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, `numpyscipy.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 $x_t$s are centered; however, this condition is not strictly necessary. (Piazza @52)

Selected clustering optimization functions  
Assume we have $n$ data points $x_1, \ldots, x_n$. (When possible, we’ll avoid using indices and refer to an arbitrary member of this set as $x$.) Assume we have a partitioning of the data points into $K$ clusters $C_1, \ldots, C_K$, which we’ll index by $j$. For each cluster $C_j$, we write $n_j$ for the number of $x$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

\[
\sum_j \sum_{x_t, x_s \in C_j, t < s} ||x_t - x_s||^2
\]

(2) Within-cluster variation

\[
\sum_j n_j \sum_{x \in C_j} ||x - r_j||^2
\]

Within-cluster scatter \[
\sum_j \max_{x \in C_j} \min_{x' \notin C_j} \{|x - x'|^2\}
\]

(4) Between-cluster variation

\[
\sum_j n_j (r_j - \mu) (r_j - \mu)^T
\]

(5) Trace

\[
\sum_j n_j \text{tr}(\Sigma_j)
\]

(6) “best-friend”

\[
\sum_j \max_{x \in C_j} \min_{x' \neq x \in C_j} \{|x - x'|^2\}
\]

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.

\[\text{tr}(M) \overset{\text{def}}{=} \sum_i M[i,i]\]
Clicker question  For any set of points $x_1, \ldots, x_N$, let $r = \frac{1}{N} \sum_t x_t$ be the centroid of the points.
Consider the following two assertions: For any point $z$,

(7) \[ \sum_t ||x_t - z||^2 = \left( \sum_t ||x_t - r||^2 \right) + ||r - z||^2 \]

(8) \[ \sum_t ||x_t - z||^2 = \left( \sum_t ||x_t - r||^2 \right) + N||r - z||^2 \]

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 $r_j^0 : j \in 1, \ldots, K$ (superscripts = iteration number, we start with $i = 0$). Repeat until “convergence”:

1. Assign each $x$ to its nearest representative $r_j^i$.

2. Set $r_j^{i+1}$ to be the centroid of the $x$ now assigned to it.

3. Increment $i$