IS622 Week8 - Clustering

James Quacinella

10/12/2015

Exercise 7.1.3 (section 7.1.4)

Suppose we have a d-dimensional Euclidean space. Consider vectors whose components are only +1 or -1 in each dimension. Note that each vector has length d, so the product of their lengths (denominator in the formula for the cosine of the angle between them) is d. If we chose each component independently, and a component is as likely to be +1 as -1, what is the distribution of the value of the numerator of the formula (i.e., the sum of the products of the corresponding components from each vector)? What can you say about the expected value of the cosine of the angle between the vectors, as d grows large?

Answer

For each component of the summation in the numerator of the formula, you can only have 2 possibilities, +1 or -1, after multiplying. Since they are equally likely, the expected value of the summation would be 0. This is like the example in the book on p243: "The numerator is 0, and as d grows, its standard deviation grows only as d. Thus, for large d, the cosine of the angle between any two vectors is almost certain to be close to 0, which means the angle is close to 90 degrees"

Exercise 7.2.1 (section 7.2.5)

Perform a hierarchical clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81, assuming clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged.

Answer

To help read the code:

- clusters is stored as a list
- print_clusters prints clusters in a nice way
- euclid_dist computes euclidean distance between to numbers, passed in as a list
- euclid_dist.centroids.wrapper is what computers the distance between two clusters by centroids, passed in as a list. NOTE: I had to use a 'closure' since I had issues with cluster not updating
- The other wrapper functions computer other cluster distance metrics as needed
- run_heir runs heirarchical clustering, given a cluster distance metric

```
# Print clusters in a nicer way
print_clusters <- function(clusters) {
  for(i in 1:length(clusters)) {
    print(paste("Cluster", i, ": ", clusters[i]))
  }
  print("==================")
}

# 1 dimensional euclid distance function (same as abs value)
euclid_dist <- function(pair) {
  return( abs(pair[[1]] - pair[[2]]) );
}

euclid_dist.centroids.wrapper <- function(clusters) {
  # Find euclidean distance between cluster centroids given indicies of clusters to compare
  euclid_dist.centroids <- function(pair_idxs) {
    x <- mean( clusters[[ pair_idxs[[1]] ]] )   # Find centroid of first cluster of pair
    y <- mean( clusters[[ pair_idxs[[2]] ]] )   # Find centroid of second cluster of pair
    return(euclid_dist(c(x,y)));
  }

  return(euclid_dist.centroids)
}

run_heir <- function(dist_func) {
  # Initial cluster assignments
  clusters <- list( c(1), c(4), c(9), c(16), c(25), c(36), c(49), c(64), c(81))

  while(length(clusters) > 1) {
    # DEBUG: print cluster state
    print("CLUSTERS:"); print_clusters(clusters)

    # Generate the indicies of the clusters we currently have
    cluster_idxs <- 1:length(clusters)
```

```
    # Minimum distance calculation to find which cluster indicies we need to merge
    pairs_of_clusters <- combn(cluster_idxs, 2, simplify = FALSE)
    distances <- unlist(lapply(pairs_of_clusters, dist_func(clusters) ))
    print( paste("Minimum dist: ", min(distances)) );
    merge_idxs <- pairs_of_clusters[[ which.min(distances) ]]

    # DEBUG: print which clusters we are merging
    print(paste("Merging cluster idx", merge_idxs[[1]], " and cluster idx", merge_idxs[[2]]))

    # Merge: store greater index into lower index; remove greater index;
    small_idx <- min(merge_idxs)
    larger_idx <- max(merge_idxs)
    clusters[[ small_idx ]] <- c( clusters[[ small_idx ]], clusters[[ larger_idx ]])
    clusters[[ larger_idx ]] <- NULL
    clusters <- clusters[!sapply(clusters, is.null)]
  }

  # Print final clustering
  print_clusters(clusters)
}

run_heir(euclid_dist.centroids.wrapper)
```

```
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  1"
## [1] "Cluster 2 :  4"
## [1] "Cluster 3 :  9"
## [1] "Cluster 4 :  16"
## [1] "Cluster 5 :  25"
## [1] "Cluster 6 :  36"
## [1] "Cluster 7 :  49"
## [1] "Cluster 8 :  64"
## [1] "Cluster 9 :  81"
## [1] "=================="
## [1] "Minimum dist:  3"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4)"
## [1] "Cluster 2 :  9"
## [1] "Cluster 3 :  16"
## [1] "Cluster 4 :  25"
## [1] "Cluster 5 :  36"
## [1] "Cluster 6 :  49"
## [1] "Cluster 7 :  64"
## [1] "Cluster 8 :  81"
## [1] "=================="
## [1] "Minimum dist:  6.5"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  16"
## [1] "Cluster 3 :  25"
## [1] "Cluster 4 :  36"
```

```
## [1] "Cluster 5 :  49"
## [1] "Cluster 6 :  64"
## [1] "Cluster 7 :  81"
## [1] "=================="
## [1] "Minimum dist:  9"
## [1] "Merging cluster idx 2  and cluster idx 3"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  c(16, 25)"
## [1] "Cluster 3 :  36"
## [1] "Cluster 4 :  49"
## [1] "Cluster 5 :  64"
## [1] "Cluster 6 :  81"
## [1] "=================="
## [1] "Minimum dist:  13"
## [1] "Merging cluster idx 3  and cluster idx 4"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  c(16, 25)"
## [1] "Cluster 3 :  c(36, 49)"
## [1] "Cluster 4 :  64"
## [1] "Cluster 5 :  81"
## [1] "=================="
## [1] "Minimum dist:  15.8333333333333"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  c(36, 49)"
## [1] "Cluster 3 :  64"
## [1] "Cluster 4 :  81"
## [1] "=================="
## [1] "Minimum dist:  17"
## [1] "Merging cluster idx 3  and cluster idx 4"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  c(36, 49)"
## [1] "Cluster 3 :  c(64, 81)"
## [1] "=================="
## [1] "Minimum dist:  30"
## [1] "Merging cluster idx 2  and cluster idx 3"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  c(36, 49, 64, 81)"
## [1] "=================="
## [1] "Minimum dist:  46.5"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25, 36, 49, 64, 81)"
## [1] "=================="
```

Exercise 7.2.2 (section 7.2.5)

How would the clustering of Example 7.2 change if we used for the distance between two clusters:

(a) The minimum of the distances between any two points, one from each cluster.
(b) The average of the distances between pairs of points, one from each of the two clusters.

Answer

Using different metrics for distances between clusters, we would expect different clustering.

(a) The minimum of the distances between any two points, one from each cluster.

I would expect this to get clustered in a simplistic way, since for this data, the minimum distance between clusters will always point to the next point in the list:

```
euclid_dist.min.wrapper <- function(clusters) {
  # Find minimum euclidean distance between all cluster points, given indicies of clusters to compare
  euclid_dist.min <- function(pair_idxs) {
    min(apply(expand.grid(clusters[[ pair_idxs[[1]] ]], clusters[[ pair_idxs[[2]] ]]), 1, euclid_dist))
  }

  return(euclid_dist.min)
}

run_heir(euclid_dist.min.wrapper)
```

```
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  1"
## [1] "Cluster 2 :  4"
## [1] "Cluster 3 :  9"
## [1] "Cluster 4 :  16"
## [1] "Cluster 5 :  25"
## [1] "Cluster 6 :  36"
## [1] "Cluster 7 :  49"
## [1] "Cluster 8 :  64"
## [1] "Cluster 9 :  81"
## [1] "=================="
## [1] "Minimum dist:  3"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4)"
## [1] "Cluster 2 :  9"
## [1] "Cluster 3 :  16"
## [1] "Cluster 4 :  25"
## [1] "Cluster 5 :  36"
## [1] "Cluster 6 :  49"
## [1] "Cluster 7 :  64"
## [1] "Cluster 8 :  81"
## [1] "=================="
## [1] "Minimum dist:  5"
```

```
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  16"
## [1] "Cluster 3 :  25"
## [1] "Cluster 4 :  36"
## [1] "Cluster 5 :  49"
## [1] "Cluster 6 :  64"
## [1] "Cluster 7 :  81"
## [1] "=================="
## [1] "Minimum dist:  7"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16)"
## [1] "Cluster 2 :  25"
## [1] "Cluster 3 :  36"
## [1] "Cluster 4 :  49"
## [1] "Cluster 5 :  64"
## [1] "Cluster 6 :  81"
## [1] "=================="
## [1] "Minimum dist:  9"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  36"
## [1] "Cluster 3 :  49"
## [1] "Cluster 4 :  64"
## [1] "Cluster 5 :  81"
## [1] "=================="
## [1] "Minimum dist:  11"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25, 36)"
## [1] "Cluster 2 :  49"
## [1] "Cluster 3 :  64"
## [1] "Cluster 4 :  81"
## [1] "=================="
## [1] "Minimum dist:  13"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25, 36, 49)"
## [1] "Cluster 2 :  64"
## [1] "Cluster 3 :  81"
## [1] "=================="
## [1] "Minimum dist:  15"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25, 36, 49, 64)"
## [1] "Cluster 2 :  81"
## [1] "=================="
## [1] "Minimum dist:  17"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25, 36, 49, 64, 81)"
## [1] "=================="
```

(b) The average of the distances between pairs of points, one from each of the two clusters.

```
euclid_dist.avg.wrapper <- function(clusters) {
  # Find average euclidean distance between all cluster points, given indicies of clusters to compare
  euclid_dist.avg <- function(pair_idxs) {
    mean(apply(expand.grid(clusters[[ pair_idxs[[1]] ]], clusters[[ pair_idxs[[2]] ]]), 1, euclid_dist))
  }

  return(euclid_dist.avg)
}

run_heir(euclid_dist.avg.wrapper)
```

```
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  1"
## [1] "Cluster 2 :  4"
## [1] "Cluster 3 :  9"
## [1] "Cluster 4 :  16"
## [1] "Cluster 5 :  25"
## [1] "Cluster 6 :  36"
## [1] "Cluster 7 :  49"
## [1] "Cluster 8 :  64"
## [1] "Cluster 9 :  81"
## [1] "=================="
## [1] "Minimum dist:  3"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4)"
## [1] "Cluster 2 :  9"
## [1] "Cluster 3 :  16"
## [1] "Cluster 4 :  25"
## [1] "Cluster 5 :  36"
## [1] "Cluster 6 :  49"
## [1] "Cluster 7 :  64"
## [1] "Cluster 8 :  81"
## [1] "=================="
## [1] "Minimum dist:  6.5"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  16"
## [1] "Cluster 3 :  25"
## [1] "Cluster 4 :  36"
## [1] "Cluster 5 :  49"
## [1] "Cluster 6 :  64"
## [1] "Cluster 7 :  81"
## [1] "=================="
## [1] "Minimum dist:  9"
## [1] "Merging cluster idx 2  and cluster idx 3"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  c(16, 25)"
## [1] "Cluster 3 :  36"
## [1] "Cluster 4 :  49"
```

```
## [1] "Cluster 5 :  64"
## [1] "Cluster 6 :  81"
## [1] "=================="
## [1] "Minimum dist:  13"
## [1] "Merging cluster idx 3  and cluster idx 4"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9)"
## [1] "Cluster 2 :  c(16, 25)"
## [1] "Cluster 3 :  c(36, 49)"
## [1] "Cluster 4 :  64"
## [1] "Cluster 5 :  81"
## [1] "=================="
## [1] "Minimum dist:  15.8333333333333"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  c(36, 49)"
## [1] "Cluster 3 :  64"
## [1] "Cluster 4 :  81"
## [1] "=================="
## [1] "Minimum dist:  17"
## [1] "Merging cluster idx 3  and cluster idx 4"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  c(36, 49)"
## [1] "Cluster 3 :  c(64, 81)"
## [1] "=================="
## [1] "Minimum dist:  30"
## [1] "Merging cluster idx 2  and cluster idx 3"
## [1] "CLUSTERS:"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25)"
## [1] "Cluster 2 :  c(36, 49, 64, 81)"
## [1] "=================="
## [1] "Minimum dist:  46.5"
## [1] "Merging cluster idx 1  and cluster idx 2"
## [1] "Cluster 1 :  c(1, 4, 9, 16, 25, 36, 49, 64, 81)"
## [1] "=================="
```