

DS2500
4/1-Tues

Admin

- no more HW!
- no more labs!

- exam #2 4/4

- second grace hw due 4/11 9pm

- project deadline 4/15 9pm

Agenda

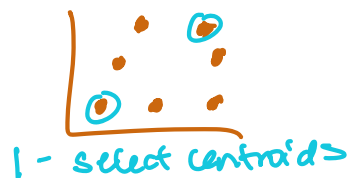
1. unsupervised learning
2. k-means clustering
3. Python

1. unsupervised learning

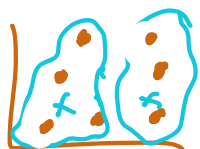
joinpd.com
eve jzw

2. K-means clustering

$K=2$



2. Assign points to closest centroid



3 - recompute centroids
arg F1, arg F2

repeat 2+3

For 3 Functions

- params?
- Pseudocode
- helper function?
- return?

1. Select Centroids

- DF subset (skip k=1)
- returns: mini DF with K rows

`df.sample()`

- Steps 1, 2, 3 are functions
- data in df

name	s	k	catch
—	—	—	—
—	—	—	—
—	—	—	—

- euclidean distance

centroids: mini DF with K rows and features

assy: centroid index #
0, 1, 2, ..., K-1

2. Assign points to closest centroid

- params: df, centroids
- for each row, compute distance to all centroids
- belongs to minimum $\rightarrow$ by index #

`dist()` $\rightarrow$ dist b/w df and a row

`reshape()` $\rightarrow$ for one row

`iterrows()` $\rightarrow$ iterate over all

3. Recompute centroids

- Compute mean of each feature within each class

`df.groupby()`

4. plot!

Z:36