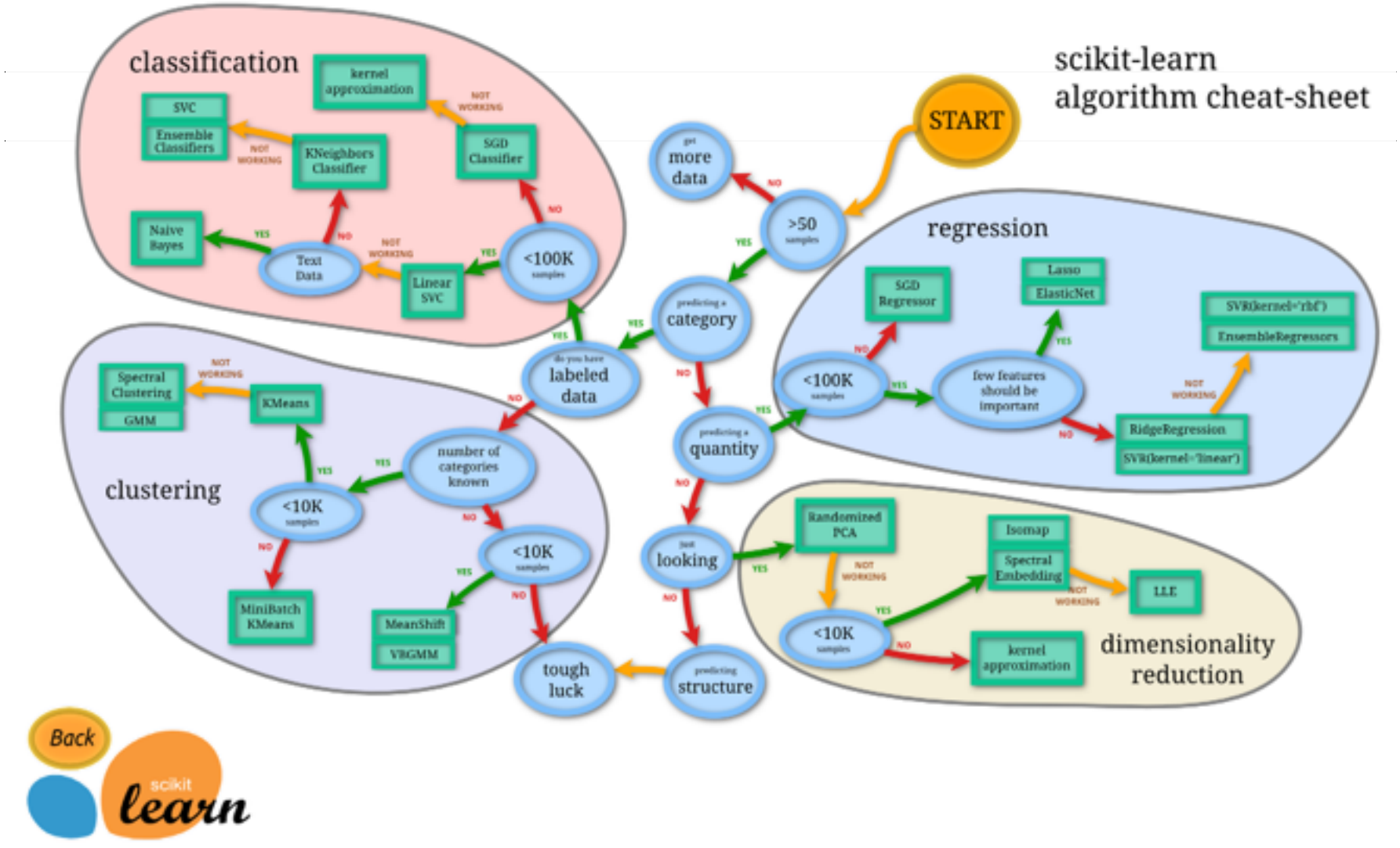
GENERAL ASSEMBLY

DATA SCIENCE
11 WEEK PART TIME COURSE

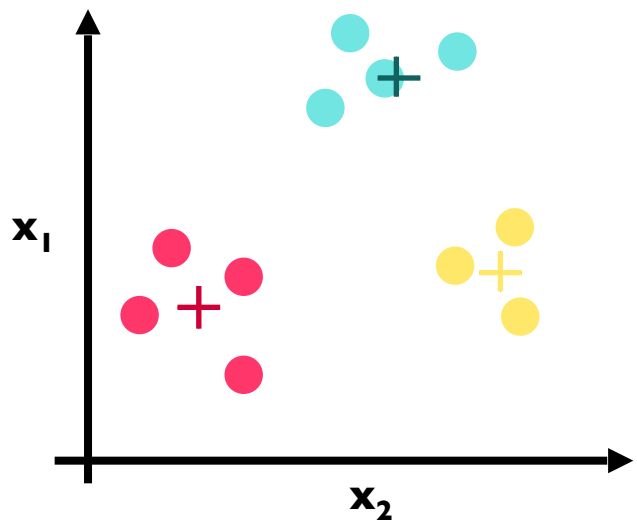
Week 4 - Clustering
Monday 18th April 2016

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

WHAT IS CLUSTERING AND WHY DO IT?


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
<10K samples
YES
MiniBatch KMeans
MeanShift
VBGMM
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

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



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.

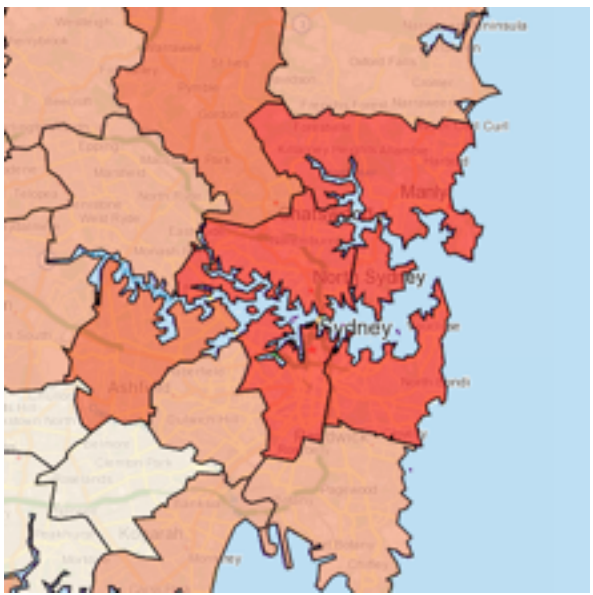
We have to decide what variables we will construct the groups on. What makes them different (or similar)?

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

Clustering provides a layer of abstraction from individual data points.

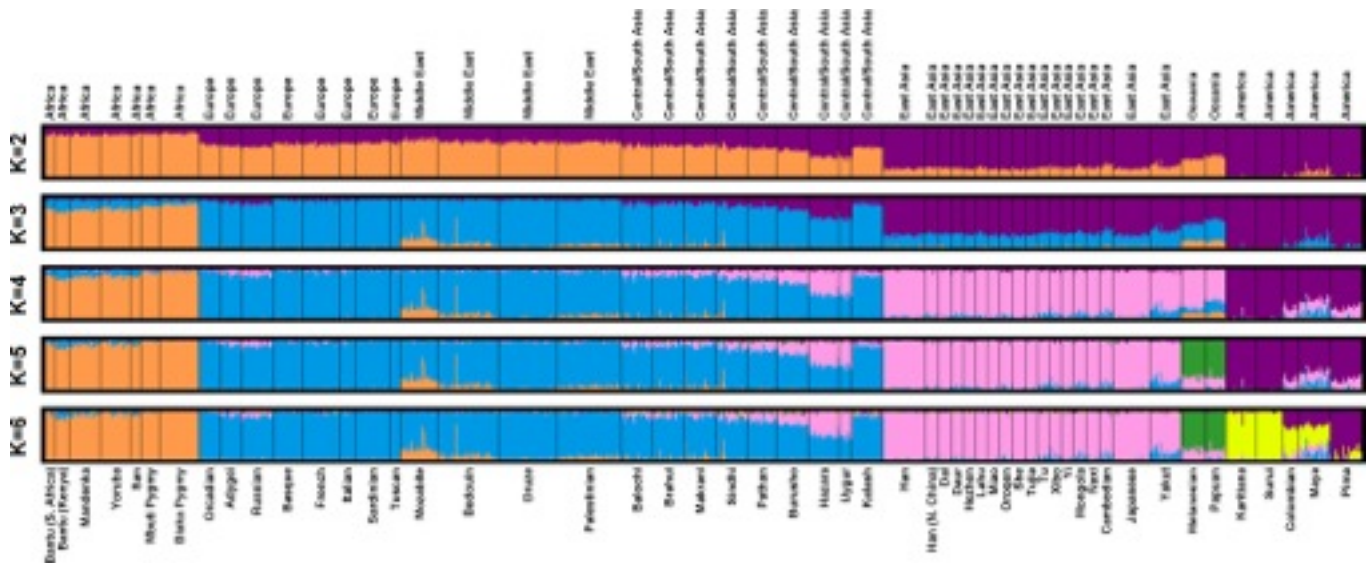
The goal is to extract and enhance the natural structure of the data

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



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

Genetics data can be clustered to identify ancestry



HOW DO WE CLUSTER DATA?

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

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)

The similarity criterion is determined by the measure we choose.

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

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

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)

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

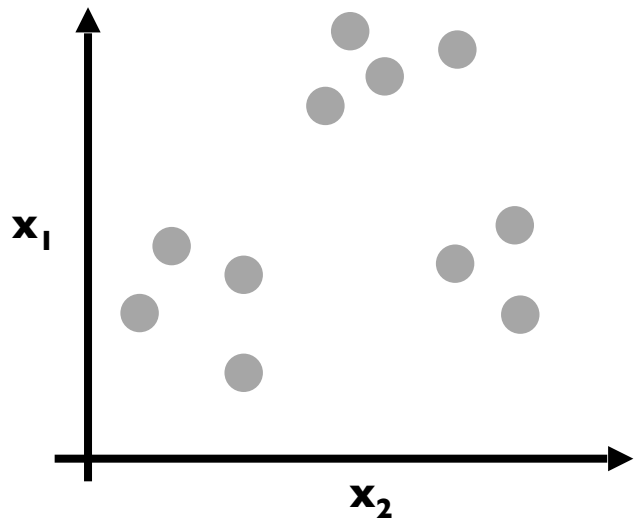
Stopping criteria can be based on the centroids (eg, if positions change by no more than ε) or on the points (eg, if no more than x% change clusters between iterations).

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

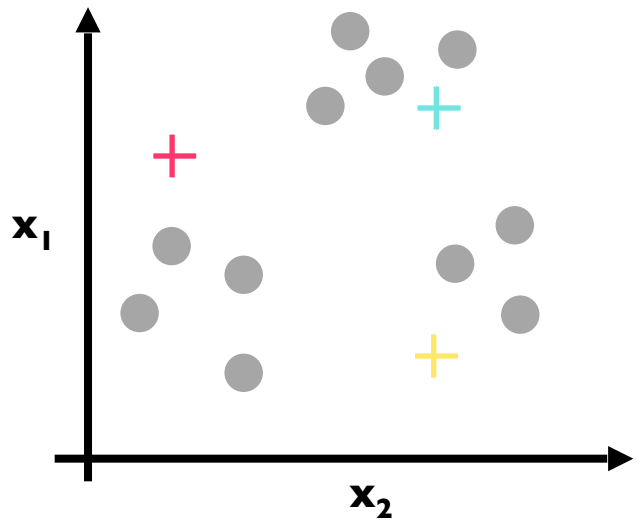


1) **Choose k initial centroids**

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

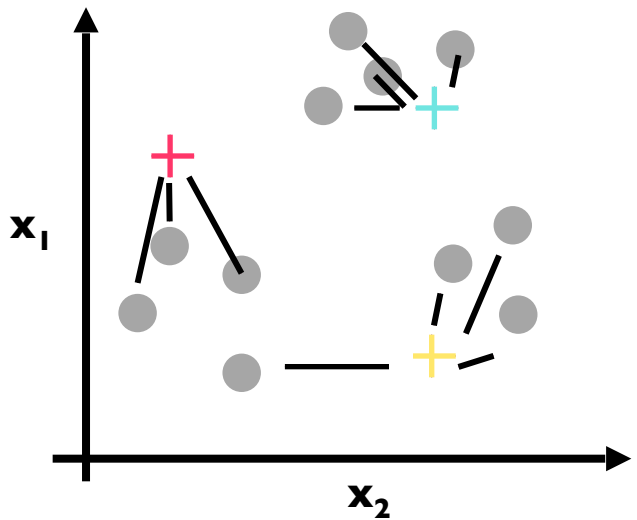


1) Choose k initial centroids

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

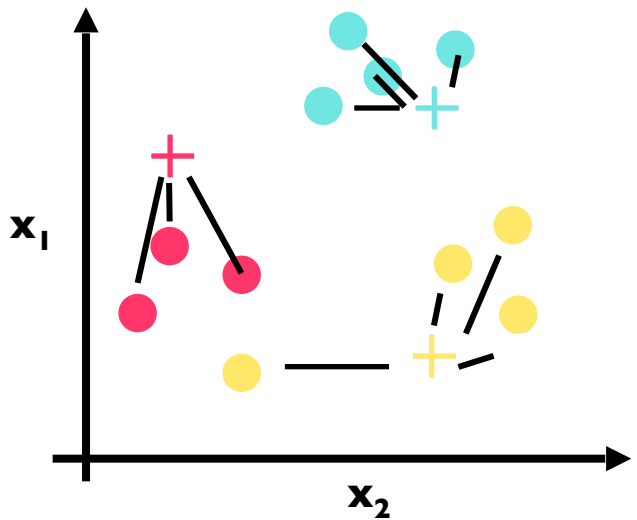


1) Choose k initial centroids

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

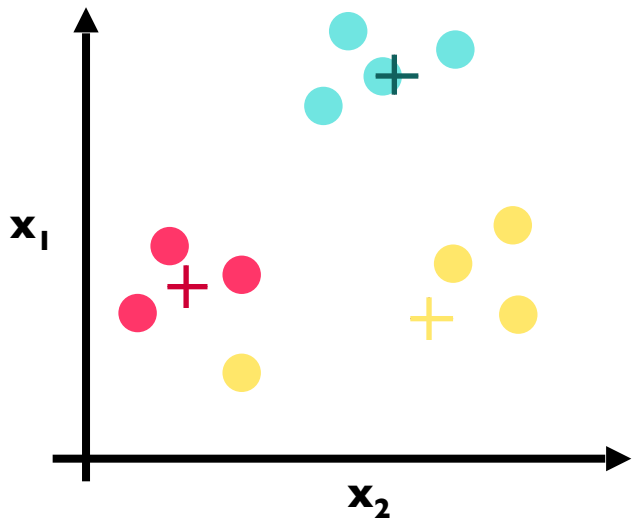


1) Choose k initial centroids

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

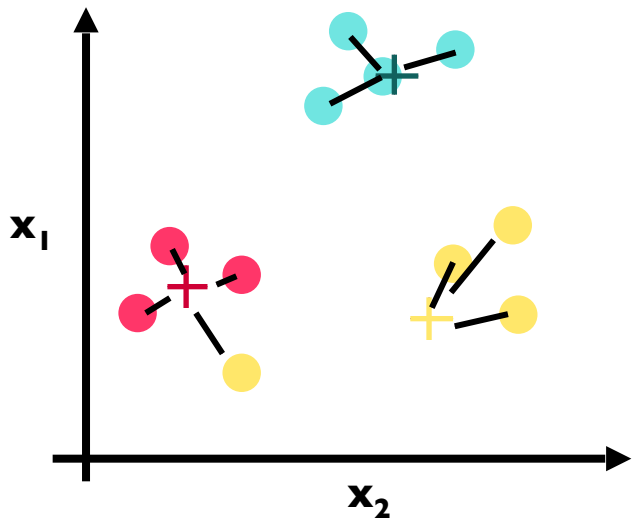


1) Choose k initial centroids

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

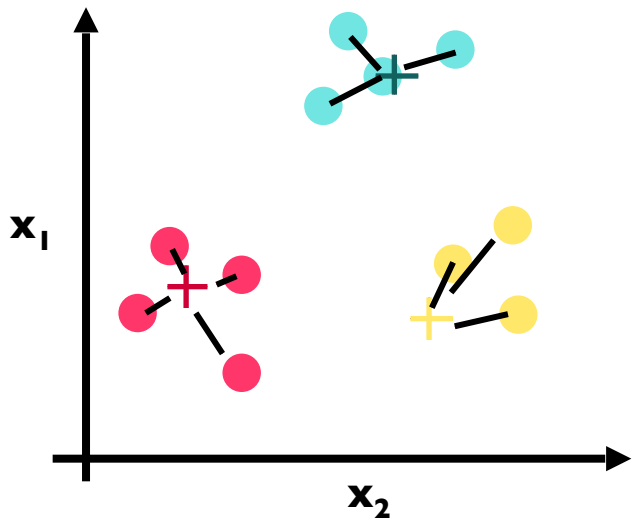


1) Choose k initial centroids

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

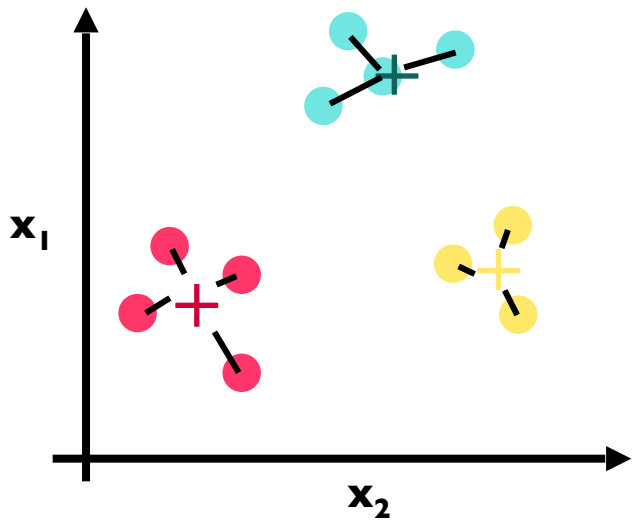


1) Choose k initial centroids

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

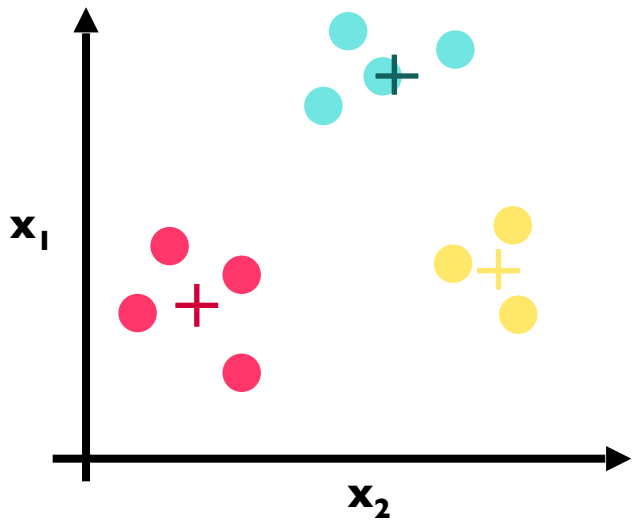


1) Choose k initial centroids

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



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.

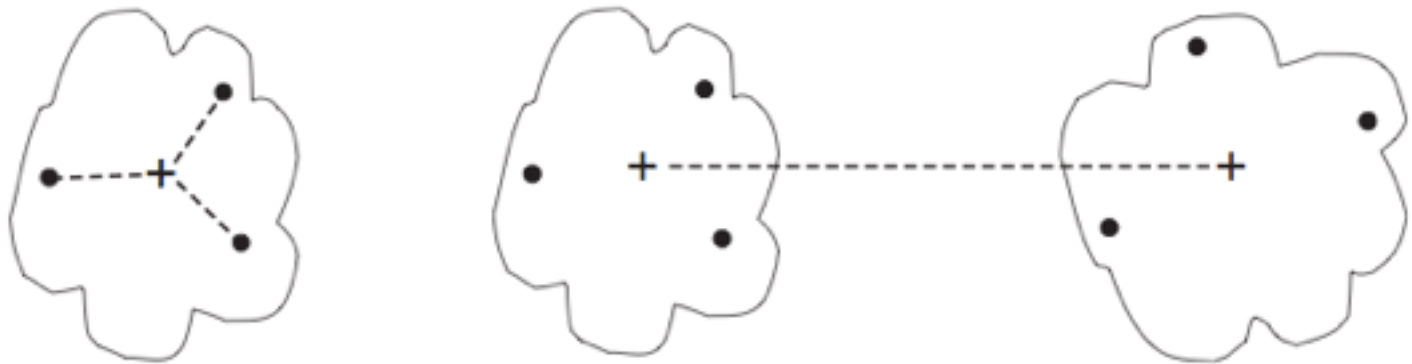
We will look at two validation metrics useful for partitional clustering, **cohesion** and **separation**.

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.

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

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

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.

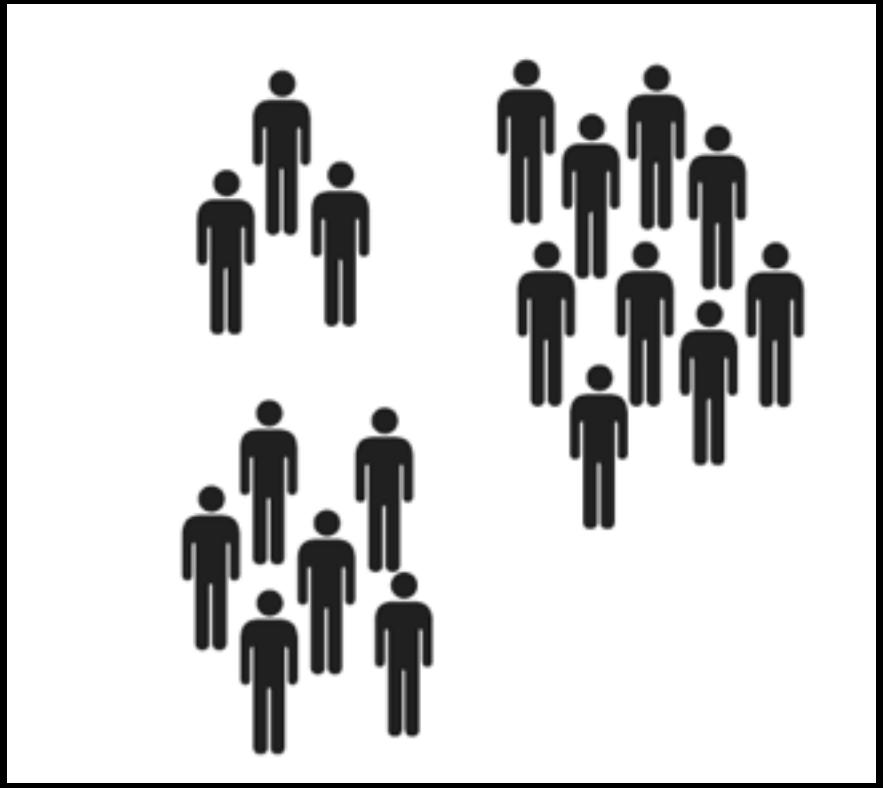
Strengths:

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

DISCUSSION TIME

- Homework 1
- Readings

DATA SCIENCE – Week 5 Day 1

HOMEWORK 1

Highlights

- Everyone mentioned the T-shaped Data Scientist !!!
- 'Diane stresses the importance of the human-computer symbiosis (i.e. let the computer do what it does best E.g. crunch numbers/models and calculations, and let humans do what they do best, E.g. Interpret models and relationships to extract meaning from which to base a decision/ recommendation.' – Arthur
- 'The most successful data scientists are those with substantial, deep expertise in at least one aspect of data science, be it statistics, big data, or business communication.' – Flavia

DATA SCIENCE – Week 5 Day 1

HOMEWORK 1

Highlights

- **‘Collabration and team skills are cruicial in succeeding as a data scientists. ’ – Louis**
- **‘There needs to be adequate integration between new data scientist hires and rest of the organization.’ – Anushi**
- **‘He did mention that at the beginning of intership he did realize that being only an statistician with not programming skills will not be enough to achieve his goals as it will be a limitation in the number of thing that he could do; therefore, worked hard in order to get those skills.’ – Claudia**

DATA SCIENCE – Week 5 Day 1

HOMEWORK 1

Highlights

- ‘That people and organizations tend to hire others like themselves. You need to try lots of things and talk to a lot of people. You need to fail sometimes and learn.’ – Jeremy
- ‘If you talk a problem though to a plastic duck, sometime you start to find the holes in your assumptions. Then you can plug them up.’ – Vijay
- ‘The most successful data scientists are those with substantial, deep expertise in at least one aspect of data science, be it statistics, big data, or business communication.’ – Jin

Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

Abstract—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

Index Terms—Data, analysis, visualization, enterprise.

1 Introduction

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals — with varied titles such as "business analyst", "data analyst" and "data scientist" — perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions within an organization impact the analysis process within the enterprise. Understanding how these issues shape analytic workflows can inform the design of future tools.

To better understand the day-to-day practices of enterprise analysts, we conducted semi-structured interviews with 35 analysts from sectors including healthcare, retail, finance, and social networking. We asked analysts to walk us through the typical tasks they perform, the tools they use, the challenges they encounter, and the organizational context in which analysis takes place.

In this paper, we present the results and analysis of these interviews. We find that our respondents are well-described by **three archetypes** that differ in terms of skill set and typical workflows. We find that ery and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. One avenue for achieving better interoperability is through systems that specify analysis or data processing operations in a high-level language, enabling retargeting across tools or platforms. We also note that the current lack of reusable workflows could be improved via less intrusive methods for recording data provenance.

2 Related Work

Many researchers have studied analysts and their processes within intelligence agencies [5, 18, 24, 25, 30]. This work characterizes intelligence analysts' process, discusses challenges within the process, and describes collaboration among analysts. Although there is much overlap in the high-level analytic process of intelligence and enterprise

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

How to actually learn any new programming concept

Essential

Changing Stuff and Seeing What Happens

O RLY?

@ThePracticalDev