Clustering

Kenneth (Kenny) Joseph

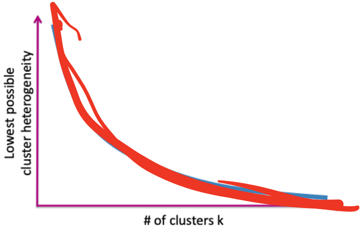
University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

3/23/22

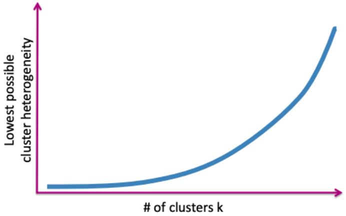
Check your understanding

Consider trying k-means with different values of k. Which of the following graphs shows how the globally optimal heterogeneity changes for each value of k?

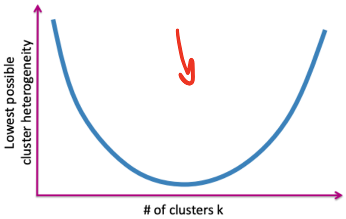
J( )



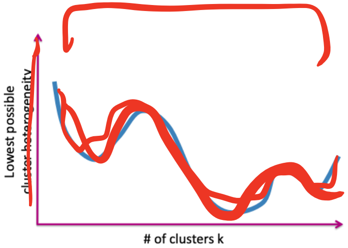
J( )



J( )

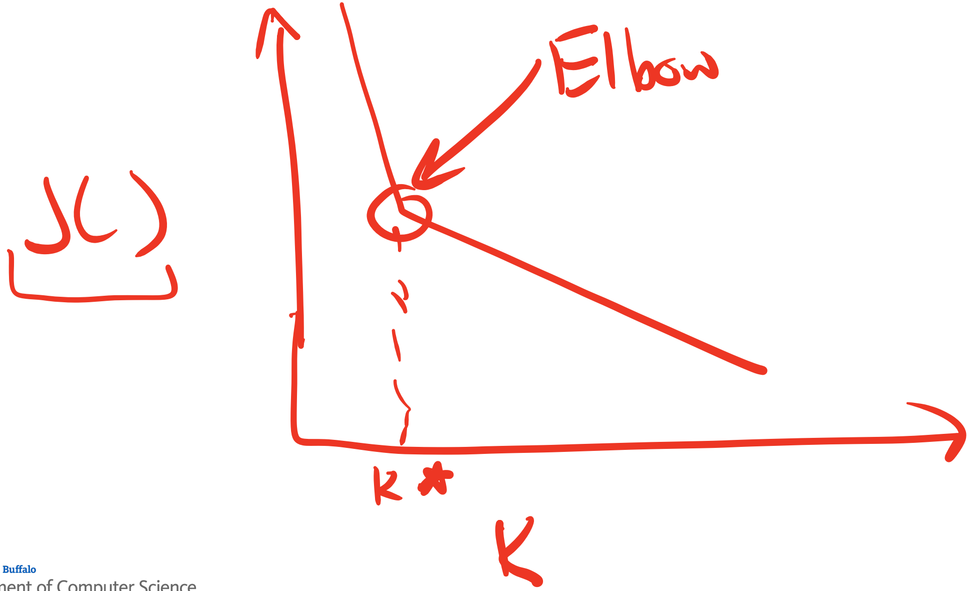


J( )

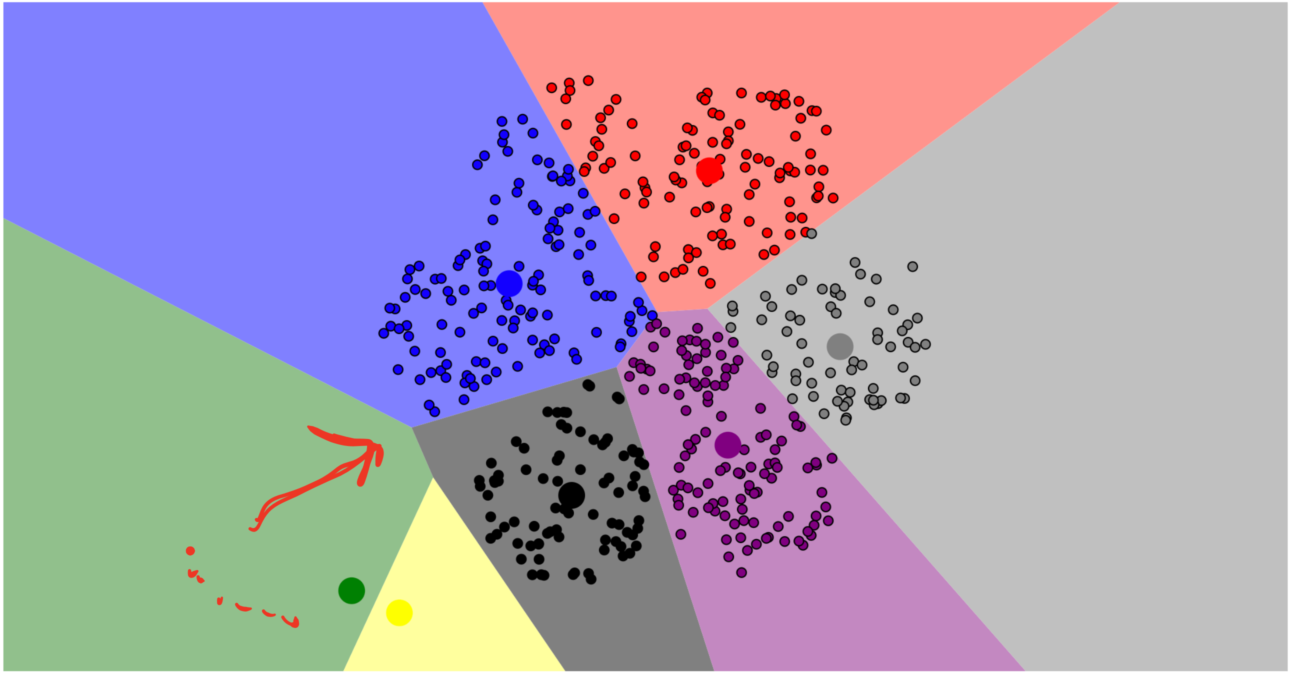


How should we pick K then?

- The "Elbow rule"

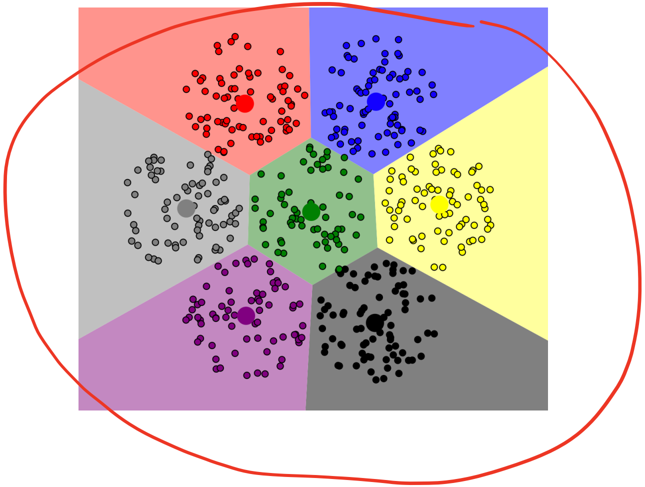


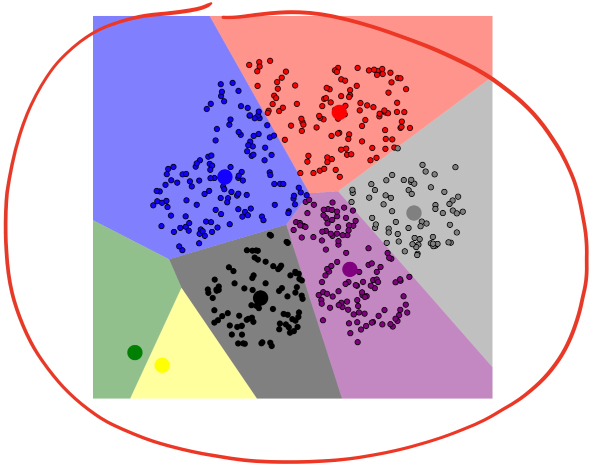
Intuition: Local Minima, simple example



How do we evaluate?

What makes one of these better than the other?



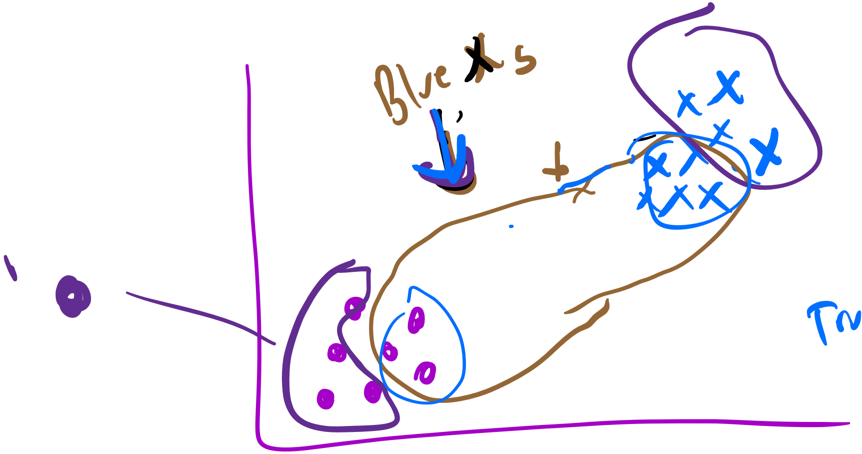


Two sets of evaluation metrics

- When the clusters are **known**
  - Can use the standard approaches, e.g. precision/recall (how?)
    - **PA4!**
  - Can use a variety of metrics
    - Mutual information-based scores
    - Entropy-based scores
    - …
- But we usually cluster when we *don't know* the labels!!

Two sets of evaluation metrics

- When the clusters are **known**
  - Can use the standard approaches, e.g. precision/recall (how?)
    - **PA4!**



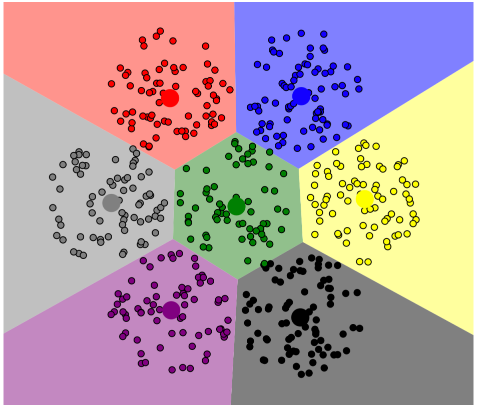
x

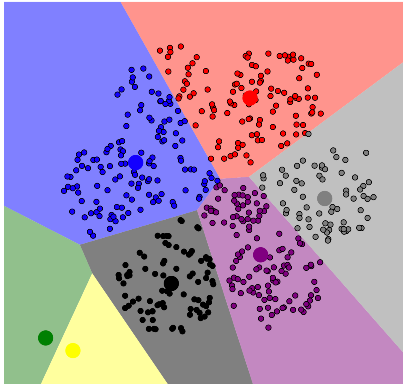
●

	Blue x	Not Blue x
Truth ●	3	4
x	5	4

How do we evaluate?

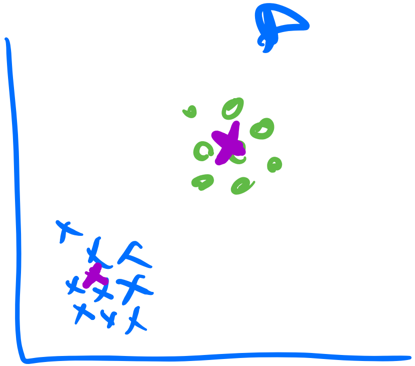
What makes one of these better than the other?

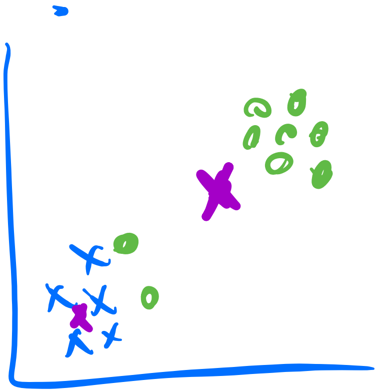




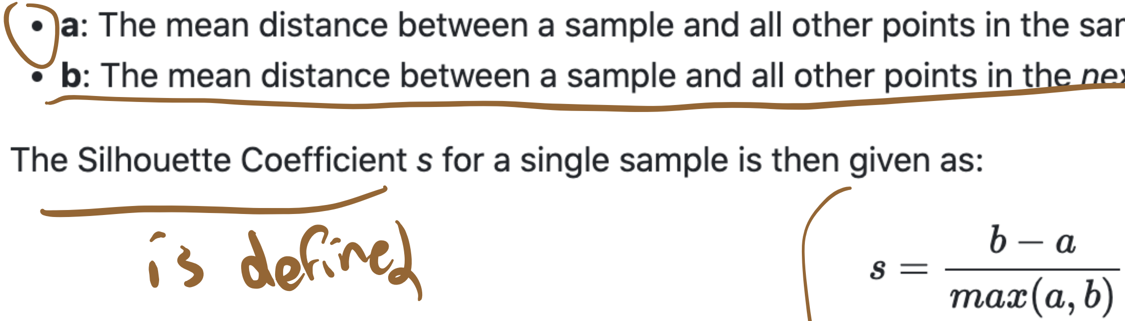
Evaluation (con't)

5 minutes: Come up with an evaluation *metric* that you could use to *quantify* your intuition. Give me a number, and how you computed it!





One Evaluation Metric – Silhouette Score



- **a**: The mean distance between a sample and all other points in the same class.
- **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

is defined for a single point!

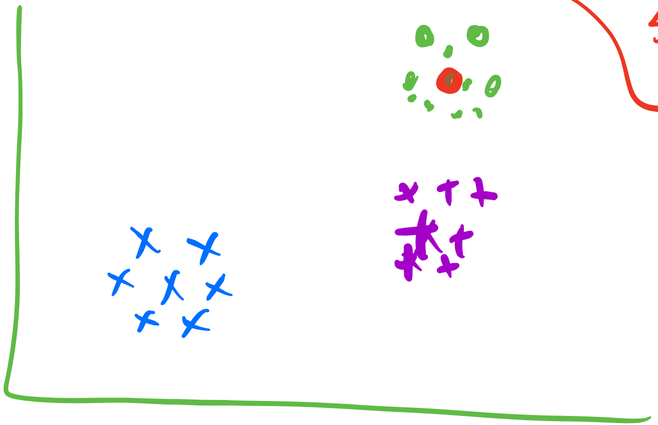
→ for dataset, just take average.

One Evaluation Metric – Silhouette Score

- **a**: The mean distance between a sample and all other points in the same class.
- **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

$$s = \frac{b-a}{max(a,b)}$$



sil score for ● ?

a: ● 1. compute distance from ● to all the ● in it's cluster

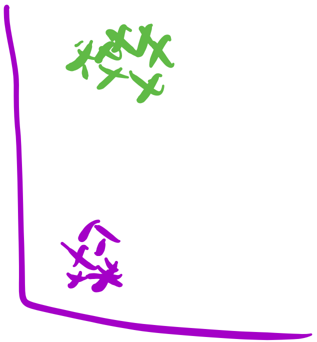
2. take mean.

b 1. Find nearest cluster: +

2. Distance from ● to all +

3. mean

Code Demo



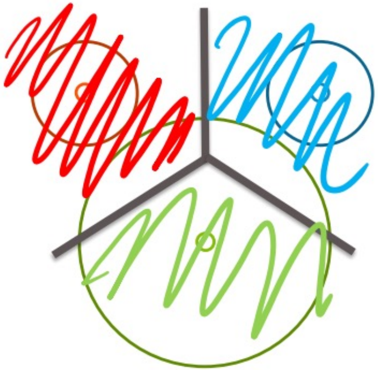
(Intracluster distance should be small

(Intercluster distance should be large

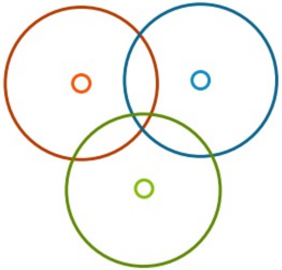
Kmeans Drawbacks: Difficulties w/ high dimensional data

- Full details: Section 3.5, CIML
- Intuition:
  - In high dimensions, distances start to become “more equal” (the variance of the distribution of distances across all points converges to a single number)
  - That’s bad, because all kmeans does is work with distances between centers and points!
  - Luckily, it’s not all that bad, because points are not distributed uniformly,

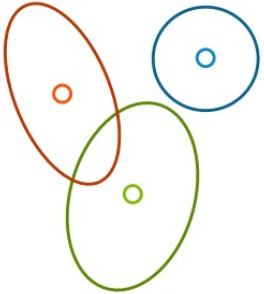
More Drawbacks to K-means



disparate cluster sizes



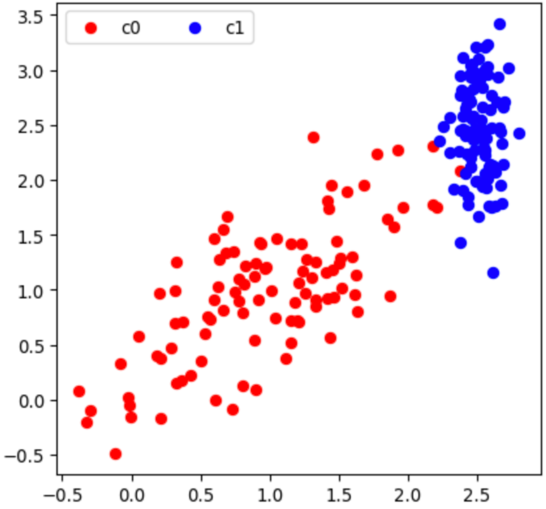
overlapping clusters

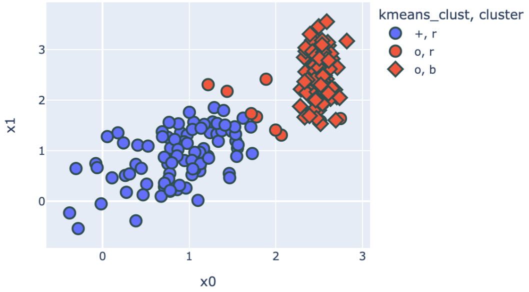


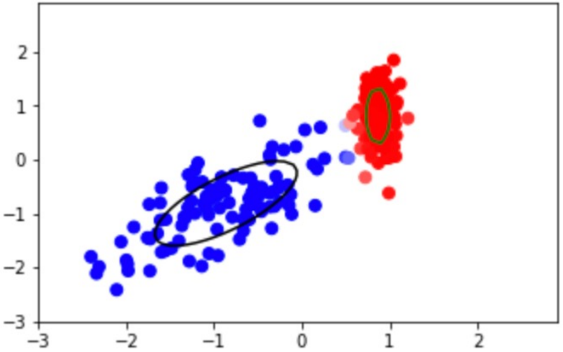
different shaped/oriented clusters

A different approach: (Gaussian) mixture modeling

- Details in notebook…


c0
c1
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


kmeans_clust, cluster
+, r
o, r
o, b
x1
x0


2
1
0
−1
−2
−3
−3
−2
−1
0
1
2