# Homework 3 Data Mining *

Alex Clemmer

March 7, 2012

## 1 Hierarchical Clustering

**A:** Mean link probably works the best on large data, since you don't have to compute pairwise norms. It is hard to say which worked "best", since that depends on your measurement (*i.e.*, inter-cluster similarity, extra-cluster similarity, whether it looks "good", etc.). That said, it is likely that the min link clustering produced the best-looking results, since the points that should be clustered together tend to be closest to the clusters they should be in.

*Single-Link Clustering:*

```
CLUST
('m', (14.02, 5.03))
('n', (16.05, 5.01))
CLUST
('a', (4.01, 15.021))
('b', (3.02, 14.031))
('c', (2.99, 12.02))
('d', (3.107, 10.04))
('e', (3.08, 8.05))
('f', (3.04, 7.03))
('g', (4.01, 6.06))
('h', (5.07, 5.06))
('i', (6.02, 4.03))
('j', (8.06, 4.09))
('k', (10.02, 4.08))
('l', (12.01, 4.07))
CLUST
('o', (12.54, 12.51))
('p', (12.03, 12.04))
('q', (11.52, 11.57))
('r', (11.03, 11.09))
('s', (10.51, 10.532))
('t', (10.01, 10.01))
('u', (12.5, 15.52))
('v', (12.06, 15.1))
('w', (11.55, 14.57))
```

*CS 6955 Data Mining; Spring 2012 Instructor: Jeff M. Phillips, University of Utah

```
('x', (11.08, 14.3))
('y', (10.52, 13.53))
('z', (10.03, 13.008))
```

*Complete-Link Clustering:*

```
CLUST
('e', (3.08, 8.05))
('f', (3.04, 7.03))
('g', (4.01, 6.06))
('h', (5.07, 5.06))
('i', (6.02, 4.03))
CLUST
('j', (8.06, 4.09))
('k', (10.02, 4.08))
('l', (12.01, 4.07))
('m', (14.02, 5.03))
('n', (16.05, 5.01))
CLUST
('a', (4.01, 15.021))
('b', (3.02, 14.031))
('c', (2.99, 12.02))
('d', (3.107, 10.04))
('o', (12.54, 12.51))
('p', (12.03, 12.04))
('q', (11.52, 11.57))
('r', (11.03, 11.09))
('s', (10.51, 10.532))
('t', (10.01, 10.01))
('u', (12.5, 15.52))
('v', (12.06, 15.1))
('w', (11.55, 14.57))
('x', (11.08, 14.3))
('y', (10.52, 13.53))
('z', (10.03, 13.008))
```

*Average-Link Clustering:*

```
CLUST
('o', (12.54, 12.51))
('p', (12.03, 12.04))
('q', (11.52, 11.57))
('r', (11.03, 11.09))
('s', (10.51, 10.532))
('t', (10.01, 10.01))
('u', (12.5, 15.52))
('v', (12.06, 15.1))
```

```
('w', (11.55, 14.57))
('x', (11.08, 14.3))
('y', (10.52, 13.53))
('z', (10.03, 13.008))
CLUST
('j', (8.06, 4.09))
('k', (10.02, 4.08))
('l', (12.01, 4.07))
('m', (14.02, 5.03))
('n', (16.05, 5.01))
CLUST
('a', (4.01, 15.021))
('b', (3.02, 14.031))
('c', (2.99, 12.02))
('d', (3.107, 10.04))
('e', (3.08, 8.05))
('f', (3.04, 7.03))
('g', (4.01, 6.06))
('h', (5.07, 5.06))
('i', (6.02, 4.03))
```

*Average-Link Clustering:*

```
CLUST
('o', (12.54, 12.51))
('p', (12.03, 12.04))
('q', (11.52, 11.57))
('r', (11.03, 11.09))
('s', (10.51, 10.532))
('t', (10.01, 10.01))
('u', (12.5, 15.52))
('v', (12.06, 15.1))
('w', (11.55, 14.57))
('x', (11.08, 14.3))
('y', (10.52, 13.53))
('z', (10.03, 13.008))
CLUST
('j', (8.06, 4.09))
('k', (10.02, 4.08))
('l', (12.01, 4.07))
('m', (14.02, 5.03))
('n', (16.05, 5.01))
CLUST
('a', (4.01, 15.021))
('b', (3.02, 14.031))
('c', (2.99, 12.02))
('d', (3.107, 10.04))
('e', (3.08, 8.05))
```

```
('f', (3.04, 7.03))
('g', (4.01, 6.06))
('h', (5.07, 5.06))
('i', (6.02, 4.03))
```

**B:** Begin with the definition of the average-link cluster:

$$\frac{1}{|S_1||S_2|}\sum_{(s_1,s_2)\in S_1\times S_2}||s_1-s_2||_2 = \frac{1}{|S_1||S_2|}\sum_{(s_1,s_2)\in S_1\times S_2}\sqrt{(s_1-s_2)^2} \tag{1}$$

$$= \frac{1}{|S_1||S_2|}\sqrt{(\sum_{s_1\in S_1} s_1 - \sum_{s_2\in S_2} s_2)^2} \tag{2}$$

This is clearly equivalent to the mean cluster, as long as it's $\mathbb{R}^1$.

# 2 Point Assignment Clustering

*MinMax:* Clusters centers are the first level of indent; the items clustered around them are indented four spaces.

```
('g', (2.0, 3.0))
    ('g', (2.0, 3.0))
    ('i', (2.1, 3.8))
    ('h', (2.4, 3.4))
    ('k', (2.6, 3.5))
    ('j', (2.9, 3.1))
    ('m', (2.2, 3.9))
    ('l', (2.5, 3.6))
    ('o', (2.3, 3.4))
    ('n', (2.7, 3.5))
    ('q', (1.1, 7.8))
    ('p', (1.0, 7.0))
    ('s', (1.5, 7.9))
    ('r', (1.2, 7.3))
    ('u', (1.5, 7.1))
    ('t', (1.3, 7.2))
    ('w', (1.9, 7.7))
    ('v', (1.4, 7.2))
    ('y', (1.6, 7.8))
    ('x', (1.3, 7.1))
('z', (16.0, 2.0))
    ('z', (16.0, 2.0))
('a', (11.0, 12.0))
    ('a', (11.0, 12.0))
    ('c', (11.8, 12.6))
    ('b', (11.2, 12.3))
    ('e', (11.3, 12.4))
```

```
    ('d', (11.1, 12.9))
    ('f', (11.7, 12.1))
```

*KMeans++* The same notation is used.

```
('g', (2.0, 3.0))
    ('g', (2.0, 3.0))
    ('i', (2.1, 3.8))
    ('h', (2.4, 3.4))
    ('k', (2.6, 3.5))
    ('j', (2.9, 3.1))
    ('m', (2.2, 3.9))
    ('l', (2.5, 3.6))
    ('o', (2.3, 3.4))
    ('n', (2.7, 3.5))
    ('q', (1.1, 7.8))
    ('p', (1.0, 7.0))
    ('s', (1.5, 7.9))
    ('r', (1.2, 7.3))
    ('u', (1.5, 7.1))
    ('t', (1.3, 7.2))
    ('w', (1.9, 7.7))
    ('v', (1.4, 7.2))
    ('y', (1.6, 7.8))
    ('x', (1.3, 7.1))
('z', (16.0, 2.0))
    ('z', (16.0, 2.0))
('a', (11.0, 12.0))
    ('a', (11.0, 12.0))
    ('c', (11.8, 12.6))
    ('b', (11.2, 12.3))
    ('e', (11.3, 12.4))
    ('d', (11.1, 12.9))
    ('f', (11.7, 12.1))
```

**B:** The average distance of a point $p$ to its closest centroid $c$ is given by:

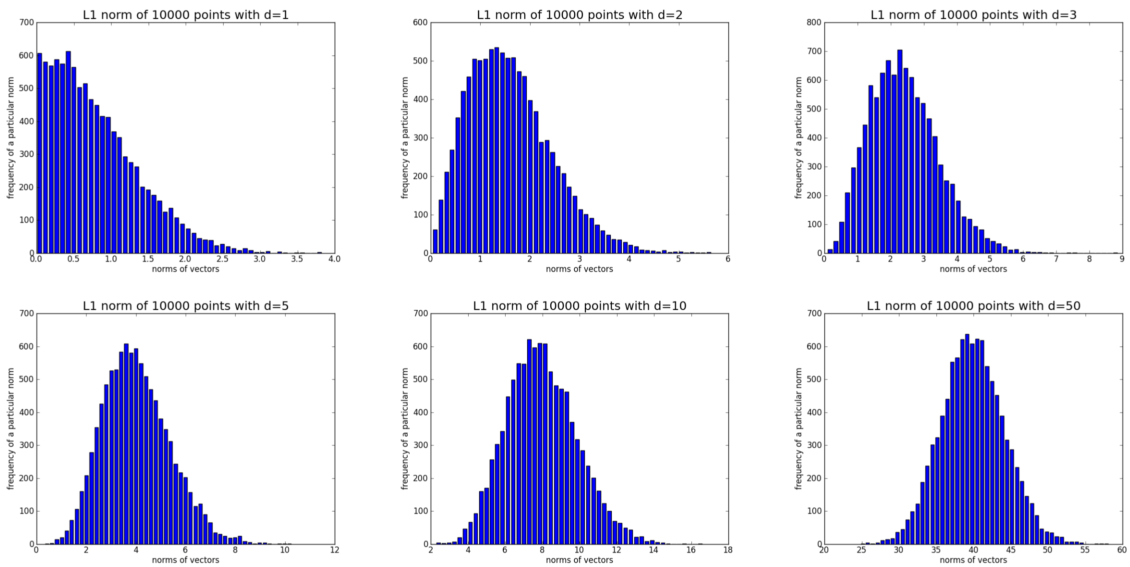
$$\frac{\sum_{p \in P} d(p, [\arg\min_{c \in C} d(p, c)])}{|P|} \tag{3}$$

The average distance of a point to its closest cluster should always decrease if we are converging. If a cluster moves in some direction, then the gain in average distance loss must be greater than the loss incurred by staying. By the principle of contraction, the average distance gained by any points lost by the cluster in the next evaluation step are less then the average distance lost.

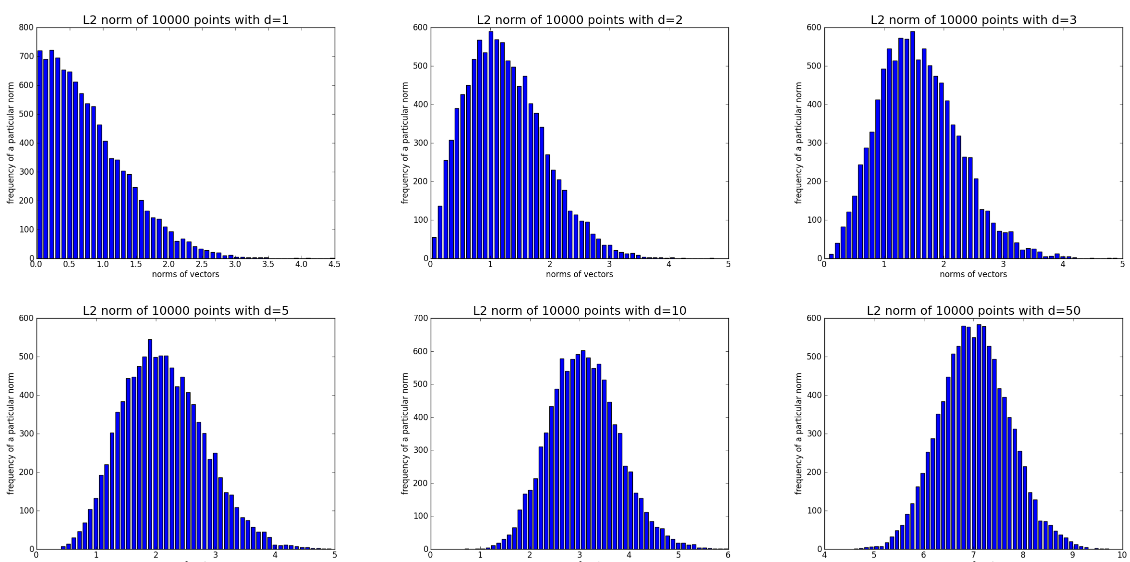
Since the average distance of a point to the closest cluster always decreases, we converge.

# 3 High Dimensions

**A:** I used $n = 10,000$ for my histograms. **First**, the histograms for the $L_1$ norms for increadingly high dimensionality $d$:




And second, the histograms for the $L_2$ norms on increasingly high dimensionality $d$:




**B:** I estimated the probability by generating $n = 10,000$ samples, norming them, and calculating at the empirical $\mathbf{Pr}(L_P(\vec{x}) < 1)$ for all $\vec{x}$ in my generated observations.

The tables are broken apart by dimension $d$ (*i.e.*, the first table is for $d = 1$, the second for $d = 2$, and so on). The *index* column is just the number of the experiment—it's useful because there should be 30 in all, which we can plainly see. The $P$ column refers to which norm $L_P$ we're using, *e.g.*, $L_1$, $L_2$, and so on.

Finally, (and most importantly) the last column refers to the empirical probability that the norm is less 1. We calculate this by taking the number of samples for which this condition is true, and dividing that number by the total.

In answer to the second part of the question as $P$ approaches 0, we expect this probability also to approach 0.

For $d = 1$:

| *index* | $d$ | $P$ in $L_P$ | $\mathbf{Pr}(< 1)$ |
|---|---|---|---|
| 1 | 1 | 0.5 | 1.000000 |
| 2 | 1 | 1 | 1.000000 |
| 3 | 1 | 2 | 1.000000 |
| 4 | 1 | 3 | 1.000000 |
| 5 | inf | 3.0 | 1.000000 |

For $d = 2$:

| *index* | $d$ | $P$ in $L_P$ | $\mathbf{Pr}(< 1)$ |
|---|---|---|---|
| 6 | 2 | 0.5 | 0.166200 |
| 7 | 2 | 1 | 0.501400 |
| 8 | 2 | 2 | 0.785800 |
| 9 | 2 | 3 | 0.883900 |
| 10 | inf | 3.0 | 0.883900 |

For $d = 3$:

| *index* | $d$ | $P$ in $L_P$ | $\mathbf{Pr}(< 1)$ |
|---|---|---|---|
| 11 | 3 | 0.5 | 0.011600 |
| 12 | 3 | 1 | 0.158100 |
| 13 | 3 | 2 | 0.523600 |
| 14 | 3 | 3 | 0.706900 |
| 15 | inf | 3.0 | 0.706900 |

For $d = 5$:

| *index* | $d$ | $P$ in $L_P$ | $\mathbf{Pr}(< 1)$ |
|---|---|---|---|
| 16 | 5 | 0.5 | 0.000000 |
| 17 | 5 | 1 | 0.008400 |
| 18 | 5 | 2 | 0.160200 |
| 19 | 5 | 3 | 0.369700 |
| 20 | inf | 3.0 | 0.369700 |

For $d = 10$:

| *index* | $d$ | $P$ in $L_P$ | $\mathbf{Pr}(< 1)$ |
|---|---|---|---|
| 21 | 10 | 0.5 | 0.000000 |
| 22 | 10 | 1 | 0.000000 |
| 23 | 10 | 2 | 0.002400 |
| 24 | 10 | 3 | 0.035600 |
| 25 | inf | 3.0 | 0.035600 |

For $d = 50$:

| *index* | $d$ | $P$ in $L_P$ | $\mathbf{Pr}(< 1)$ |
|---|---|---|---|
| 26 | 50 | 0.5 | 0.000000 |
| 27 | 50 | 1 | 0.000000 |
| 28 | 50 | 2 | 0.000000 |
| 29 | 50 | 3 | 0.000000 |
| 30 | inf | 3.0 | 0.000000 |