged after 30 iterations.
2. When the value of objective function does not change, it means that the algorithm has converged.
3. With more iterations, the objective function decreases slower.

## 2 Problem 2. K-means for sparse data

### 2.1 Results and Aanalyse

The dataset is `data_sparse_E2006.pl`

| Iteration | Time consuming(s) | Objective function |
|---|---|---|
| 10 | 2205 s | 201.33 |
| 20 | 4313 s | 169.79 |
| 30 | 6326 s | 163.54 |
| 40 | 8342 s | 162.57 |

Analyze the results from the exprement.

1. The k-means algorithm has not converged after 40 iterations. With 70-80 iterations, it will converge.

2. We can choose the first ten observations to initialize cluster centers instead of using the random points which will help make the value of objective function acceptable at the beginning.

3. The dataset is quite large, it takes more time to complete each iteration. Use some packages from python will speed up the code.

4. With more iterations, the objective function decreases slower.