• <50 people @ beginning of class! Might have got to 70 by the end.

Lec 9, 2/26/15
Intro to clustering

1 handout; have clickers out

Announcements. Remember to update the HW.

Partners: — for A! ask q. w/ ppt.

(w/ clickers)

C: want to join a group. ~~need a partner~~

A: in "group" that's happy to take one more ~~wants one~~

E: none of the above

(w/out clickers) — raise hands.

	w/ clickers	w/out clickers
A	6	+4
C	4	+2

so, should be able to match people up: ask them to meet up by back door after class.

note to self: give #'s to points drawn on board!

Outline: ~~from representations dimens~~

we've talked about dimension reduction in the feature space via projections

given data matrix $X: n \times d$ ~~[illegible]~~, objects in d dimensions.

produce ~~new~~ projected data matrix that is $n \times K$, $K \ll d$.

- this helps us improve our features when they are "wrong".
- reducing # of features.

Today: another kind of "dimension reduction":

given n datapoints, "condense" into K clusters of similar points.
(K ↑ predetermined) ~~(for now, assume~~

so, reducing # of objects.

- ~~diff ways to choose "good" clusters characteriz~~
- diff ways to characterize "good" clusters.
- < an algorithm for "one" of these characterizations >

↳ we didn't get here ~~altho~~ (which is good! That would have been too break-neck of a pace).

- clusters: let's consider as: fixed k, partition points into k disjoint groups.
  ↳ each part is disjoint from the others, and their union is all the points

  - "compression": n datapoints → k
  "visualization"/understanding:
    - better than a pointcloud
    - people like understanding things as groups of things.
      - biology taxonomy (animals vs plants, canines vs. felines vs. marsupials), → as opposed to "fluffy" and "grumpy"
      - market segmentation
        - what kind of customers are there? < credit-worthy?
        - what kinds of students are there? < good vs need help

(e 1st: what could this not include?)

[this is a broad, but not universal paradigm:
- fixed k → try to "discover" the right k?
- partition → let items belong to more than one type.
But we won't consider these for now.]

---

selected qt fs: take k=2.

ask: exaple of a good partition vs. bad clustering for (1)

ask: for (6)
(should they get the same answers?)

outline of next section.

: ~~play around~~/ ~~what we can~~ understand

- ask: does (1) have a diff. solution than (3)?

~~use lem~~

assum lemma, ~~show pf~~. mention as the consequence. (but don't prove it).

↓

and so (1); (2) are the same.

prove the lemma.

Talk about k-means.

---

vocabulary items: a clustering consists of a set of clusters

→ handout: we a # of ways to evaluate a clustering, to decide it's "good"
(i.e., to define the goal of clustering).

You should see 6 functions there. Everyone see 6?

↘ why am I asking? (again, telegraphing what's going to happen...)

Important skill, as <I think> we've mentioned before: given two proposals, can you tell when they're "the same" and when they're "different"? ←

< We've set up some notation. I tried to avoid using indices on the data points, ~~to~~ but various types of double-counting occurs if you specify index pairs as $(s \neq t)$ as opposed to $s < t$. >

- You guys look @ (1) ~~ok~~, and while I'm writing on the board, try to decide why it's called the "within-cluster scatter"

Fix k=2 throughout (just 2 clusters)

↑ today

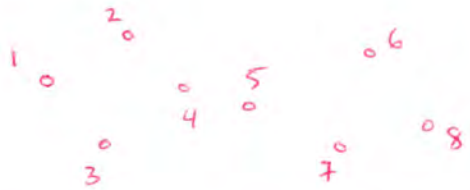
$\sum_j (\quad)$

j ↖ means we're going to break this objective into a score for each individual cluster.

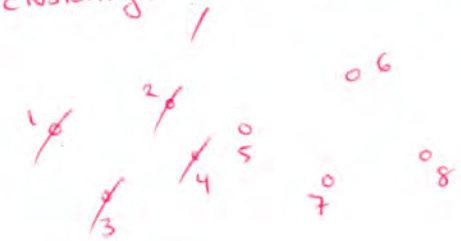
$\sum_{x_t, x_s \in C_j, t<s}$ → sum over all the ways to pick to different points in cluster C_j.

$\|x_t - x_s\|_2^2$

- $\|\cdot\|_2$ → Euclidean distance
- $(\cdot)^2$ → square is convenient and important.
- $\|x_t - x_s\|$ → distance btwn that pair of points.



A sample clustering: "x" in the 1st cluster, "no x" means ~~not~~ in the other cluster.



~~"pay"~~ for

(1,2)	(5,6)
(1,3)	(5,7)
(1,4)	(5,8)
(2,3)	(6,7)
(2,4)	(6,8)
(3,4)	(7,8)

this should do "not bad" according to (1). but not, for example, (4,5).

q: wouldn't it be better to "strike" 5 as well?

a: good question. (I didn't say this is optimal)
It removes (5,6), (5,7), (5,8) from contributing to the score,
but adds in (1,5), (2,5), (3,5), (4,5).
... and I'm not gonna compute in my head which of these is actually better, but you get the idea.

q: so does this like clusters that are the same size?

a: exactly, b/c of how each point-pair contributes to the score.

(1) "likes" <to anthropomorphize> similar-size, spherical clusters.

et's jump to (6) now, which has not <as of ~~time of~~ time of lecture> been given a name.

Does this "like" the same things?

(lots of consultation among the students.
No volunteers for giving an intuitive name).

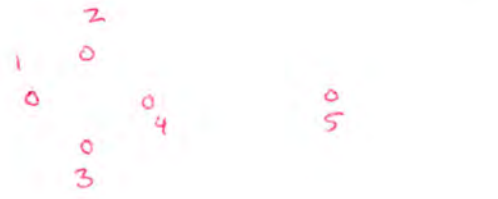
parsing: $\sum_j ($ $)$

= breaking into per-cluster scores

$\max_{x \in C_j}$ ↖ let's spse we pick a particular x.

$\min_{x' \neq x \text{ in } C_j} \|x - x'\|_2^2$

↳ ok to have "duplicates" pairs b/c we're taking a min.



- closest thing to 1 is, say 2.
  + (1,2) distance
- closest thing to 2 is, say 1,
  + about same
- closest thing to 3
  + about same
- closest thing to 4
  + about same
- closest thing to 5 = 4

bigger quantity gets added in

~~total cost~~.

this is the max.

~~what~~

likes: clusters where each point has something in the cluster that's close to it

even if there is also something in the cluster that's far away from it.

(6) "likes" much better than (1) likes.

So (6) $\not\equiv$ (1).

q: can we call this 'single-link'

a: it's true that this is the criterion that underlies the single-link clustering alg.

Now let's consider (3); my question is: do you think it differs from (1)?

While I write the clicker responses on board, you ponder:

Are ~~Do~~ (1); (3) <to maximize> ~~have the same~~ equivalent?

↓

<to minimize>

clickers

(A) Yes! 19 (46%)

(B) No!! 16 (39%)

(C) I don't know! 6 (15%)

- ask non-clickers to predict majority answer, in a pseudo Bayesian-truth-serum experiment
- good participation.

Followup question: how many tried to answer this w/ a math. proof? (two hands!).

Ask for representative reason for (A):

(A) if you move a point to a diff. cluster, what was an intra-cluster distance becomes a between-cluster ~~diff~~ distance, and vice versa.

Ask for representative reason for (B)

(B) You could make the clusters far apart without making the within-scatter small.

I think both intuitions are reasonable on their face, so let's try a proof <or @ least resort to some math>:

~~What is (1)~~

What is (1)+(3)? Total sum of distances between pairs of objects

↓

defined as (x_t, x_s) where $t < s$

- if we wrote $t \neq s$, we would double-count.

(It's important to get a feeling of what differences matter and what don't. Here probably doesn't change the optima.)

a constant! across diff. clusterings.

So, if clustering1 improves (1) compared to clustering2

then .. .. (3)

also: as (1) changes, (3) changes the other way

SO: one pair of (1)-(6) turns out to be equivalent

<not a lot of time left>

Let's now talk about (5), ~~b/c this allows us to draw a connection with the projection unit we~~ and how it relates to (2)

- (said) this allows us to draw a connection with the unit on projection that we just finished
- (said later: also, a technique for coming up w/ other criteria that might be useful)
- (didn't say) it ~~is~~ - this is the ~~pedagogical~~ 'attack' in order to move from (1) which seems like a very general criterion – to (2)
- (the reappearance of the covariance matrix!)

- notational note: you can see why it's helpful to put the box around the cov. matrix: there's a Σ_j that means something else in the same statement! (in class I said "right to the left", which is pretty absurd...)

Once again, this breaks down into a term for each cluster; here the terms involve the trace of the covariance matrix just for the data items in cluster C_j.

~~it's the~~ it's the sum...
<some diagonals swipes from audience>
of the items on the diagonals.

q: since this is the covariance matrix, what are the diagonal entries?

a: <came from class> the variances. ~~scaled by~~ multiplicative factor, ~~entries are~~ diagonal entries are: $\frac{1}{n_j}\sum_t (x_t - r_j)^2$

q: and if we look at what the components of (2) are,

$= \|x_t - r_j\|_2^2$

So $\boxed{(2) = (5)}$.

So right, again we have the covariance matrix making an appearance.

Is that spooky? Will it haunt your dreams? :)

And note that a "natural" matrix function led to an informative clustering criterion.

So you could imagine trying out other things, like the determinant, and this is sometimes used.

Note: in class I said something totally false about the determinant. It is not invariant under linear transformation; I clearly had something else in mind, but no longer recall what. (Altho' there's a scale independence, apparently).

~~Let's now see if we can collapse some of these groupings~~

Now ~~we~~ would be a good idea to see if we can collapse some of these groupings.

(2) ≡ (5) (1) ≡ (3)

Let's look @ (1) & 2.

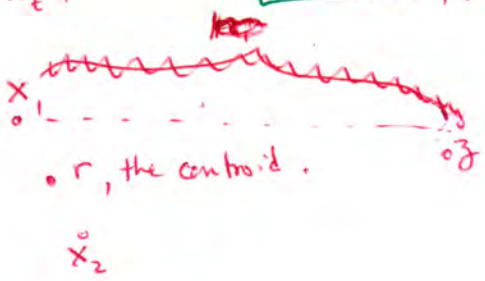
As a first step, let's see what you think of the clicker question on the back of your handout.

A: only 7 twc : 6 = 18%
B: only 8 twc 11 = 33%
C: both 1 = 3%
D: neither, by Δ in = 13 = 37%
E: don't know

~~only~~ (low # clicks, needed more time)

For cluster points x_t.

for ~~some~~ any point z



looks like the Δ in ~~∈~~

what fact (8) ~~let's~~ lets us do is relate (1) to (2)

↳ ~~more sensible/good~~ more general-looking criterion

↳ more apparently amenable to developing an algorithm for.