

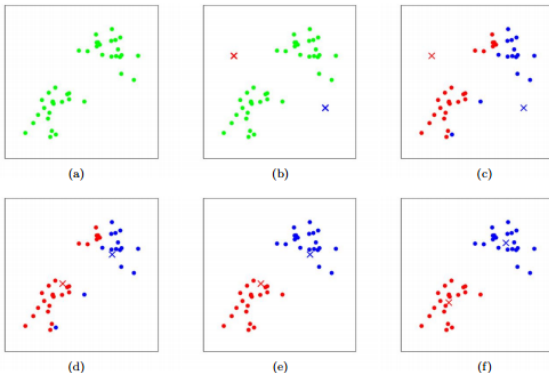
A2 Overview

Chris Kauffman

*Last Updated:
Thu Mar 31 09:35:56 AM CDT 2022*

K-Means Clustering

- ▶ A standard ML / Data Mining / Stats problem
- ▶ Input: data + #of clusters desired
- ▶ Output: assignment of each data to a cluster + cluster centers
- ▶ Algorithm: Iterates between
 1. Calculate cluster centers
 2. Calculate cluster assignments



Source: K-Means by Chris Piech. Based on a handout by Andrew Ng.

Determine Cluster Centers

```
1 # DETERMINE NEW CLUSTER CENTERS
2 for c in range(clust.nclust):      # reset cluster centers to 0.0
3     for d in range(clust.dim):
4         clust.features[c][d] = 0.0
5
6 for i in range(data.ndata):        # sum up data in each cluster
7     c = data.assigned[i]
8     for d in range(clust.dim):
9         clust.features[c][d] += data.features[i][d]
10
11 for c in range(clust.nclust):      # divide by ndatas of data to
12     if clust.counts[c] > 0:         # get mean of cluster center
13         for d in range(clust.dim):
14             clust.features[c][d] = clust.features[c][d] / clust.counts[c]
```

Determine Cluster Assignments

```
1 # DETERMINE NEW CLUSTER ASSIGNMENTS FOR EACH DATA
2 for c in range(clust.nclust):      # reset cluster counts to 0
3     clust.counts[c] = 0
4
5 nchanges = 0
6 for i in range(data.ndata):        # iterate over all data
7     best_clust = None
8     best_distsq = float("inf")
9     for c in range(clust.nclust):  # compare data clusters, assign closest
10         distsq = 0.0
11         for d in range(clust.dim): # calculate squared distance to each dim
12             diff = data.features[i][d] - clust.features[c][d]
13             distsq += diff*diff
14         if distsq < best_distsq:    # if closer to this cluster than
15             best_clust = c          # current best
16             best_distsq = distsq
17     clust.counts[best_clust] += 1
18     if best_clust != data.assigs[i]: # assigning data to a different cluster?
19         nchanges += 1               # indicate cluster assignment has changed
20     data.assigs[i] = best_clust     # assign to new cluster
```

Overall

```
# ASSIGN RANDOM CLUSTER TO EACH DATA
...

maxiter = 100          # bounds the iterations
curiter = 1            # current iteration
nchanges = ndata       # count changes in assignment each iter

while nchanges > 0 and curiter <= maxiter: # loop until convergence
    # DETERMINE NEW CLUSTER CENTERS
    ...

    # DETERMINE NEW CLUSTER ASSIGNMENTS FOR EACH DATA
    ...
```

Distributed Memory Parallel Versions

- ▶ Algorithm deals with Data and Clusters, each a matrixy thing
- ▶ How would you divide up this data in a distributed parallel version?
- ▶ Would data redistribution be required in your scheme?
- ▶ What information needs to be exchanged at each iteration?

Shared Memory Parallel Versions

- ▶ Determine which loops Can and Should be parallelized in a shared memory system
- ▶ Is any coordination required for loop iterations? Reductions needed?
- ▶ Suggest some places to put OpenMP `#pragma` directives