GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 4 - Clustering
Wednesday 13th January 2016

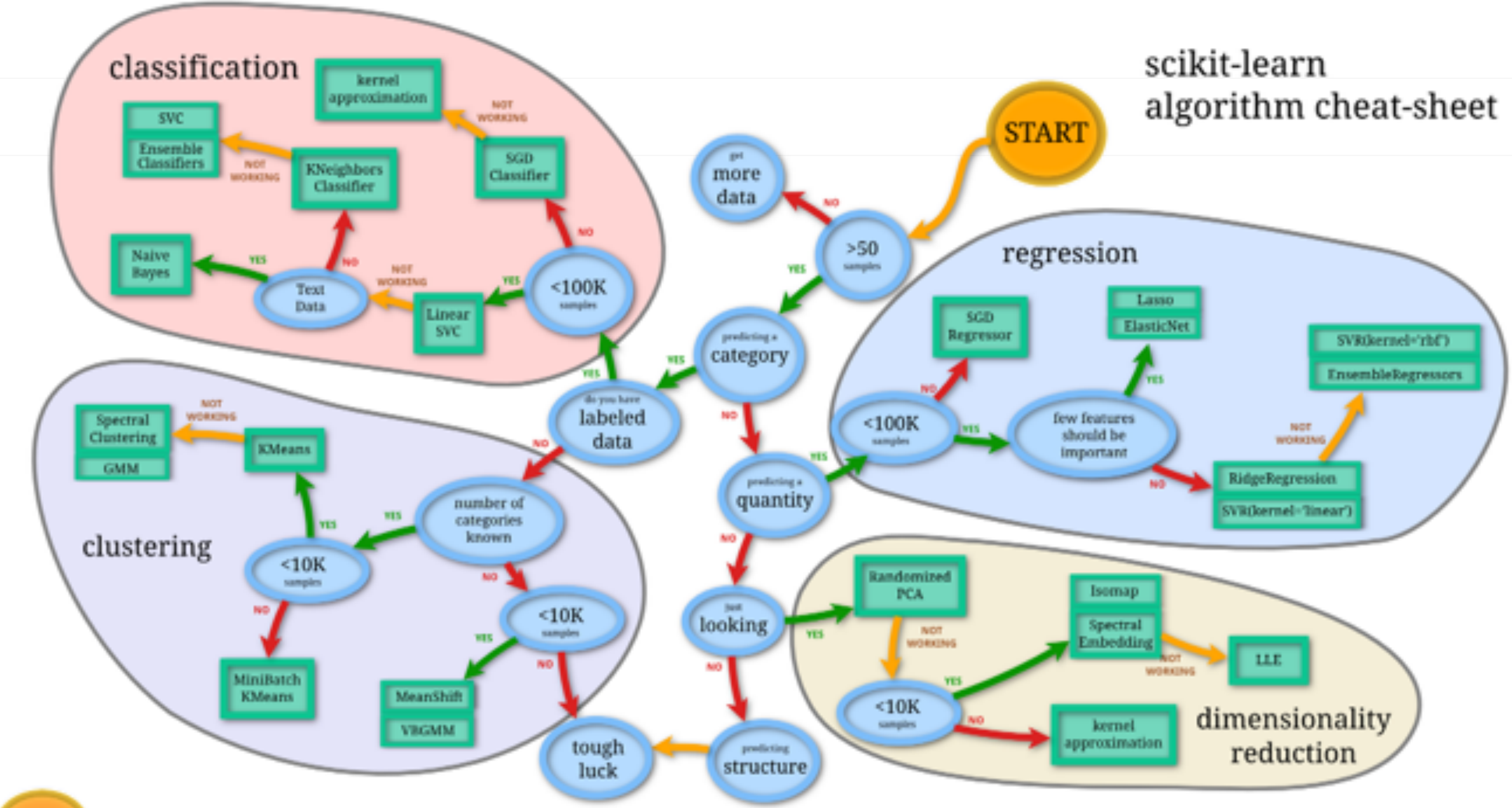
AGENDA

1. Motivation / Review
2. What is Clustering?
3. What is K-Means and how does it work?
4. Lab
5. Discussion - Homework, Project, Kaggle

DATA SCIENCE PART TIME COURSE

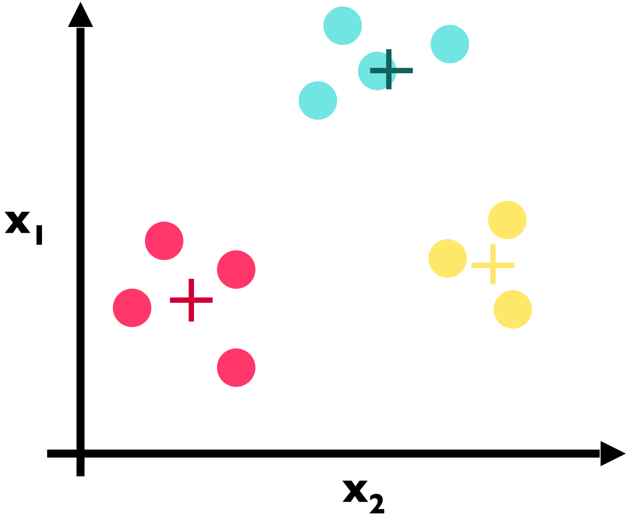
WHAT IS CLUSTERING AND WHY DO IT?

scikit-learn algorithm cheat-sheet



Back

- What is a Cluster?
- Why would we do this?
- What is K-Means?



WHAT IS A CLUSTER?

Recall unsupervised learning is when we are trying to find interesting patterns or groups in our data. We don't have a variable we are trying to predict (a Y value).

Clustering aims to discover subgroups in our data where the points are similar to each other. So we have a collection of groups and all points belonging to the same group are similar. Points in different groups are different to each other.

We have to decide what variables we will construct the groups on. What makes them different (or similar)?

WHY WOULD WE CLUSTER DATA?

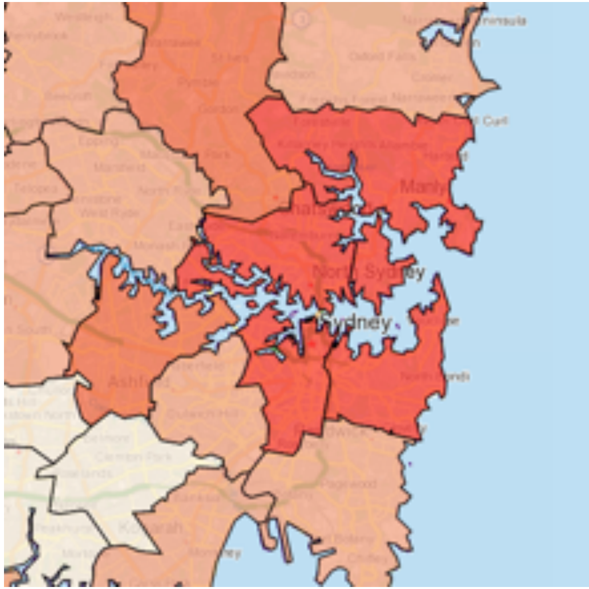
To enhance our understanding of a dataset by dividing the data into groups.

Clustering provides a layer of abstraction from individual data points.

The goal is to extract and enhance the natural structure of the data

WHY WOULD WE CLUSTER DATA?

Marketing teams might want to group customers into like groups as a way of summarising the data

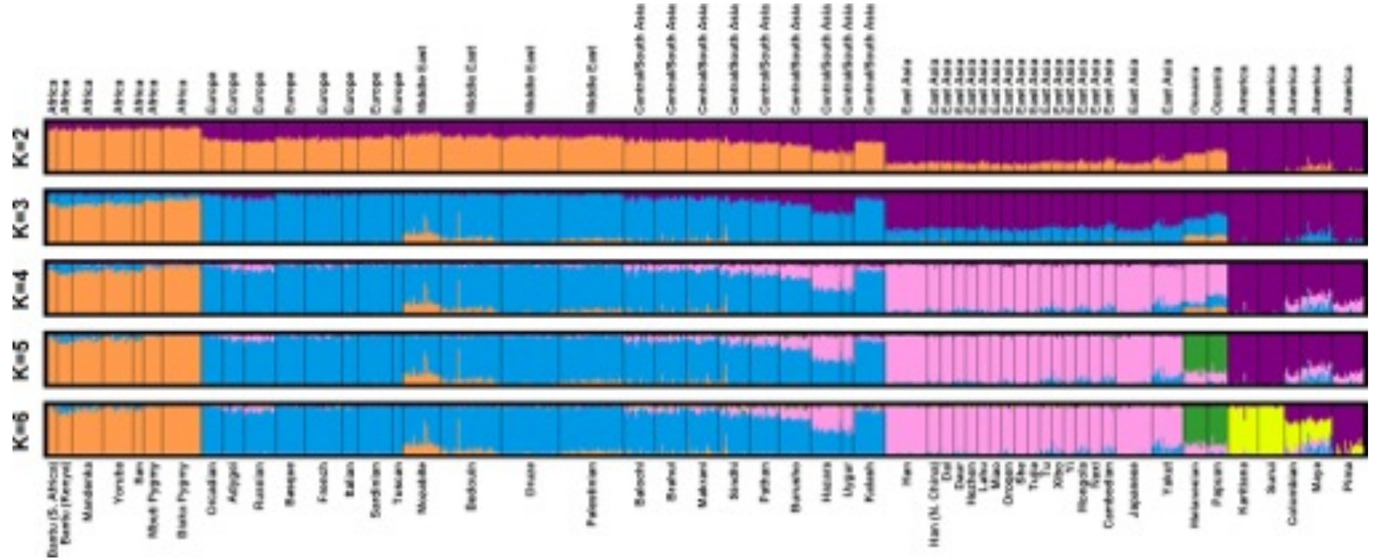


WHY WOULD WE CLUSTER DATA?

Financial groups may want to group transactions into like groups as a way to find unusual payments

WHY WOULD WE CLUSTER DATA?

Genetics data can be clustered to identify ancestry



DATA SCIENCE PART TIME COURSE

HOW DO WE CLUSTER DATA?

KMEANS ALGORITHM

1) Choose k initial centroids (note that k is an input)

2) For each point:
    - find distance to each centroid
    - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

STEP 1 - CHOOSE CENTROIDS

There are several options:

- randomly (but may yield divergent behavior)
- perform alternative clustering task, use resulting centroids as initial k-means centroids
- start with global centroid, choose point at max distance, repeat (but might select outlier)

STEP 2 - ASSESS SIMILARITY

The similarity criterion is determined by the measure we choose.

In the case of k-means clustering, the similarity metric is the **Euclidian distance:**

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{N} (x_{1i} - x_{2i})^2}$$

STEP 3 - RECALCULATE CENTROID POSITIONS

Q: How do we re-compute the positions of the centres at each iteration of the algorithm?

A: By calculating the centroid (i.e., the geometric centre)

STEP 4 - CONVERGENCE

We iterate until some stopping criteria are met; in general, suitable convergence is achieved in a small number of steps.

Stopping criteria can be based on the centroids (eg, if positions change by no more than ε) or on the points (eg, if no more than x% change clusters between iterations).

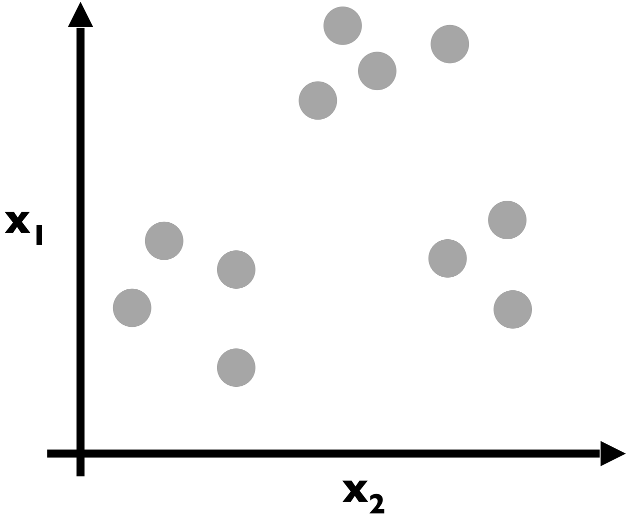
KMEANS ALGORITHM

1) Choose k initial centroids

2) For each point:
    - find distance to each centroid
    - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



1) Choose k initial centroids

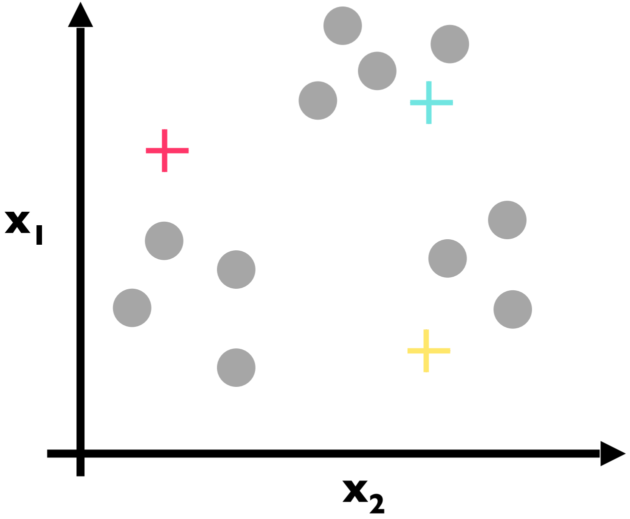
2) For each point:

  - find distance to each centroid

  - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



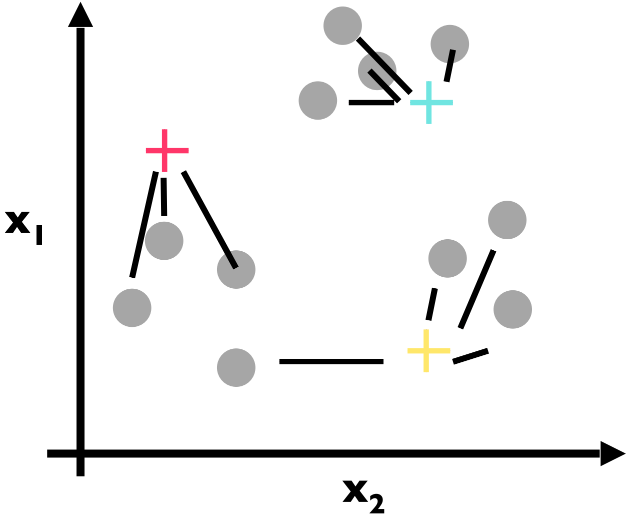
1) Choose k initial centroids

2) For each point:

- **find distance to each centroid**
- assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



KMEANS ALGORITHM

1) Choose k initial centroids

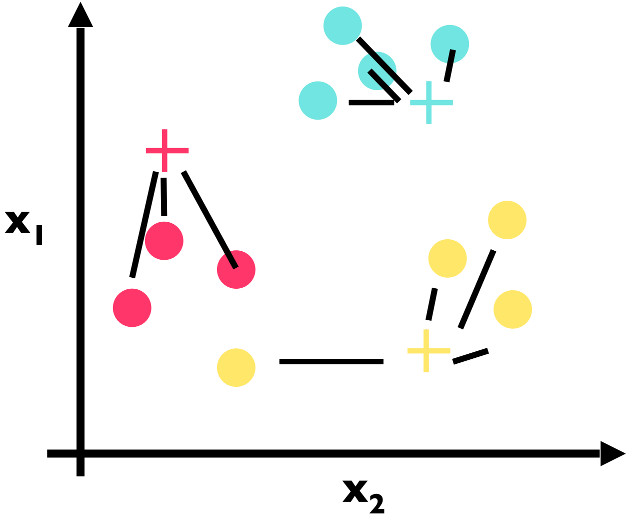
2) For each point:

    - find distance to each centroid

    - **assign point to nearest centroid**

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



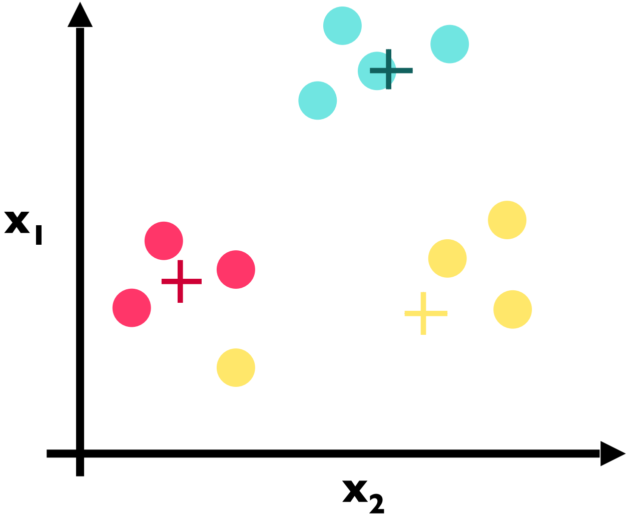
1) Choose k initial centroids

2) For each point:

- find distance to each centroid
- assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



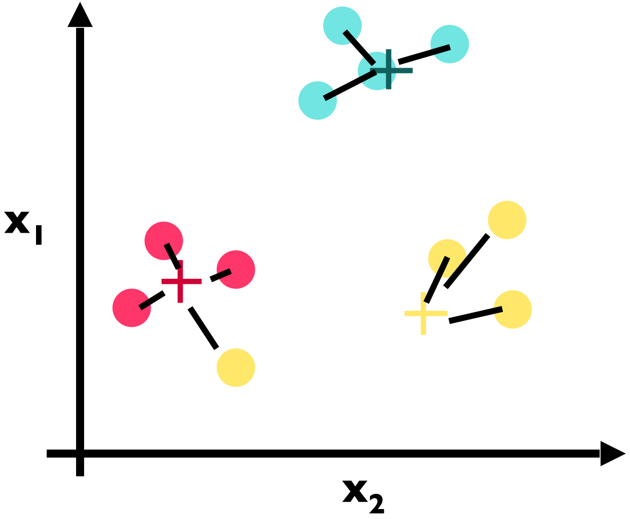
1) Choose k initial centroids

2) For each point:

- find distance to each centroid
- assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



1) Choose k initial centroids

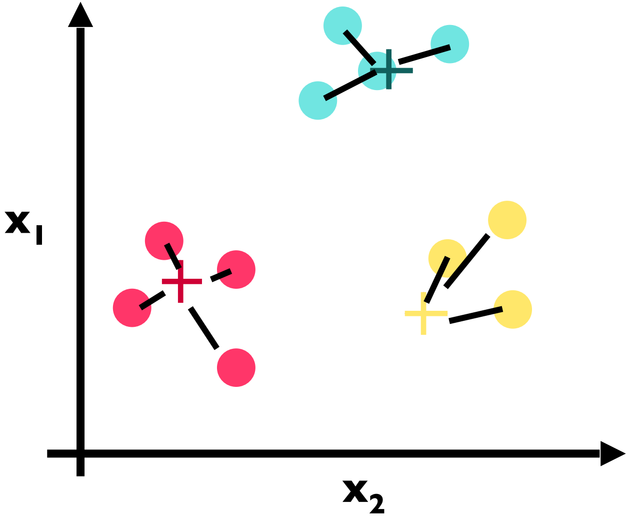
2) For each point:

    - find distance to each centroid

    - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



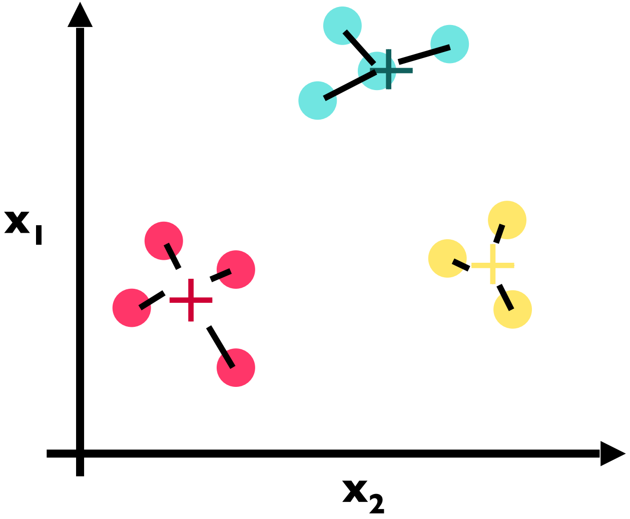
1) Choose k initial centroids

2) For each point:

- find distance to each centroid
- assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



KMEANS ALGORITHM

1) Choose k initial centroids

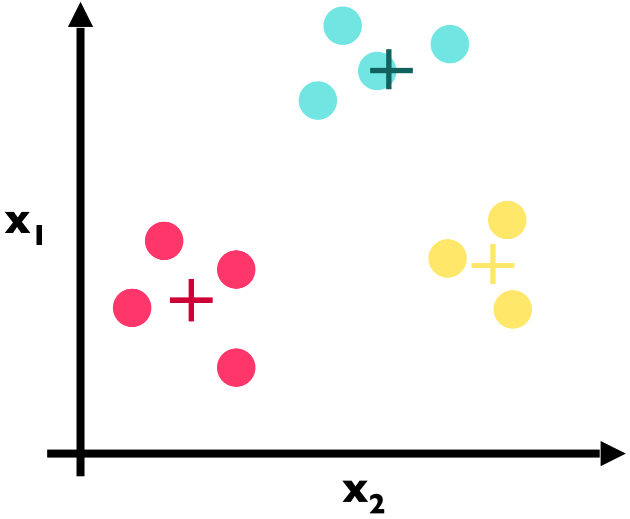
2) For each point:

   - find distance to each centroid

   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



DATA SCIENCE PART TIME COURSE

HOW DO WE KNOW OUR CLUSTERS ARE ANY GOOD?

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**.

CLUSTER VALIDATION

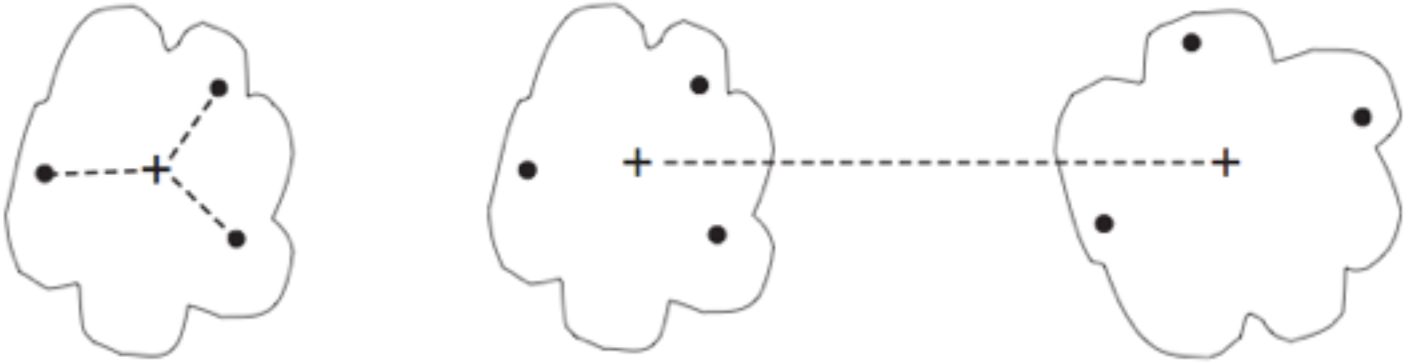
Cohesion measures clustering effectiveness within a cluster.

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

Separation measures clustering effectiveness between clusters.

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

CLUSTER VALIDATION



(a) Cohesion.

(b) Separation.

Figure 8.28. Prototype-based view of cluster cohesion and separation.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

CLUSTER VALIDATION

One useful measure than combines the ideas of cohesion and separation is the silhouette coefficient. For point x_i, this is given by:

$$SC_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

such that:

a_i = average in-cluster distance to x_i

b_{ij} = average between-cluster distance to x_i

$b_i = \min_j(b_{ij})$

The silhouette coefficient can take values between -1 and 1.

In general, we want separation to be high and cohesion to be low. This corresponds to a value of SC close to +1.

A negative silhouette coefficient means the cluster radius is larger than the space between clusters, and thus clusters overlap

The silhouette coefficient for the cluster C_i is given by the average silhouette coefficient across all points in C_i:

$$SC(C_i) = \frac{1}{m_i} \sum_{x \in C_i} SC_i$$

The overall silhouette coefficient is given by the average silhouette coefficient across all clusters:

$$SC_{total} = \frac{1}{k} \sum_{1}^{k} SC(C_i)$$

One useful application of cluster validation is to determine the best number of clusters for your dataset.

Q: How would you do this?

A: By computing the SSE or SC for different values of k.

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

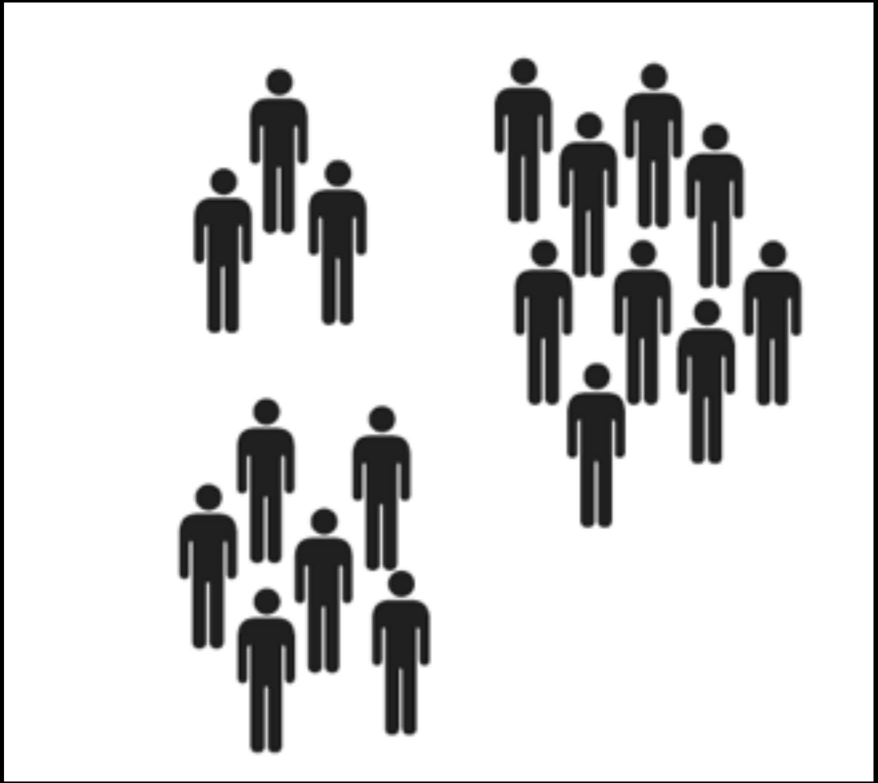
Strengths:

K-means is a popular algorithm because of its computational efficiency and simple and intuitive nature.

Weaknesses:

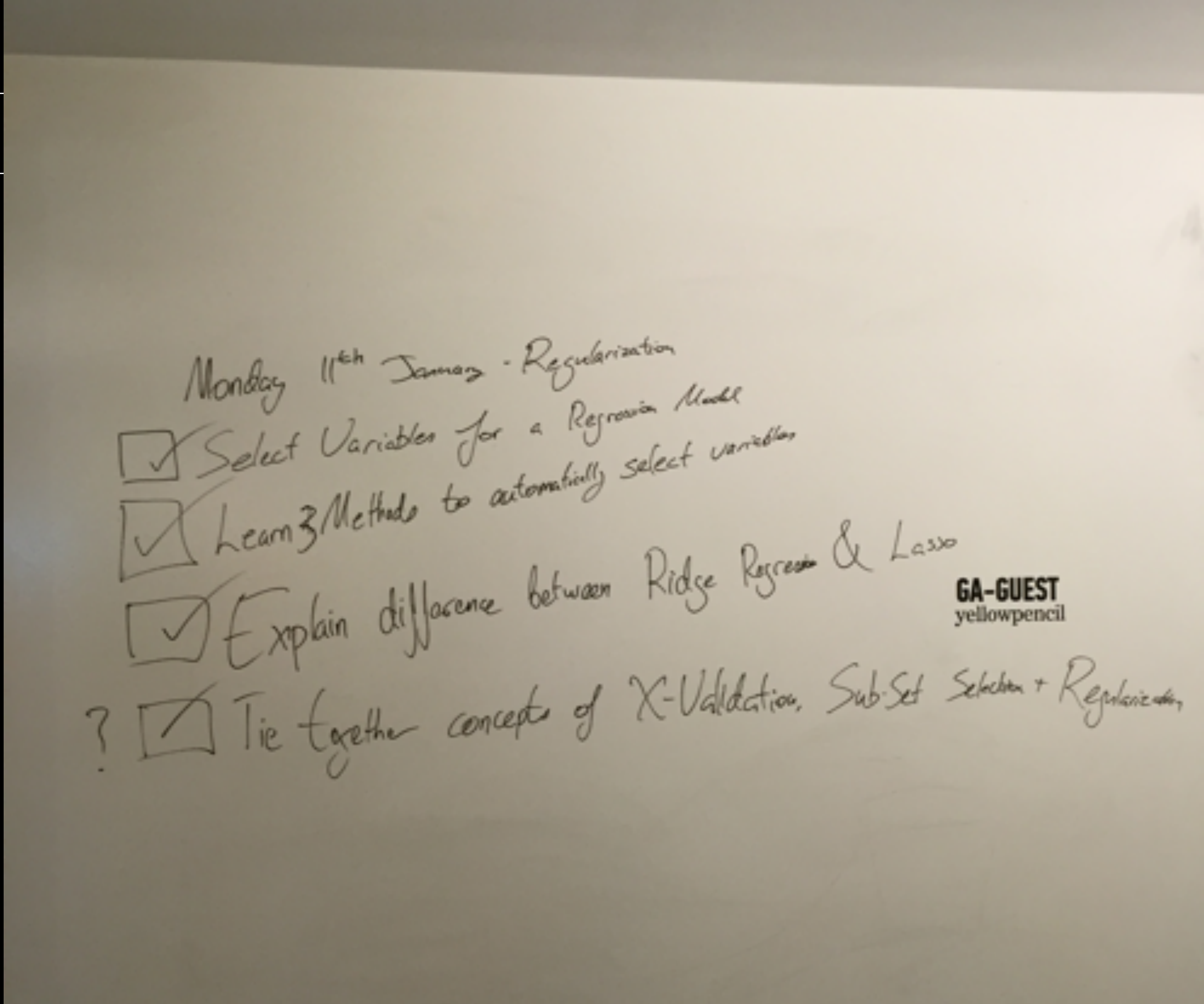
However, K-means is highly scale dependent, and is not suitable for data with widely varying shapes and densities.

LAB



```
git remote -v
git remote add upstream https://github.com/ihansel/SYD_DAT_3.git
git remote -v
git fetch upstream
git checkout master
git merge upstream/master
OR git reset –hard upstream/master
```

WEEK 4 Monday



Monday 11th January - Regularization

☑ Select Variables for a Regression Model

☑ Learn 3 Methods to automatically select variables

☑ Explain difference between Ridge Regression & Lasso

? ☑ Tie together concepts of X-Validation, Sub-Set Selection + Regularization

GA-GUEST
yellowpencil

DISCUSSION TIME

- Homework 2
- Homework 3
- Presentations next Wednesday – 5 slides on a data science topic of your choice
- Project
- Kaggle

DATA SCIENCE - Week 3 Day 1

HOMEWORK 2

Highlights

- Good communication on the basics of what regression is, some might have been a little too technical but that’s ok
- Most were able to interpret the outputs of the linear regression and calculate the Y variables by hand
- Nice reviews of Small Multiple visualisations
- More in depth thinking about the projects

HOMEWORK 3

Tasks

- Prepare a 5 minute presentation for class next Wednesday (details in homework notebook)
- Data in github for your project (or transferred to me in some form)
- Read your data in Python
- Visualise your data in Python
- One of the following with your data; Linear or Logistic Regression (with regularization), or Clustering

READINGS

Read the following before class on Monday

- http://www.slideshare.net/MrChrisJohnson/algorithmic-music-recommendations-at-spotify
- http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html