Mihir Kelkar
September 14, 2014

CMSC 678 Homework 2

Implementing K-means clustering Algorithm

The code for the K-means clustering Algorithm is attached at the end. The general idea in the assignment was to calculate a series of centroids from a given set of 10000 instances which had class labels attached, group instances into cluster depending on distance from Centroid and then repeat process until all the instances are grouped into clusters which are represented by the same centroids. At this point, it can be said that all instances have converged.

We had to group the instances into 10 clusters with randomly selected initial centroids, 5 clusters with randomly selected initial centroids and 10 clusters with centroids selected such that they represented each instance label.

(a) The results from clustering the instances into 10 clusters where initial centroids were instances which were randomly selected are as follows:

```
============================================================
------- This is Iteration no 1 ----------------------
============================================================
The label for the 1 centroid is 7 with 311 elements
-----------------------------
The label for the 2 centroid is 0 with 795 elements
-----------------------------
The label for the 3 centroid is 1 with 1098 elements
-----------------------------
The label for the 4 centroid is 8 with 621 elements
-----------------------------
The label for the 5 centroid is 6 with 461 elements
-----------------------------
The label for the 6 centroid is 4 with 431 elements
-----------------------------
The label for the 7 centroid is 3 with 721 elements
-----------------------------
The label for the 8 centroid is 2 with 743 elements
-----------------------------
The label for the 9 centroid is 7 with 497 elements
-----------------------------
The label for the 10 centroid is 6 with 416 elements
-----------------------------
============================================================
------- This is Iteration no 2 ----------------------
============================================================
The label for the 1 centroid is 2 with 724 elements
```

```
----------------------------
The label for the 2 centroid is 1 with 1100 elements
----------------------------
The label for the 3 centroid is 8 with 610 elements
----------------------------
The label for the 4 centroid is 4 with 384 elements
----------------------------
The label for the 5 centroid is 7 with 428 elements
----------------------------
The label for the 6 centroid is 5 with 302 elements
----------------------------
The label for the 7 centroid is 3 with 665 elements
----------------------------
The label for the 8 centroid is 0 with 702 elements
----------------------------
The label for the 9 centroid is 9 with 477 elements
----------------------------
The label for the 10 centroid is 6 with 701 elements
----------------------------
============================================================
 ------- This is Iteration no 3 ----------------------
============================================================
The label for the 1 centroid is 8 with 603 elements
----------------------------
The label for the 2 centroid is 7 with 657 elements
----------------------------
The label for the 3 centroid is 1 with 663 elements
----------------------------
The label for the 4 centroid is 5 with 271 elements
----------------------------
The label for the 5 centroid is 3 with 736 elements
----------------------------
The label for the 6 centroid is 9 with 539 elements
----------------------------
The label for the 7 centroid is 1 with 465 elements
----------------------------
The label for the 8 centroid is 4 with 335 elements
----------------------------
The label for the 9 centroid is 0 with 808 elements
----------------------------
The label for the 10 centroid is 6 with 709 elements
----------------------------
============================================================
 ------- This is Iteration no 4 ----------------------
============================================================
The label for the 1 centroid is 3 with 631 elements
----------------------------
The label for the 2 centroid is 0 with 260 elements
----------------------------
The label for the 3 centroid is 5 with 316 elements
----------------------------
The label for the 4 centroid is 0 with 546 elements
----------------------------
The label for the 5 centroid is 4 with 539 elements
----------------------------
The label for the 6 centroid is 8 with 576 elements
----------------------------
The label for the 7 centroid is 1 with 1101 elements
----------------------------
The label for the 8 centroid is 7 with 631 elements
----------------------------
The label for the 9 centroid is 6 with 671 elements
----------------------------
The label for the 10 centroid is 2 with 721 elements
----------------------------
============================================================
 ------- This is Iteration no 5 ----------------------
============================================================
The label for the 1 centroid is 2 with 692 elements
----------------------------
The label for the 2 centroid is 5 with 244 elements
```

```
----------------------------
The label for the 3 centroid is 1 with 489 elements
----------------------------
The label for the 4 centroid is 0 with 675 elements
----------------------------
The label for the 5 centroid is 1 with 640 elements
----------------------------
The label for the 6 centroid is 3 with 715 elements
----------------------------
The label for the 7 centroid is 7 with 643 elements
----------------------------
The label for the 8 centroid is 4 with 562 elements
----------------------------
The label for the 9 centroid is 4 with 345 elements
----------------------------
The label for the 10 centroid is 6 with 721 elements
----------------------------
============================================================
 ------- This is Iteration no 6 ----------------------
============================================================
The label for the 1 centroid is 5 with 306 elements
----------------------------
The label for the 2 centroid is 4 with 522 elements
----------------------------
The label for the 3 centroid is 1 with 1097 elements
----------------------------
The label for the 4 centroid is 4 with 302 elements
----------------------------
The label for the 5 centroid is 6 with 646 elements
----------------------------
The label for the 6 centroid is 8 with 605 elements
----------------------------
The label for the 7 centroid is 3 with 694 elements
----------------------------
The label for the 8 centroid is 7 with 514 elements
----------------------------
The label for the 9 centroid is 0 with 462 elements
----------------------------
The label for the 10 centroid is 0 with 382 elements
----------------------------
============================================================
 ------- This is Iteration no 7 ----------------------
============================================================
The label for the 1 centroid is 2 with 695 elements
----------------------------
The label for the 2 centroid is 8 with 394 elements
----------------------------
The label for the 3 centroid is 1 with 643 elements
----------------------------
The label for the 4 centroid is 1 with 486 elements
----------------------------
The label for the 5 centroid is 3 with 584 elements
----------------------------
The label for the 6 centroid is 4 with 561 elements
----------------------------
The label for the 7 centroid is 0 with 792 elements
----------------------------
The label for the 8 centroid is 7 with 628 elements
----------------------------
The label for the 9 centroid is 8 with 348 elements
----------------------------
The label for the 10 centroid is 6 with 721 elements
----------------------------
============================================================
 ------- This is Iteration no 8 ----------------------
============================================================
The label for the 1 centroid is 1 with 648 elements
----------------------------
The label for the 2 centroid is 4 with 294 elements
----------------------------
The label for the 3 centroid is 8 with 527 elements
```

```
----------------------------
The label for the 4 centroid is 0 with 821 elements
----------------------------
The label for the 5 centroid is 6 with 787 elements
----------------------------
The label for the 6 centroid is 1 with 481 elements
----------------------------
The label for the 7 centroid is 4 with 378 elements
----------------------------
The label for the 8 centroid is 7 with 628 elements
----------------------------
The label for the 9 centroid is 3 with 742 elements
----------------------------
The label for the 10 centroid is 9 with 424 elements
----------------------------
============================================================
 ------- This is Iteration no 9 ----------------------
============================================================
The label for the 1 centroid is 2 with 718 elements
----------------------------
The label for the 2 centroid is 7 with 648 elements
----------------------------
The label for the 3 centroid is 0 with 615 elements
----------------------------
The label for the 4 centroid is 1 with 633 elements
----------------------------
The label for the 5 centroid is 1 with 496 elements
----------------------------
The label for the 6 centroid is 3 with 809 elements
----------------------------
The label for the 7 centroid is 8 with 373 elements
----------------------------
The label for the 8 centroid is 0 with 238 elements
----------------------------
The label for the 9 centroid is 4 with 550 elements
----------------------------
The label for the 10 centroid is 6 with 670 elements
----------------------------
============================================================
 ------- This is Iteration no 10 ----------------------
============================================================
The label for the 1 centroid is 3 with 593 elements
----------------------------
The label for the 2 centroid is 4 with 380 elements
----------------------------
The label for the 3 centroid is 8 with 480 elements
----------------------------
The label for the 4 centroid is 7 with 476 elements
----------------------------
The label for the 5 centroid is 2 with 670 elements
----------------------------
The label for the 6 centroid is 1 with 643 elements
----------------------------
The label for the 7 centroid is 6 with 786 elements
----------------------------
The label for the 8 centroid is 0 with 820 elements
----------------------------
The label for the 9 centroid is 7 with 367 elements
----------------------------
The label for the 10 centroid is 1 with 486 elements
----------------------------
```

• My observations from creating 10 clusters from instances are as follows:

• The average number of iterations for convergence were about 90 with about 5 instance conversions were truncated when 8 of the 10 centroids became constant.

• The distribution of points across the clusters were fairly equal. This points to the fact that good clusters could not be achieved, probably due to the fat that selected centroids often represented a very skewed representation of the instances.

• Most points were distributed in equal measures across all clusters. the size of clusters were almost the same in most instances.

(b) The results from clustering the instances into 5 clusters where initial centroids were instances randomly selected are as follows:

```
============================================================
------- This is Iteration no 1 ----------------------
============================================================
The label for the 1 centroid is 3 with 871 elements
----------------------------
The label for the 2 centroid is 6 with 806 elements
----------------------------
The label for the 3 centroid is 0 with 853 elements
----------------------------
The label for the 4 centroid is 9 with 885 elements
----------------------------
The label for the 5 centroid is 1 with 1126 elements
----------------------------
============================================================
------- This is Iteration no 2 ----------------------
============================================================
The label for the 1 centroid is 3 with 868 elements
----------------------------
The label for the 2 centroid is 6 with 803 elements
----------------------------
The label for the 3 centroid is 0 with 844 elements
----------------------------
The label for the 4 centroid is 9 with 882 elements
----------------------------
The label for the 5 centroid is 1 with 1126 elements
----------------------------
============================================================
------- This is Iteration no 3 ----------------------
============================================================
The label for the 1 centroid is 6 with 802 elements
----------------------------
The label for the 2 centroid is 3 with 867 elements
----------------------------
The label for the 3 centroid is 0 with 848 elements
----------------------------
```

The label for the 4 centroid is 9 with 880 elements

- -

The label for the 5 centroid is 1 with 1126 elements

- -

==========

- - - - - - - This is Iteration no 4 ----------

==========

The label for the 1 centroid is 3 with 850 elements

- -

The label for the 2 centroid is 0 with 828 elements

- -

The label for the 3 centroid is 9 with 861 elements

- -

The label for the 4 centroid is 6 with 668 elements

- -

The label for the 5 centroid is 1 with 1124 elements

- -

==========

- - - - - - - This is Iteration no 5 ----------

==========

The label for the 1 centroid is 9 with 880 elements

- -

The label for the 2 centroid is 0 with 854 elements

- -

The label for the 3 centroid is 3 with 868 elements

- -

The label for the 4 centroid is 6 with 808 elements

- -

The label for the 5 centroid is 1 with 1126 elements

- -

==========

- - - - - - - This is Iteration no 6 ----------

==========

The label for the 1 centroid is 9 with 867 elements

- -

The label for the 2 centroid is 2 with 701 elements

- -

The label for the 3 centroid is 0 with 879 elements

- -

The label for the 4 centroid is 3 with 798 elements

- -

The label for the 5 centroid is 1 with 997 elements

- -

==========

- - - - - - - This is Iteration no 7 ----------

==========

The label for the 1 centroid is 3 with 856 elements

```
----------------------------
The label for the 2 centroid is 0 with 907 elements
----------------------------
The label for the 3 centroid is 7 with 640 elements
----------------------------
The label for the 4 centroid is 4 with 587 elements
----------------------------
The label for the 5 centroid is 1 with 1127 elements
----------------------------
============================================================
 ------- This is Iteration no 8 ----------------------
============================================================
The label for the 1 centroid is 6 with 802 elements
----------------------------
The label for the 2 centroid is 3 with 870 elements
----------------------------
The label for the 3 centroid is 0 with 851 elements
----------------------------
The label for the 4 centroid is 9 with 889 elements
----------------------------
The label for the 5 centroid is 1 with 1126 elements
----------------------------
============================================================
 ------- This is Iteration no 9 ----------------------
============================================================
The label for the 1 centroid is 3 with 855 elements
----------------------------
The label for the 2 centroid is 4 with 574 elements
----------------------------
The label for the 3 centroid is 0 with 911 elements
----------------------------
The label for the 4 centroid is 7 with 631 elements
----------------------------
The label for the 5 centroid is 1 with 1127 elements
----------------------------
============================================================
 ------- This is Iteration no 10 ----------------------
============================================================
The label for the 1 centroid is 6 with 805 elements
----------------------------
The label for the 2 centroid is 1 with 1132 elements
----------------------------
The label for the 3 centroid is 4 with 545 elements
----------------------------
The label for the 4 centroid is 7 with 625 elements
----------------------------
The label for the 5 centroid is 0 with 868 elements
```

```
----------------------------
```

In this scenario we selected 5 random centroids and then grouped all instances into 5 clusters. I did this computation 10 times, and I hypothesize the following from my observations:

- The average number of iterations to convergence was 73, with two instances hitting the maximum threshold set in the code.
- On an average the first cluster usually consisted of class label 3 with around 790 instances, only in one case was this class label 9.
- The second cluster shows a fairly uniform distribution of data points from class labels 0, 3, 6. Thus, the class label sways among the following class labels.
- The third cluster in most cases had a class label of 0. Thus on a co-ordinate axis, cluster 2 must lie between 1 and 3 and its class label changes by the proximity of cluster 2's centroid to cluster 3 or cluster 1.
- The label for cluster 4 varies amongst either 6 or 9
- In most cases , cluster 5 is almost a precise cluster of class label 1 with roughly 950 instances. It sways to a different class label only once in the entire 10 iterations.
- Classification into 5 clusters somehow provides a better generalization than into 10 clusters.

(c) The results from the clustering instances into 10 clusters where initial centroids were instances from each class label are as follows:

```
============================================================
 ------- This is Iteration no 1 ————————————
============================================================
The label for the 1 centroid is 5 with 297 elements
----------------------------
The label for the 2 centroid is 9 with 384 elements
----------------------------
The label for the 3 centroid is 4 with 394 elements
----------------------------
The label for the 4 centroid is 2 with 736 elements
----------------------------
The label for the 5 centroid is 1 with 1069 elements
----------------------------
The label for the 6 centroid is 0 with 753 elements
----------------------------
The label for the 7 centroid is 3 with 701 elements
----------------------------
The label for the 8 centroid is 7 with 563 elements
```

```
----------------------------
The label for the 9 centroid is 4 with 287 elements
----------------------------
The label for the 10 centroid is 6 with 667 elements
----------------------------
============================================================
 ------- This is Iteration no 2 ----------------------
============================================================
The label for the 1 centroid is 3 with 701 elements
----------------------------
The label for the 2 centroid is 9 with 384 elements
----------------------------
The label for the 3 centroid is 1 with 1069 elements
----------------------------
The label for the 4 centroid is 4 with 287 elements
----------------------------
The label for the 5 centroid is 2 with 736 elements
----------------------------
The label for the 6 centroid is 6 with 667 elements
----------------------------
The label for the 7 centroid is 4 with 394 elements
----------------------------
The label for the 8 centroid is 0 with 753 elements
----------------------------
The label for the 9 centroid is 5 with 297 elements
----------------------------
The label for the 10 centroid is 7 with 563 elements
-  ----------------------------


============================================================
 ------- This is Iteration no 10 ----------------------
============================================================
The label for the 1 centroid is 3 with 701 elements
----------------------------
The label for the 2 centroid is 4 with 394 elements
----------------------------
The label for the 3 centroid is 9 with 384 elements
----------------------------
The label for the 4 centroid is 2 with 736 elements
----------------------------
The label for the 5 centroid is 6 with 667 elements
----------------------------
The label for the 6 centroid is 4 with 287 elements
----------------------------
The label for the 7 centroid is 1 with 1069 elements
----------------------------
The label for the 8 centroid is 7 with 563 elements
----------------------------
The label for the 9 centroid is 0 with 753 elements
----------------------------
The label for the 10 centroid is 5 with 297 elements
-  ----------------------------
```

Truncated to Results of Three Iterations.

In this case, when the iterations converged, I observed that the number of instances which belonged to particular class label remained almost constant, despite the fact the centroid may shift slightly in position. In mu opinion this points to the fact that if the initial pair of centroids is a an extremely accurate representation of the instances, then the instances group exactly with the centroid that is similar to its class label. This also means that the instances with similar class label

have an almost precise and perfect grouping. On an average the number of computations required to converge in each of the ten iterations were approximately 42. Only one of the iterations needed an anomalous 68 iterations till convergence.