Introduce to Machine Learning Program Assignment #2

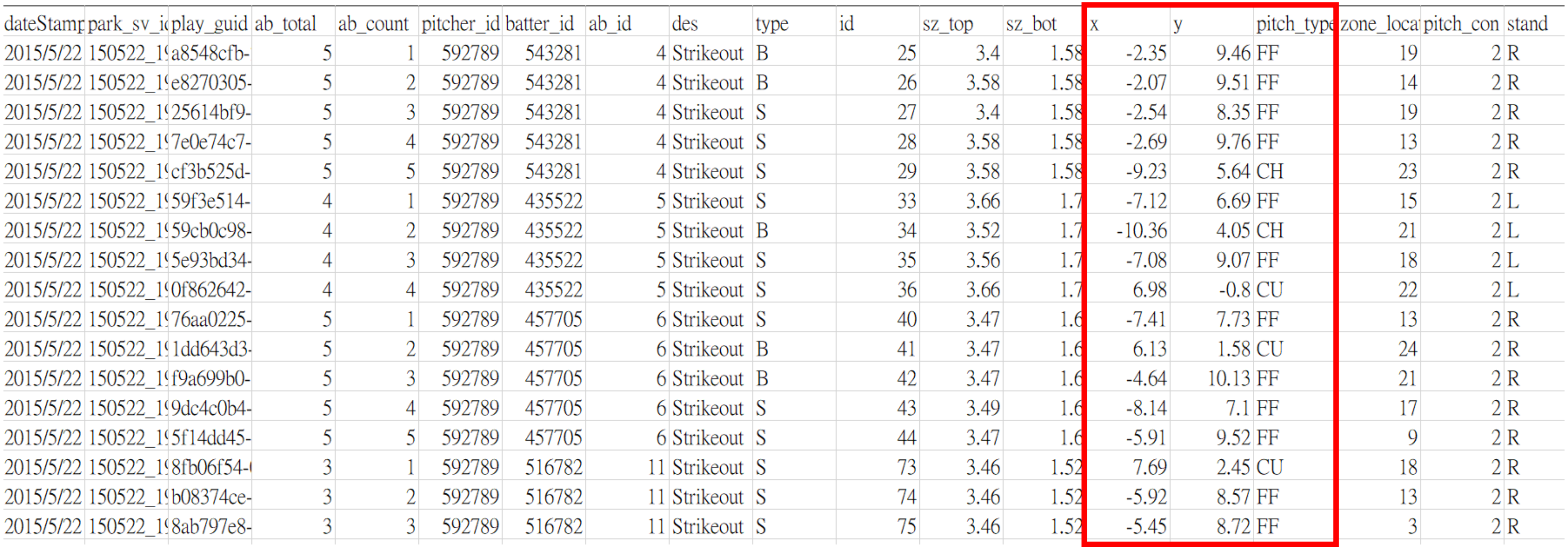
If you have any questions, send me an e-mail.

- TA's name: 林裕庭
- TA's email: kartglin.iie07g@nctu.edu.tw

K-means

You will get a dataset (**data_noah.csv**). It is Noah Syndergaard’s pitches that have been tracked by the PITCHf/x system in the MLB Regular Season.

X is horizontal movement ; y is vertical movement



dateStamp	park_sv_i	play_guid	ab_total	ab_count	pitcher_id	batter_id	ab_id	des	type	id	sz_top	sz_bot	x	y	pitch_type	zone_loca	pitch_con	stand
2015/5/22	150522_1	a8548cfb-	5	1	592789	543281	4	Strikeout	B	25	3.4	1.58	-2.35	9.46	FF	19	2	R
2015/5/22	150522_1	e8270305-	5	2	592789	543281	4	Strikeout	B	26	3.58	1.58	-2.07	9.51	FF	14	2	R
2015/5/22	150522_1	25614bf9-	5	3	592789	543281	4	Strikeout	S	27	3.4	1.58	-2.54	8.35	FF	19	2	R
2015/5/22	150522_1	7e0e74c7-	5	4	592789	543281	4	Strikeout	S	28	3.58	1.58	-2.69	9.76	FF	13	2	R
2015/5/22	150522_1	cf3b525d-	5	5	592789	543281	4	Strikeout	S	29	3.58	1.58	-9.23	5.64	CH	23	2	R
2015/5/22	150522_1	59f3e514-	4	1	592789	435522	5	Strikeout	S	33	3.66	1.7	-7.12	6.69	FF	15	2	L
2015/5/22	150522_1	59cb0c98-	4	2	592789	435522	5	Strikeout	B	34	3.52	1.7	-10.36	4.05	CH	21	2	L
2015/5/22	150522_1	5e93bd34-	4	3	592789	435522	5	Strikeout	S	35	3.56	1.7	-7.08	9.07	FF	18	2	L
2015/5/22	150522_1	0f862642-	4	4	592789	435522	5	Strikeout	S	36	3.66	1.7	6.98	-0.8	CU	22	2	L
2015/5/22	150522_1	76aa0225-	5	1	592789	457705	6	Strikeout	S	40	3.47	1.6	-7.41	7.73	FF	13	2	R
2015/5/22	150522_1	1dd643d3-	5	2	592789	457705	6	Strikeout	B	41	3.47	1.6	6.13	1.58	CU	24	2	R
2015/5/22	150522_1	f9a699b0-	5	3	592789	457705	6	Strikeout	B	42	3.47	1.6	-4.64	10.13	FF	21	2	R
2015/5/22	150522_1	9dc4c0b4-	5	4	592789	457705	6	Strikeout	S	43	3.49	1.6	-8.14	7.1	FF	17	2	R
2015/5/22	150522_1	5f14dd45-	5	5	592789	457705	6	Strikeout	S	44	3.47	1.6	-5.91	9.52	FF	9	2	R
2015/5/22	150522_1	8fb06f54-	3	1	592789	516782	11	Strikeout	S	73	3.46	1.52	7.69	2.45	CU	18	2	R
2015/5/22	150522_1	b08374ce-	3	2	592789	516782	11	Strikeout	S	74	3.46	1.52	-5.92	8.57	FF	13	2	R
2015/5/22	150522_1	8ab797e8-	3	3	592789	516782	11	Strikeout	S	75	3.46	1.52	-5.45	8.72	FF	3	2	R

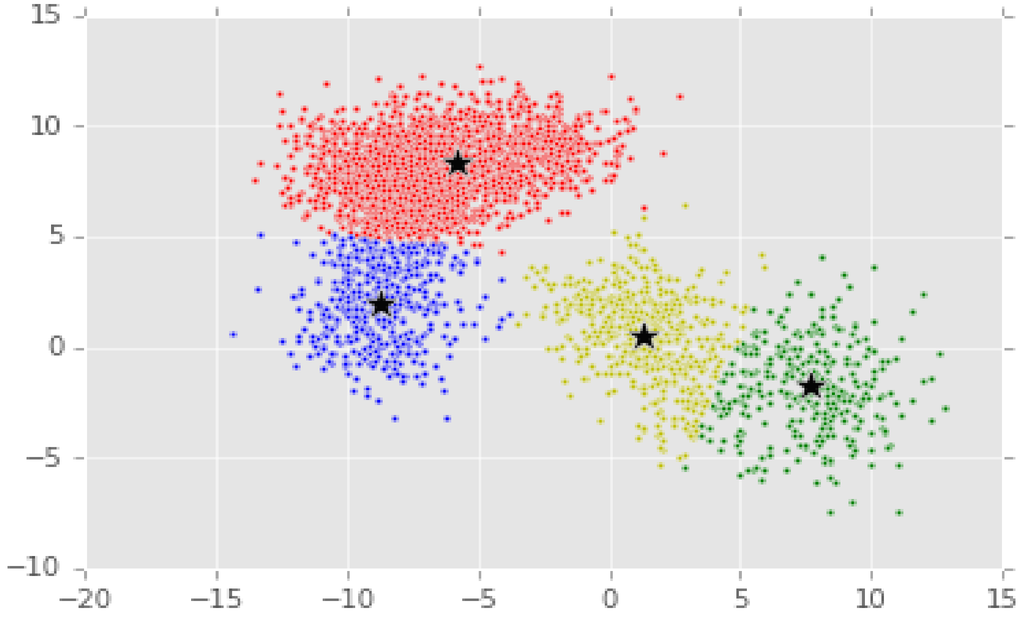
K-means

- Use **Attribute x** (horizontal movement) and **y** (vertical movement) to partition these pitches into 3 clusters.
- FF (four-seam fastball), CH (changeup) and CU (curveball)
- ***Don't*** **use the library related to K-means.**

(i.e. Construct a K-means function by yourself).

K-means

- **Construct a cost function to check the accuracy of pitch types.**
- **Generate a figure** to show the result of K-Means clustering.



K-means

- Try to use another two or more attributes (like speed) to partition.

  Don’t worry whether the accuracy is high or not!

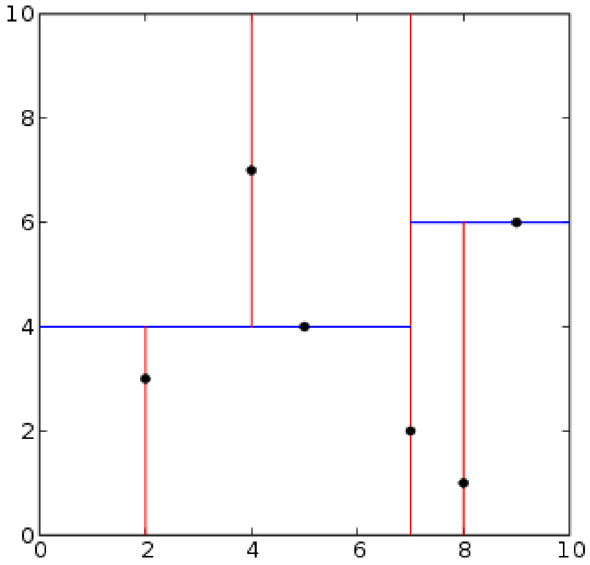
- Show your **code**, **accuracy** and the result of K-Means clustering **(figure) in your report.**

Kd-tree

- You will get a set of points (**points.txt**) in the unit square (all points have x-coordinates and y-coordinates).
- **You *can* use the library related to Kd-tree.**

Kd-tree

- **Draw a 2d-tree divides the unit square (Use two colors).**
- Show your **code** and the result of 2d-tree **(figure) in your report**.



Report & Scoring

- This is a team-based program assignment, so **one team should only submit one report and one source code to E3**.
- The report should contain the following:
  - What environments the members are using (5%)
  - K-means code (30%)
  - Cost function and accuracy (15%)
  - The result of K-Means clustering (15%)
  - Use another two or more attributes to partition (5%)
  - Kd-tree code (15%)
  - The result of Kd-tree (15%)

Some rules

- C / C++ / Java / Python / Matlab are allowed to use.
  For visualization, Excel or other programs are allowed.
- Report format should be **PDF**.
- **Attach your code when you are submitting.**
- **Delay： Your score *= 0.8**
- No cheating and plagiarizing.

I have uploaded assignment #2 description to E3