

INTRO to DATA SCIENCE

LECTURE 15: K-MEANS

DATA EXPLORATION

SUPERVISED LEARNING: REGRESSION

SUPERVISED LEARNING: CLASSIFICATION

UNSUPERVISED LEARNING

VARIOUS TOPICS

LOGISTIC REGRESSION

NAIVE BAYES

RANDOM FORESTS

SUPPORT VECTOR MACHINES

COMPETITION (LAST CLASS)

Questions?

DATA EXPLORATION

SUPERVISED LEARNING: REGRESSION

SUPERVISED LEARNING: CLASSIFICATION

UNSUPERVISED LEARNING

VARIOUS TOPICS

CLUSTERING (TODAY)
DIMENSION REDUCTION

DATA EXPLORATION

SUPERVISED LEARNING: REGRESSION

SUPERVISED LEARNING: CLASSIFICATION

UNSUPERVISED LEARNING

VARIOUS TOPICS

**Data exploration presentations
are held next lesson!**

CLUSTERING (TODAY)
DIMENSION REDUCTION

I. ...

II. ...

III. ...

- **DESCRIBE WHAT UNSUPERVISED LEARNING IS**
- **DESCRIBE WHAT CLUSTERING IS**
- **DESCRIBE WHAT K-MEANS DOES**
- **...**

I. CLUSTER ANALYSIS

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

supervised
unsupervised

making predictions
discovering patterns

Q: What is a cluster?

Q: What is a cluster?

*A: A group of **similar** data points.*

Q: What is a cluster?

*A: A group of **similar** data points.*

The concept of similarity is central to the definition of a cluster, and therefore to cluster analysis.

Examples: distance between points, number of common words, etc.

Q: What is the purpose of cluster analysis?

Q: What is the purpose of cluster analysis?

A: To enhance our understanding of a dataset by dividing the data into groups.

People You May Know



Kamal Kumar

1 mutual friend

[Add to My Friends](#)



MrsI F

1 mutual friend

[Add to My Friends](#)



**Imran
Memmedov**

1 mutual friend

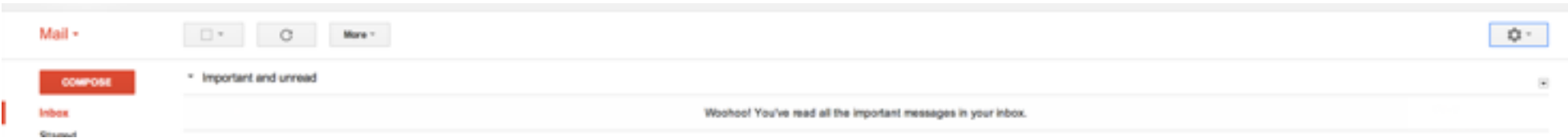
[Add to My Friends](#)



Rick Cruz

1 mutual friend

[Add to My Friends](#)



Priority Inbox: Unsupervised Learning

Group mails into groups and decide which group represents important mails

Q: How do you solve a clustering problem?

Q: How do you solve a clustering problem?

A: Think of a cluster as a “potential class”; then the solution to a clustering problem is to programatically determine these classes.

II. K-MEANS CLUSTERING

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

Q: What is k -means clustering?

Q: What is k -means clustering?

*A: A **greedy** learner that **partitions** a data set into k clusters.*

Q: What is k -means clustering?

*A: A **greedy** learner that **partitions** a data set into k clusters.*

greedy – *captures local structure (depends on initial conditions)*

partition – *each point belongs to exactly one cluster*

Q: What is k -means clustering?

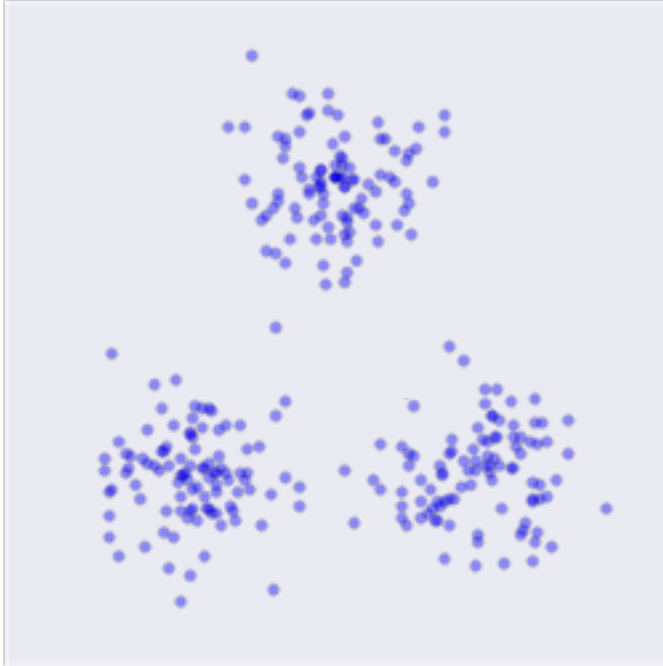
*A: A **greedy** learner that **partitions** a data set into k clusters.*

greedy – *captures local structure (depends on initial conditions)*

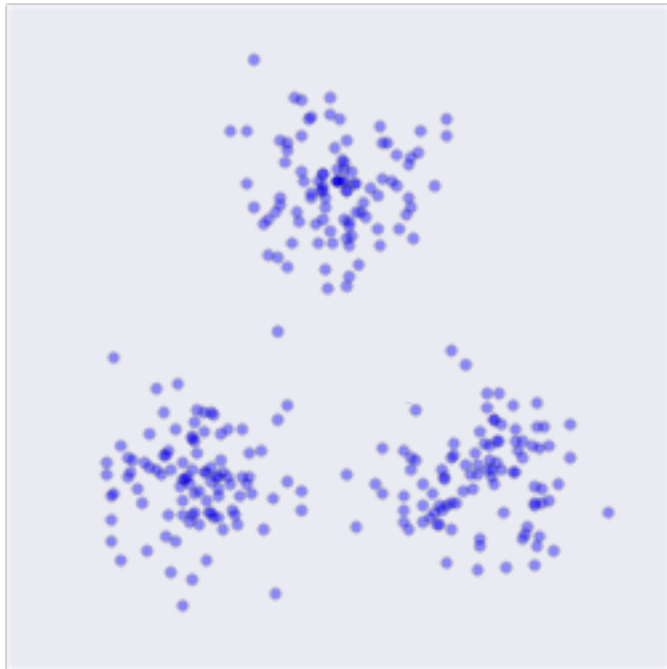
partition – *each point belongs to exactly one cluster*

K-means is algorithmically pretty efficient

(time & space complexity is linear in number of records)



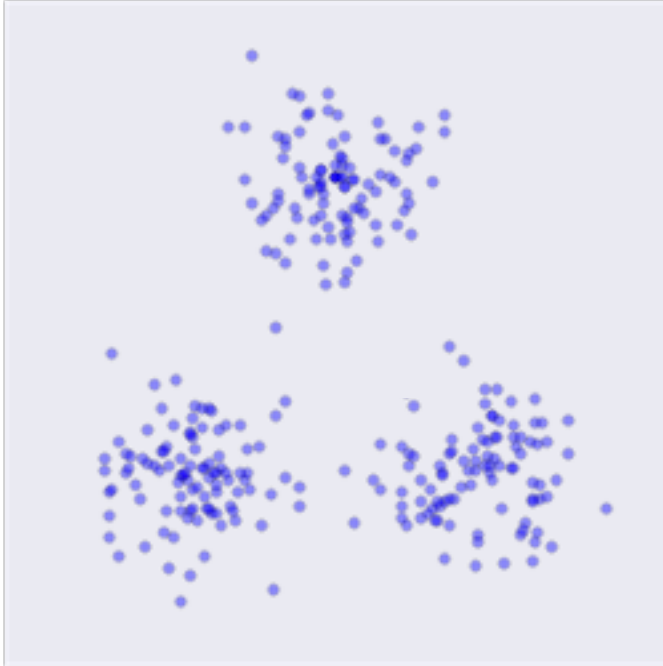
*Suppose we are given some unsupervised data
(i.e., no class labels)*

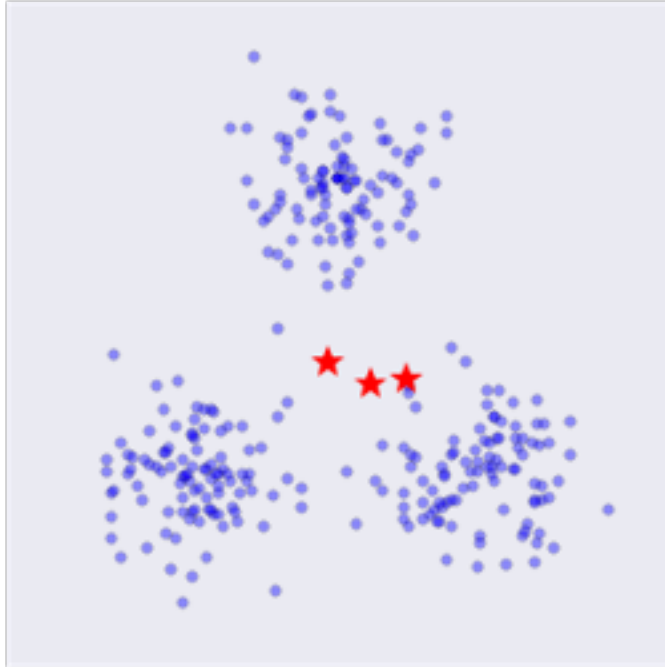


*Suppose we are given some unsupervised data
(i.e., no class labels)*

*We could like to infer class labels from the data,
i.e., cluster the data into similar groups*

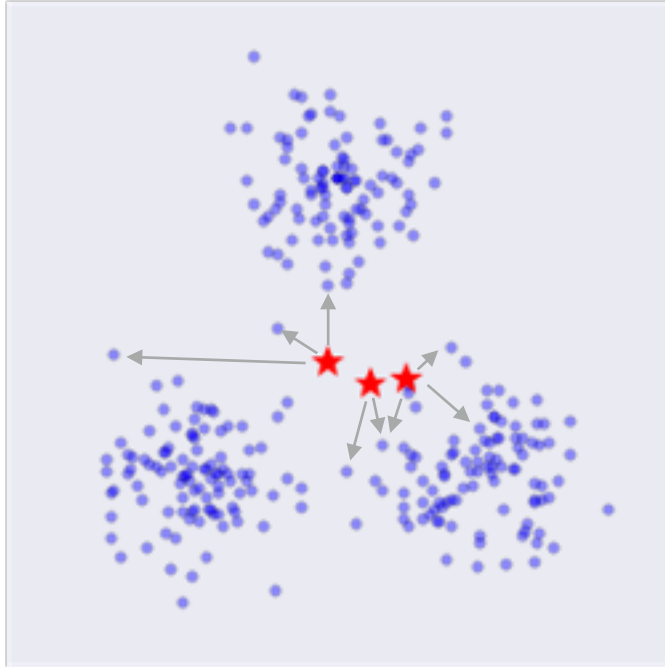
Steps of k-means algorithm





Steps of k -means algorithm

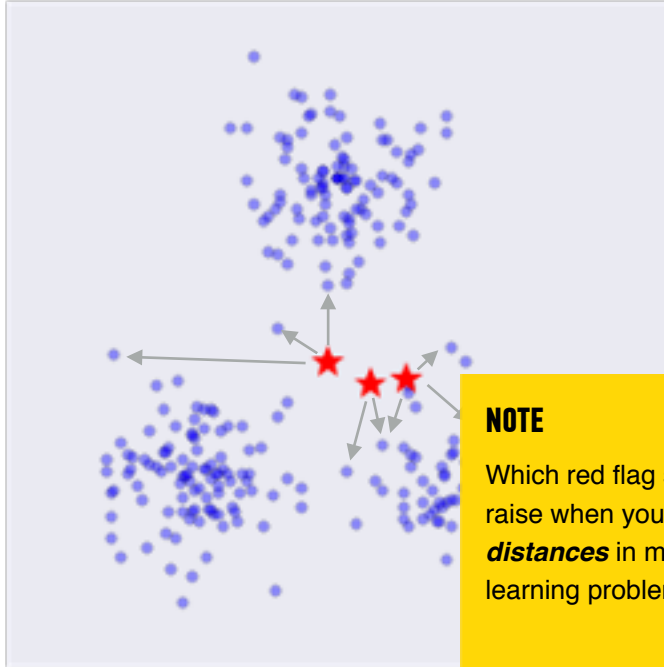
Start with k cluster centers chosen at random



Steps of k-means algorithm

Start with k cluster centers chosen at random

- 1. Compute distances from each point to centers*



Steps of k -means algorithm

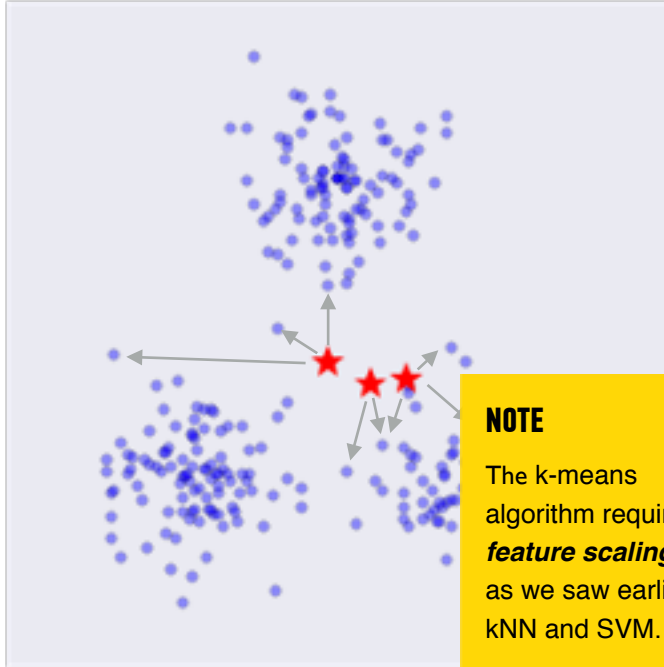
Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*

NOTE



Which red flag should raise when you use ***distances*** in machine learning problems?



NOTE

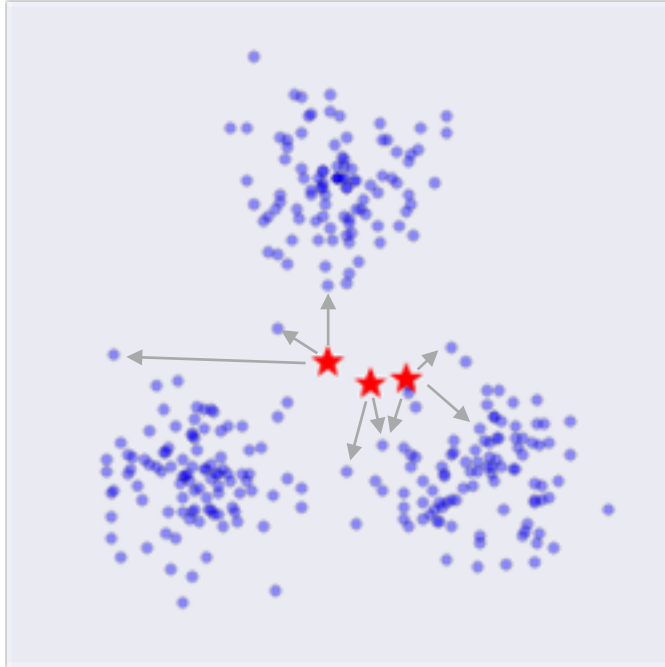


The k-means algorithm requires **feature scaling**, as we saw earlier with kNN and SVM.

Steps of k-means algorithm

Start with k cluster centers chosen at random

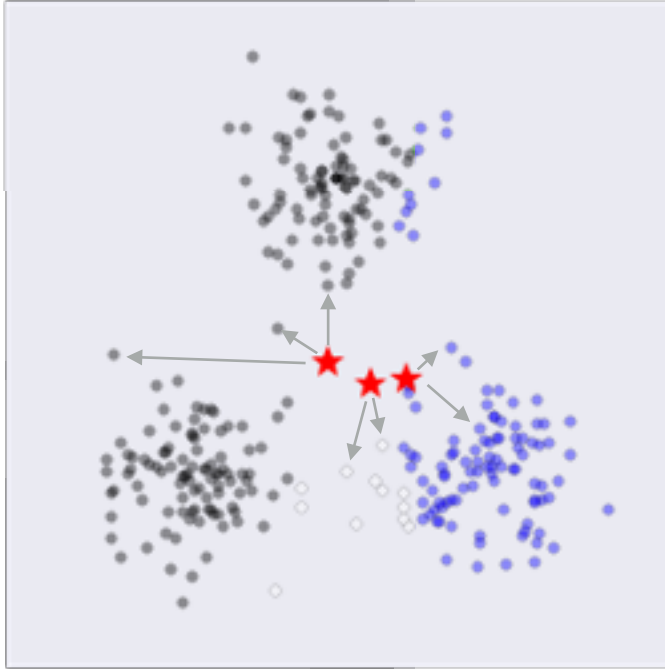
1. *Compute distances from each point to centers*



Steps of k-means algorithm

Start with k cluster centers chosen at random

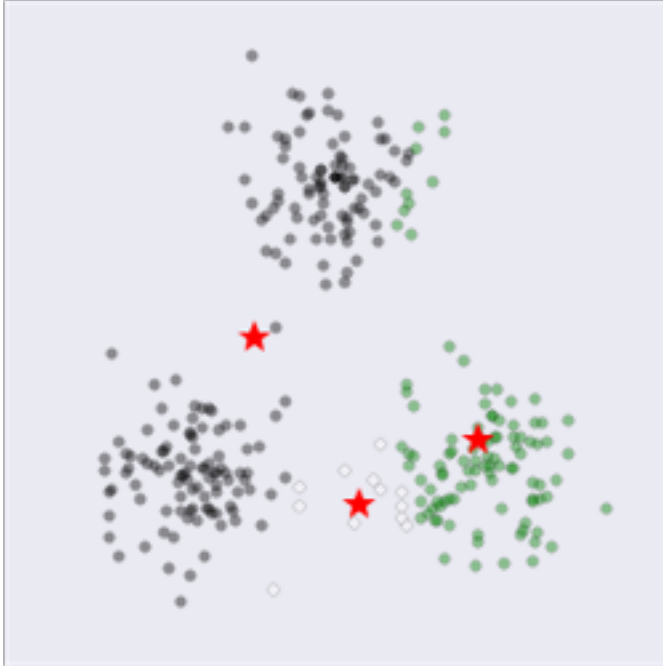
- 1. Compute distances from each point to centers*



Steps of k-means algorithm

Start with k cluster centers chosen at random

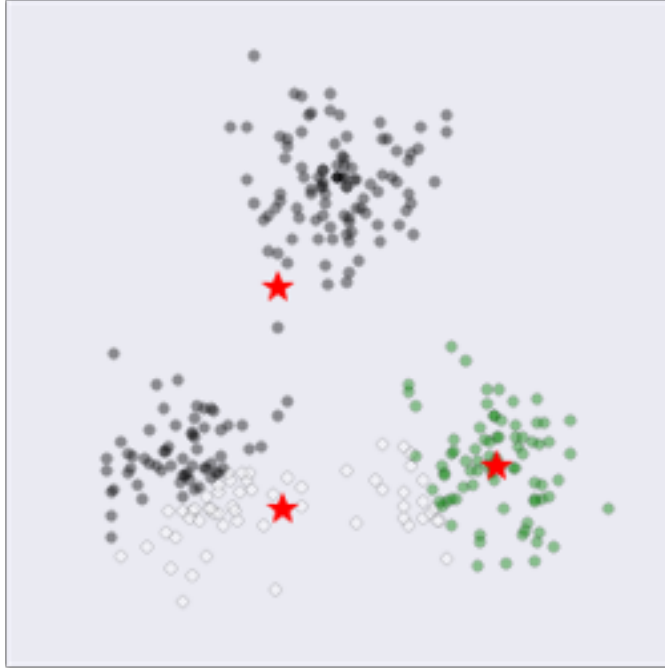
- 1. Compute distances from each point to centers*
- 2. Label data according to their closest cluster*



Steps of k-means algorithm

Start with k cluster centers chosen at random

- 1. Compute distances from each point to centers*
- 2. Label data according to their closest cluster*
- 3. Recompute cluster centers*

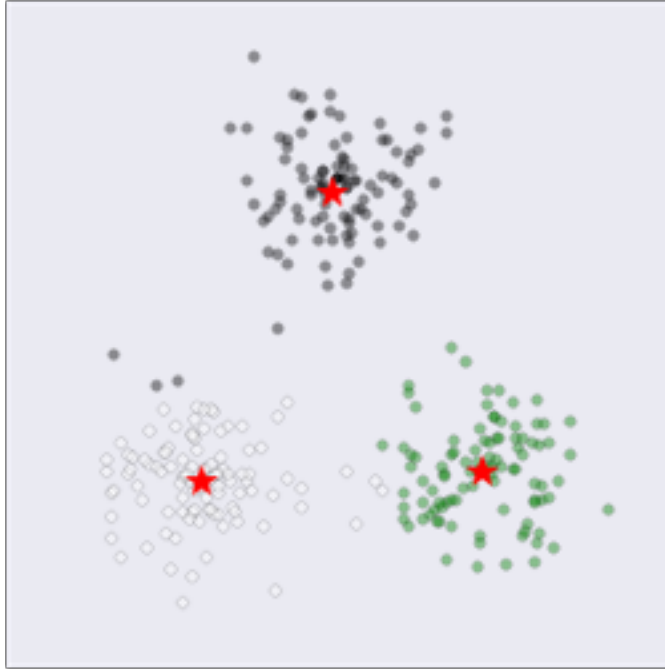


Steps of k-means algorithm

Start with k cluster centers chosen at random

- 1. Compute distances from each point to centers*
- 2. Label data according to their closest cluster*
- 3. Recompute cluster centers*

*Repeat 1-3 until labels don't change
(or some maximum iteration has been reached)*

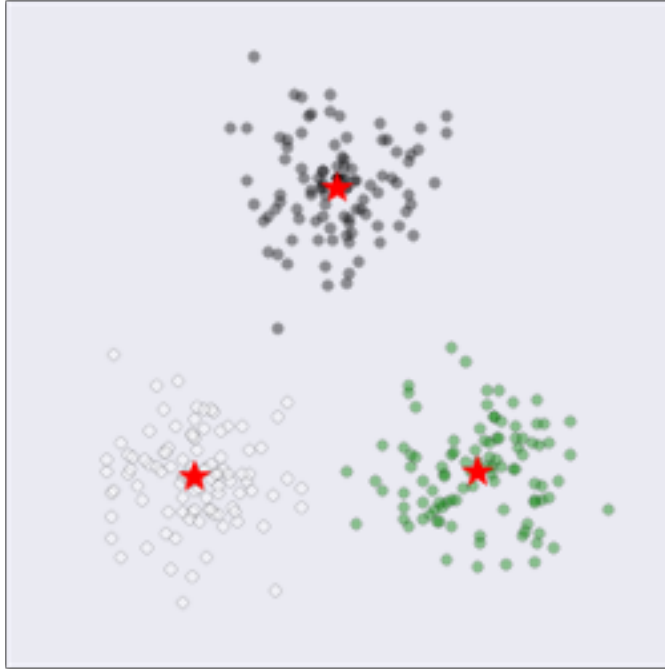


Steps of k-means algorithm

Start with k cluster centers chosen at random

- 1. Compute distances from each point to centers*
- 2. Label data according to their closest cluster*
- 3. Recompute cluster centers*

*Repeat 1-3 until labels don't change
(or some maximum iteration has been reached)*



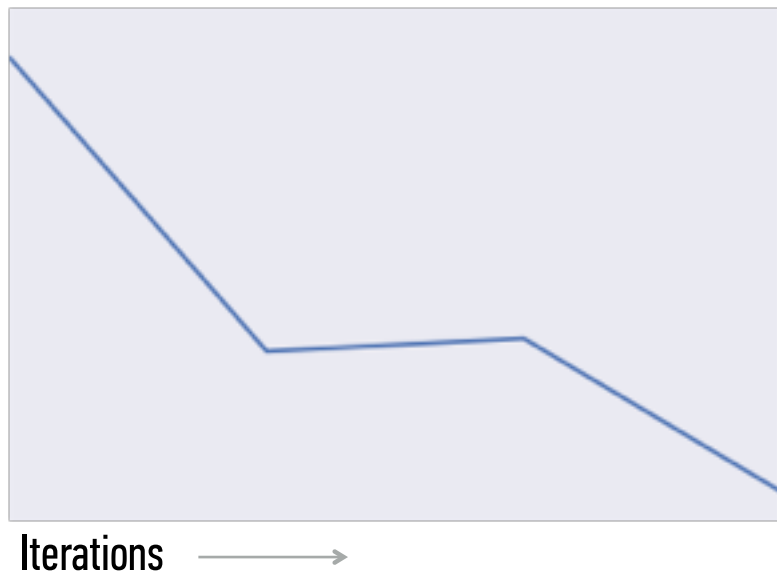
Steps of k-means algorithm

Start with k cluster centers chosen at random

- 1. Compute distances from each point to centers*
- 2. Label data according to their closest cluster*
- 3. Recompute cluster centers*

*Repeat 1-3 until labels don't change
(or some maximum iteration has been reached)*

Average distance to closest cluster



*At each step, we compute the **average distance** to the closest cluster center as its 'cost'*

Average distance to closest cluster



Iterations

*At each step, we compute the **average distance** to the closest cluster center as its 'cost'*

*Sometimes you'd see the **sum of squared distances**, which optimizes identically*

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} d(x, c_i)^2$$

Average distance to closest cluster



Iterations

*At each step, we compute the **average distance** to the closest cluster center as its 'cost'*

*Sometimes you'd see the **sum of squared distances**, which optimizes identically*

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} d(x, c_i)^2$$

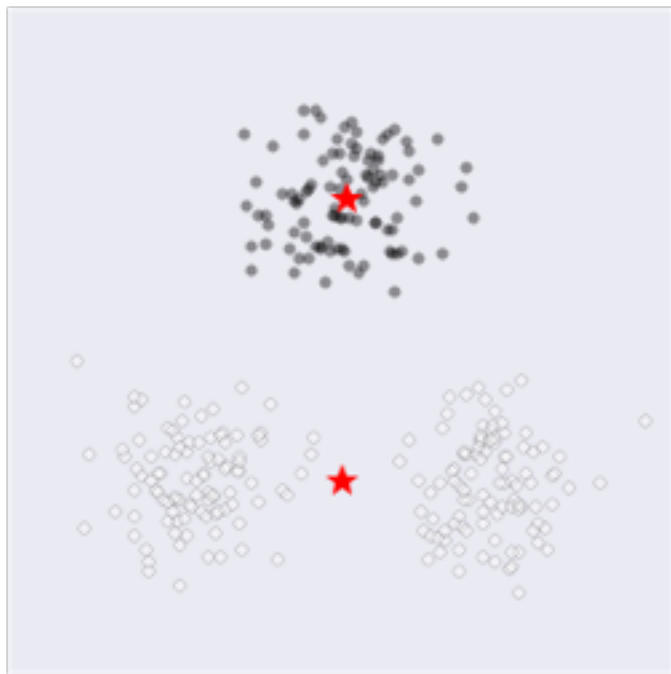
*As you see already, the cost function does **not** necessarily always decrease*

III. CLUSTER VALIDATION

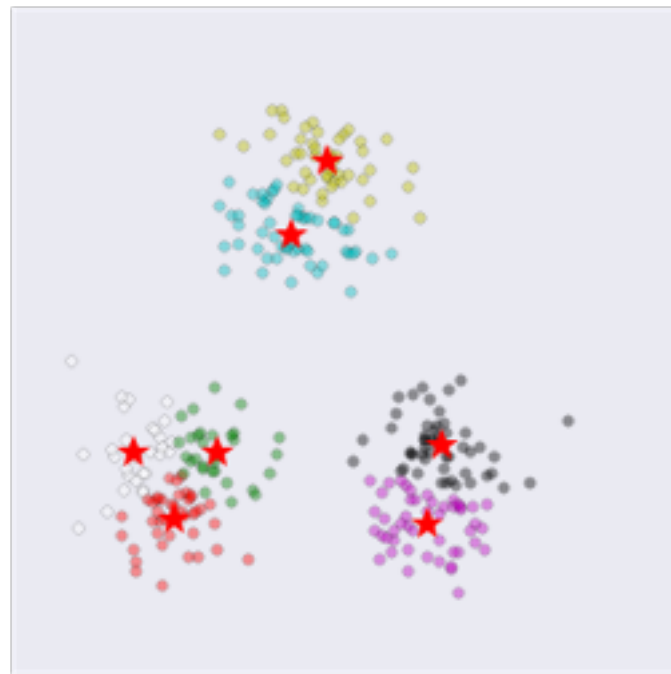
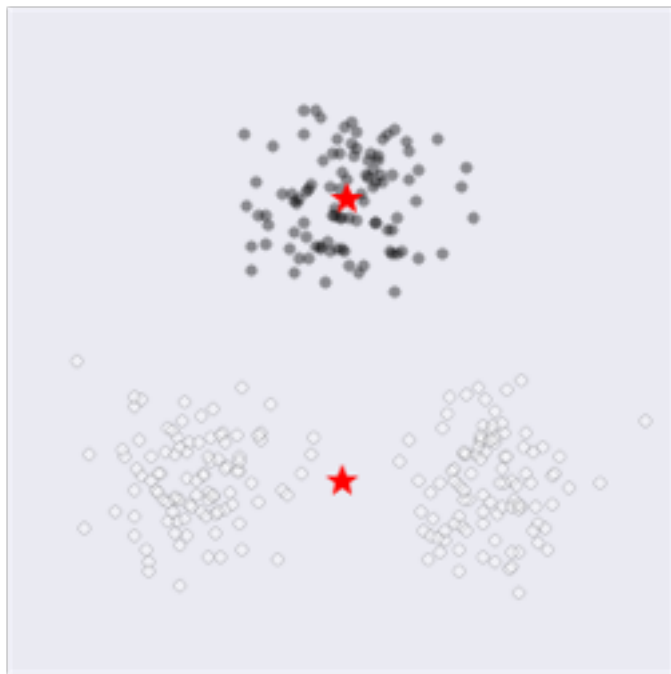
<i>supervised</i> <i>unsupervised</i>	<i>test out your predictions</i> <i>can't really</i>
--	---

How do we choose k ?

How do we choose k ?



How do we choose k ?



In general, k -means will converge to a solution and return a partition of k clusters, even if no natural clusters exist in the data.

In general, k -means will converge to a solution and return a partition of k clusters, even if no natural clusters exist in the data.

*We will look at two validation metrics useful for partitional clustering, **cohesion and separation**.*

Cohesion *measures clustering effectiveness within a cluster.*

Cohesion *measures clustering effectiveness within a cluster.*



Cohesion *measures clustering effectiveness within a cluster.*



$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

Cohesion *measures clustering effectiveness within a cluster.*



$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

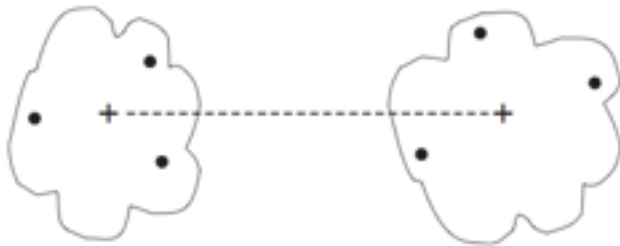
Separation *measures clustering effectiveness between clusters.*

Cohesion *measures clustering effectiveness within a cluster.*



$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

Separation *measures clustering effectiveness between clusters.*



Cohesion *measures clustering effectiveness within a cluster.*



$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

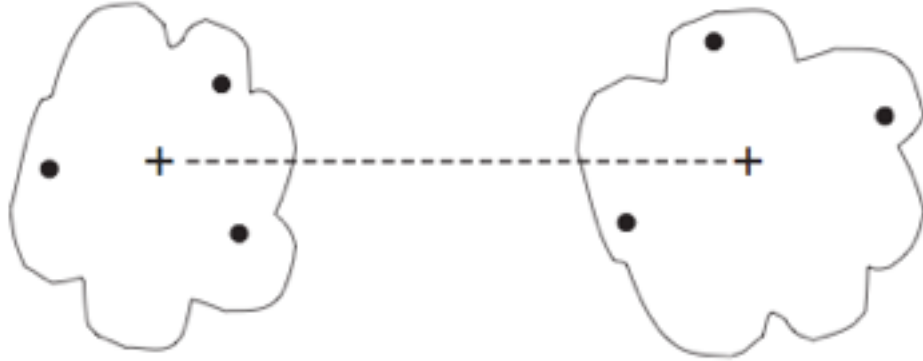
Separation *measures clustering effectiveness between clusters.*



$$\hat{S}(C_i, C_j) = d(c_i, c_j)$$



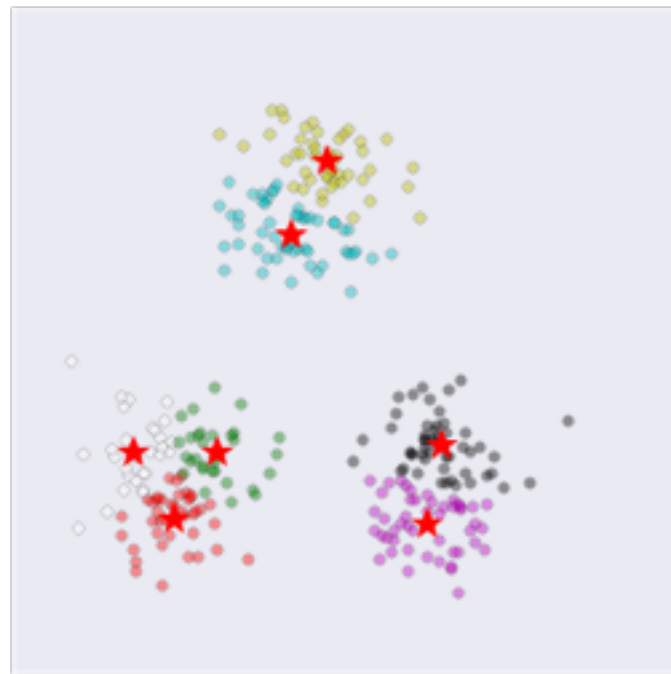
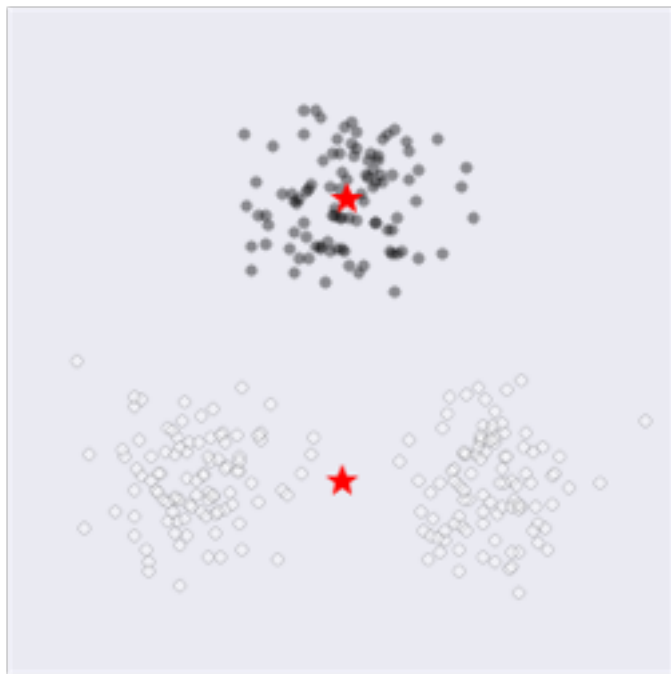
(a) Cohesion.



(b) Separation.

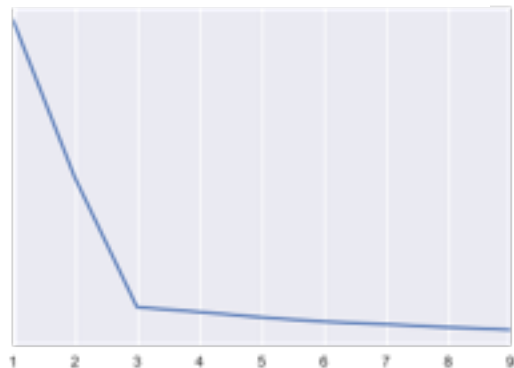
How do we choose k ?

How do we choose k ?

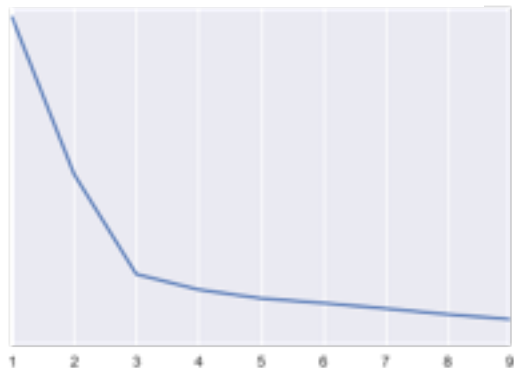


How do we choose k ?

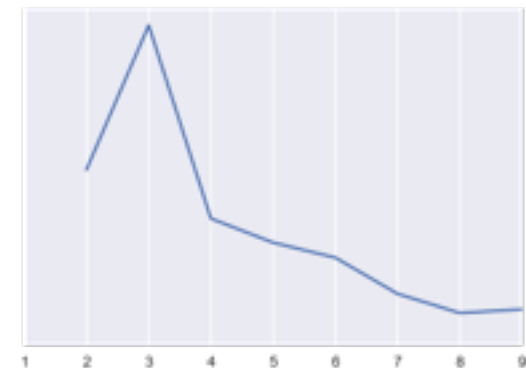
Average distance to closest cluster



Average cohesion within clusters

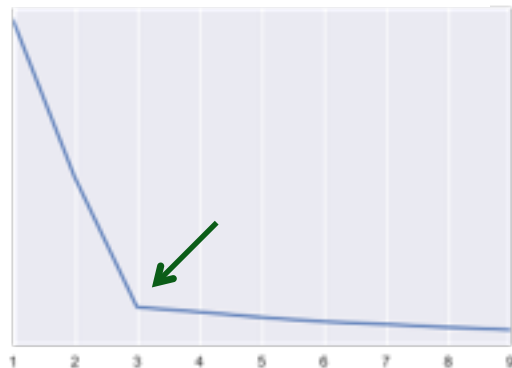


Average separation between clusters

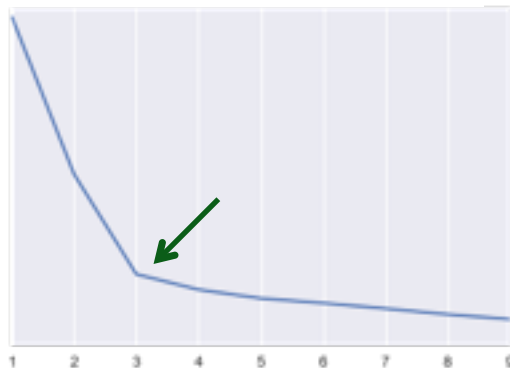


How do we choose k ?

Average distance to closest cluster



Average cohesion within clusters



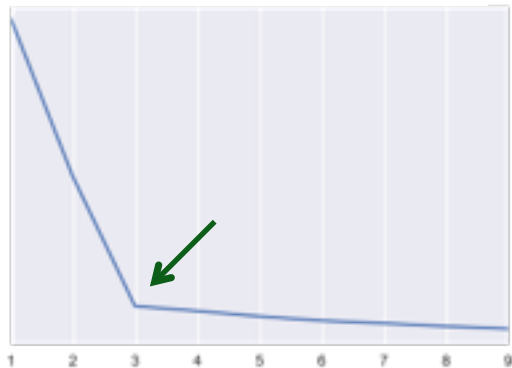
Average separation between clusters



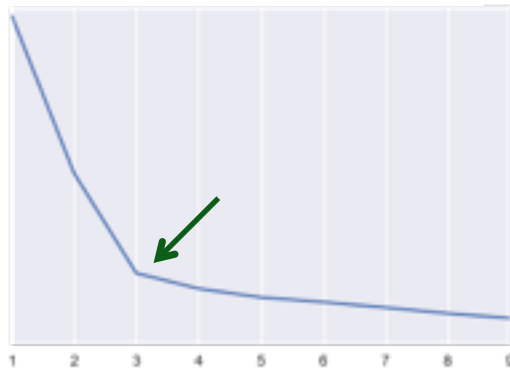
*Look for the **largest kink** in the cost curve (this is called the **elbow method**)*

How do we choose k ?

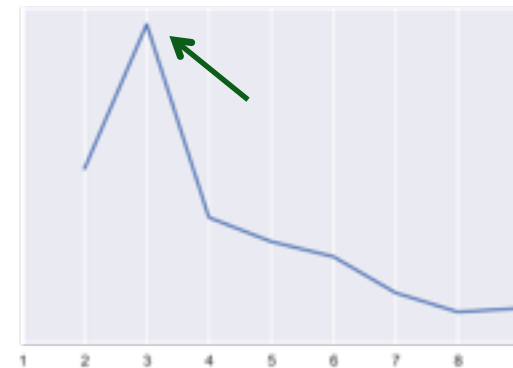
Average distance to closest cluster



Average cohesion within clusters



Average separation between clusters



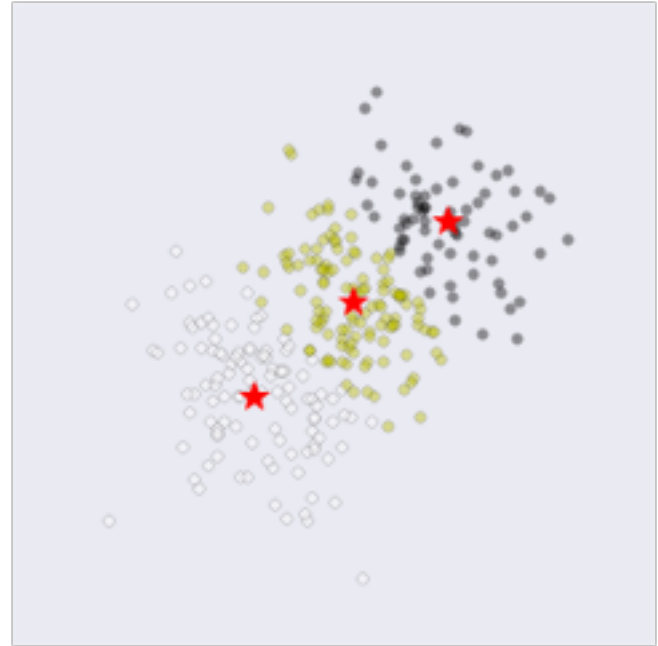
*Look for the **largest kink** in the cost curve (this is called the **elbow method**)*

*Or look for the **largest separation** between clusters*

In practice, you'd choose k with a certain application in mind

In practice, you'd choose k with a certain application in mind

*For example, you'd like to
manufacture three sizes of
clothing: small, medium or large*



Ultimately, cluster validation and clustering in general are suggestive techniques that rely on human interpretation to be meaningful.

INTRO TO DATA SCIENCE

DISCUSSION