CS4786/5786: Machine Learning for Data Science, Spring 2015
2/26/2015: Lecture 9 handout: Introduction to clustering

---

**Announcements**

1. Speed/memory problems: note that if you only need to project into $K$ dimensions, you don't need to compute all eigenvectors, but just the top $K$; consider `eigs` for Matlab, ~~`numpy`~~`scipy.linalg.eigh` in Python. If you decide to use the SVD instead, as has been alluded to in class and on Piazza, make sure you understand what matrix you should apply the SVD to. (Data = rows or columns? Center the data first? etc.) (Piazza @45 followup, "Numpy very slow")

2. A PCA example in C++ has been posted to the lecture 3 materials.

3. Homework updates (will be propagated to HW posted online some time today)

   (a) A1 Q1.2 $Y_I$ and $Y_{II}$ should be equal *up to sign*, as opposed to "strictly equal", as indicated by the subsequent assignment question.

   (b) A1 Q1.3: you may, for simplicity, assume that the $\mathbf{x}_t$s are centered; however, this condition is not strictly necessary. (Piazza @52)

**Selected clustering optimization functions** Assume we have $n$ data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. (When possible, we'll avoid using indices and refer to an arbitrary member of this set as $\mathbf{x}$.) Assume we have a partitioning of the data points into $K$ *clusters* $C_1, \ldots, C_K$, which we'll index by $j$.[1] For each cluster $C_j$, we write $n_j$ for the number of $\mathbf{x}$s in $C_j$.

Some of these should be maximized and some of these should be minimized. Can you convince yourself which is which?

| | | | |
|---|---|---|---|
| (1) | Within-cluster scatter | $\sum_j \sum_{\mathbf{x}_t, \mathbf{x}_s \in C_j, t<s} \|\|\mathbf{x}_t - \mathbf{x}_s\|\|_2^2$ | |
| (2) | Within-cluster variation | $\sum_j n_j \sum_{\mathbf{x} \in C_j} \|\|\mathbf{x} - \mathbf{r}_j\|\|_2^2$ | centroid $\mathbf{r}_j \stackrel{def}{=} \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$ |
| (3) | Between-cluster scatter | ~~$\sum_j$~~ $\sum_{\mathbf{x}_t, \mathbf{x}_s \text{ in different clusters}, t<s} \|\|\mathbf{x}_t - \mathbf{x}_s\|\|_2^2$ | |
| (4) | Between-cluster variation | $\sum_j n_j (\mathbf{r}_j - \mu)(\mathbf{r}_j - \mu)^T$ | $\mu = \frac{1}{n} \sum_{\mathbf{x}} \mathbf{x}$ |
| (5) | Trace | $\sum_j n_j \mathrm{tr}(\boxed{\Sigma_j})$ | $\boxed{\Sigma_j}$ : covariance for $C_j$<br>$\mathrm{tr}(M) \stackrel{def}{=} \sum_i M[i,i]$ |
| (6) | "best-friend" | $\sum_j \max_{\mathbf{x} \in C_j} \min_{\mathbf{x}' \neq \mathbf{x} \in C_j} \{\|\|\mathbf{x} - \mathbf{x}'\|\|_2^2\}$ | |

Post-lecture update: Actually, the "best-friend" criterion has to be modified to handle the case of clusters containing a single element. More on this next lecture.

---

[1] Using $k$ as index variable might seem like a good idea, but my $k$s and $K$s are indistinguishable on the board.

**Clicker question** For any set of points $\mathbf{x}_1, \ldots, \mathbf{x}_N$, let $\mathbf{r} = \frac{1}{N}\sum_t \mathbf{x}_t$ be the *centroid* of the points.
Consider the following two assertions: For any point $\mathbf{z}$,

$$(7) \qquad \sum_t ||\mathbf{x}_t - \mathbf{z}||_2^2 = \left(\sum_t ||\mathbf{x}_t - \mathbf{r}||_2^2\right) + ||\mathbf{r} - \mathbf{z}||_2^2$$

$$(8) \qquad \sum_t ||\mathbf{x}_t - \mathbf{z}||_2^2 = \left(\sum_t ||\mathbf{x}_t - \mathbf{r}||_2^2\right) + N||\mathbf{r} - \mathbf{z}||_2^2$$

A. Only (7) is true
B. Only (8) is true
C. Both are true
D. Neither are true, by triangle inequality
E. I really don't know

**k-means algorithm** Start with some initialization $\mathbf{r}_j^0$, $j \in 1, \ldots, K$ (superscripts = iteration number, we start with $i = 0$). Repeat until "convergence":

1. Assign each $\mathbf{x}$ to its nearest *representative* $\mathbf{r}_j^i$.
2. Set $\mathbf{r}_j^{i+1}$ to be the centroid of the $\mathbf{x}$s now assigned to it.
3. Increment $i$