Today (Week 9 Day 1) Agenda

- Announcements
- Quick review from last week
- Finish ML & ethics discussion / activity
- Slides Lecture: Introduction to k-means clustering
- R Lecture: get your feet wet with k-means in R
- Next time: more detail on k-means, how to chose “k”, pros and cons of this approach
- Homework will be assigned next class

Quick review

Supervised learning in practice

Observe: training data

Infer: the process that generated the data, relationships in the data

Predict: use this information to predict patterns in 'new' (or 'test') data

Unsupervised learning in practice

Observe: data → Use algorithms to identify patterns in the data → Outcomes: create classes

DATA

DGP collection sample design ethics scraping sourcing
domain expertise industry bias selection

EXPLORE

clean wrangle process visualize patterns
reduce features reduce complexity unsupervised learning

MODEL

$y = f(x)$ prediction v. inference selected features
test v. training data supervised learning

Conclusions (with caveats)

Let’s complete our group activity from last time

Cluster Analysis

DSI-EDA

Dr. Dorff

Week 9 Day 1

A little history…

SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS

J. MacQUEEN
University of California, Los Angeles

1. Introduction

The main purpose of this paper is to describe a process for partitioning N-dimensional population into k sets on the basis of a sample. The proc which is called 'k-means,' appears to give partitions which are reasona efficient in the sense of within-class variance. That is, if p is the probability m function for the population, $S = \{S_1, S_2, \cdots, S_k\}$ is a partition of E_N, and $i = 1, 2, \cdots, k$, is the conditional mean of p over the set S_i, then $w^2(S)$ $\sum_{i=1}^{k} \int_{S_i} |z - u_i|^2 \, dp(z)$ tends to be low for the partitions S generated by method. We say 'tends to be low,' primarily because of intuitive consideratio

[James B. MacQueen]
Published 1967

Go read the proofs!

$\cdots$. *If* $\lim_{n\to\infty} \sum_{i=1}^{k} p(S_i(y^n))|y_i^n - u_i(y^n)| = 0$, *then* $\sum_{i=1}^{k} p(S_i(x))|x_i - u_i(x^n)| = 0$ *and each point* x_i *in the k-tuple* $(x_1, x_2, \cdots, x_k)$ *is distinct from the others.*

Lemmas 1 and 2 above are primarily technical in nature. The heart of the proofs of theorems 1 and 2 is the following application of martingale theory.

LEMMA 3. *Let* $t_1, t_2, \cdots$, *and* $\xi_1, \xi_2, \cdots$, *be given sequences of random vari* *and for each* $n = 1, 2, \cdots$, *let* t_n *and* ξ_n *be measurable with respect to* β_n *v* $\beta_1 \subset \beta_2 \subset \cdots$ *is a monotone increasing sequence of* σ*-fields (belonging to the u* *lying probability sp* (i) $|t_n| \lesssim K < \infty$, (i *quences of random* $s_n = \sum_{i=1}^{n} (t_i - E($

$$(2.8) \qquad E(W(x^{n+1})|\omega_n] \leq W(x^n) - \sum_{j=1}^{k} |x_j^n - u_j^n|^2 (p_j^n)^2 (2w_j^n + 1)/(w_j^n + 1)^2 + \sum_{j=1}^{k} \sigma_{n,j}^2 (p_j^n)^2/(w_j^n + 1)^2,$$

where $\sigma_{n,j}^2 = \int_{S_j^n} |z - u_j^n|^2 \, dp(z)/p_j^n$.

Since we are assuming $p(R) = 1$, certainly $W(x^n)$ is a.s. bounded, as is $\sigma_{n,j}^2$. We now show that

$$(2.9) \qquad \sum_n (p_j^n)^2/(w_j^n + 1)^2$$

converges a.s. for each $j = 1, 2, \cdots, k$, thereby showing that

$$(2.10) \qquad \sum_n \left(\sum_{j=1}^{k} [\sigma_{n,j}^2 (p_j^n)^2/(w_j^n + 1)^2] \right)$$

converges a.s. Then lemma 3 can be applied with $t_n = W(x^n)$ and $\xi_n = \sum_{j=1}^{k} \sigma_{n,j}^2 (p_j^n)^2/(w_j^n + 1)^2$.

It suffices to consider the convergence of

$$(2.11) \qquad \sum_{n \geq 2} (p_j^n)^2/[(\beta + 1 + w_j^n)(\beta + 1 + w_j^{n+1})]$$

$$(2.5) \qquad E[W(x^{n+1})|\omega_n] = E\left[\sum_{i=1}^{k} \int_{S_i^{n+1}} |z - x_i^{n+1}|^2 \, dp(z)|\omega_n \right]$$

$$\leq E\left[\sum_{i=1}^{k} \int_{S_i^n} |z - x_i^{n+1}|^2 \, dp(z)|\omega_n \right]$$

$$= \sum_{j=1}^{k} E\left[\sum_{i=1}^{k} \int_{S_i^n} |z - x_i^{n+1}|^2 \, dp(z)|A_j^n, \omega_n \right] p_j^n.$$

r $i \neq j$. Thus we obtain

$$\leq W(x^n) - \sum_{j=1}^{k} \left(\int_{S_j^n} |z - x_j^n|^2 \, dp(z) \right) p_j^n$$

$$+ \sum_{j=1}^{k} E\left[\int_{S_j^n} |z - x_j^{n+1}|^2 \, dp(z)|A_j^n, \omega_n \right] p_j^n.$$

General Motivation for K-means

- You want to learn more about groupings in your data.
- You want your to see if your data can be easily divided into groups that are meaningful, useful, or both.
- You want to find groupings in your data that capture the ‘natural’ structure of your data.
    - Can be very useful to speak to colleagues and or researchers!

Real world cluster analysis examples

- Cancer research: for classifying patients into subgroups according their gene expression profile
- Biology: cluster analysis to analyze large amounts of genetic information to find groups of genes with similar functions
- City planning: for identifying groups of houses according to their type, value, and location
- Marketing: for market segmentation, aka identifying subgroups of customers or users with similar profiles

Basic Concept

- Say you are given a data set where each observed example has a set of features, but has **no** labels

- We can find groups of data in our dataset which are similar to one another -- what we call **clusters**.

- K-Means is an algorithm that takes a dataset and a constant, **k,** and returns **k** # of centroids (which define clusters of data in the dataset with data points that are similar to one another).

K-means Algorithm

1. Input: *K* and a set of points x_1…..x_n
2. Place centroids at random locations ("random partitioning")
3. Minimize distance between centroids and points
4. Move centroids to 'center' of points (assign points to centroids)
5. Minimize distance between centroids (again)
6. Repeat until convergence

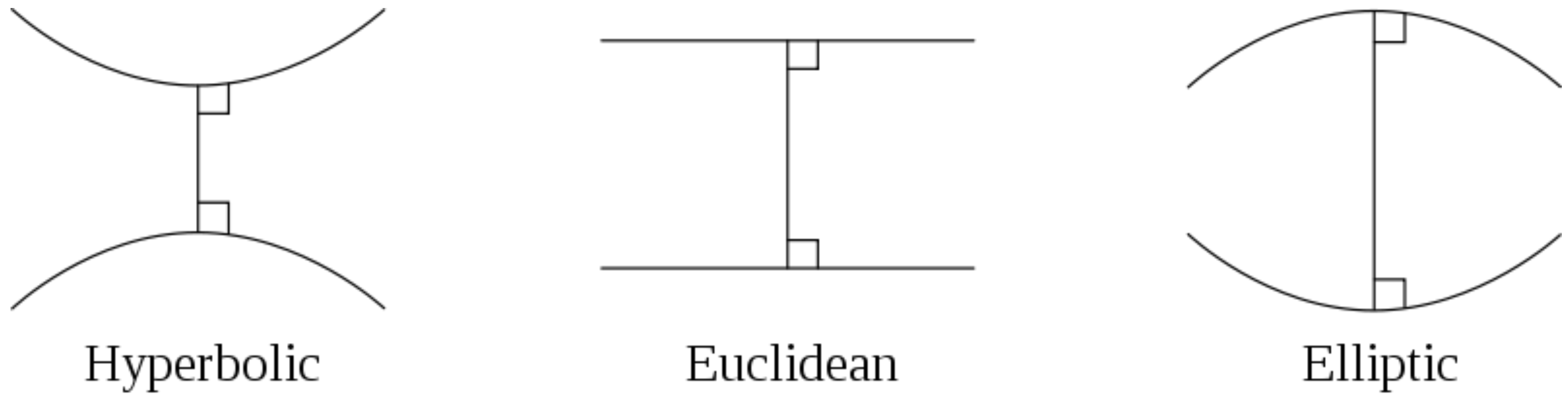
- Summary: to process the learning data, the K-means algorithm starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

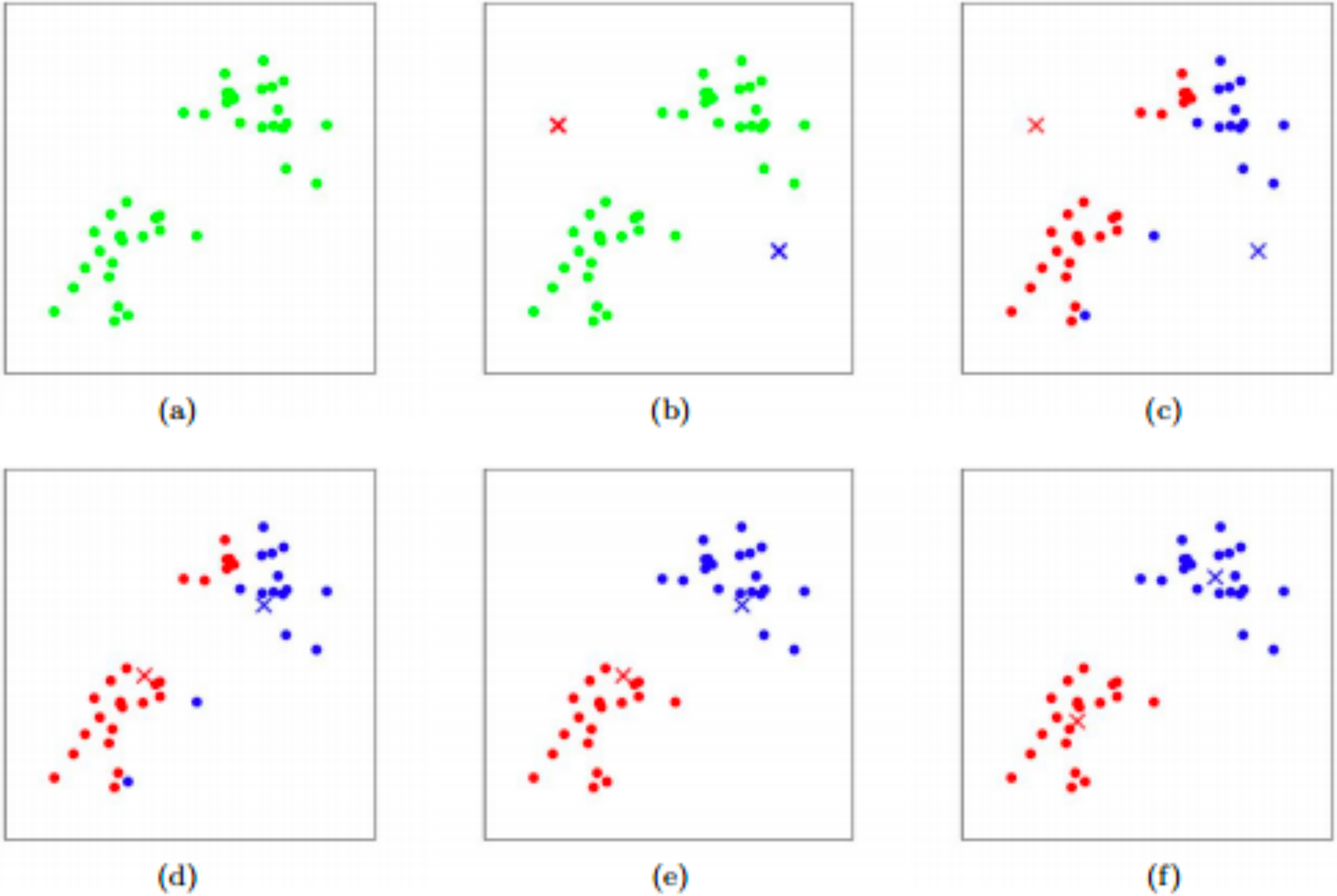
K-means converges

The algorithm stops creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

Side note: we are referring to ‘distance’ in Euclidean space

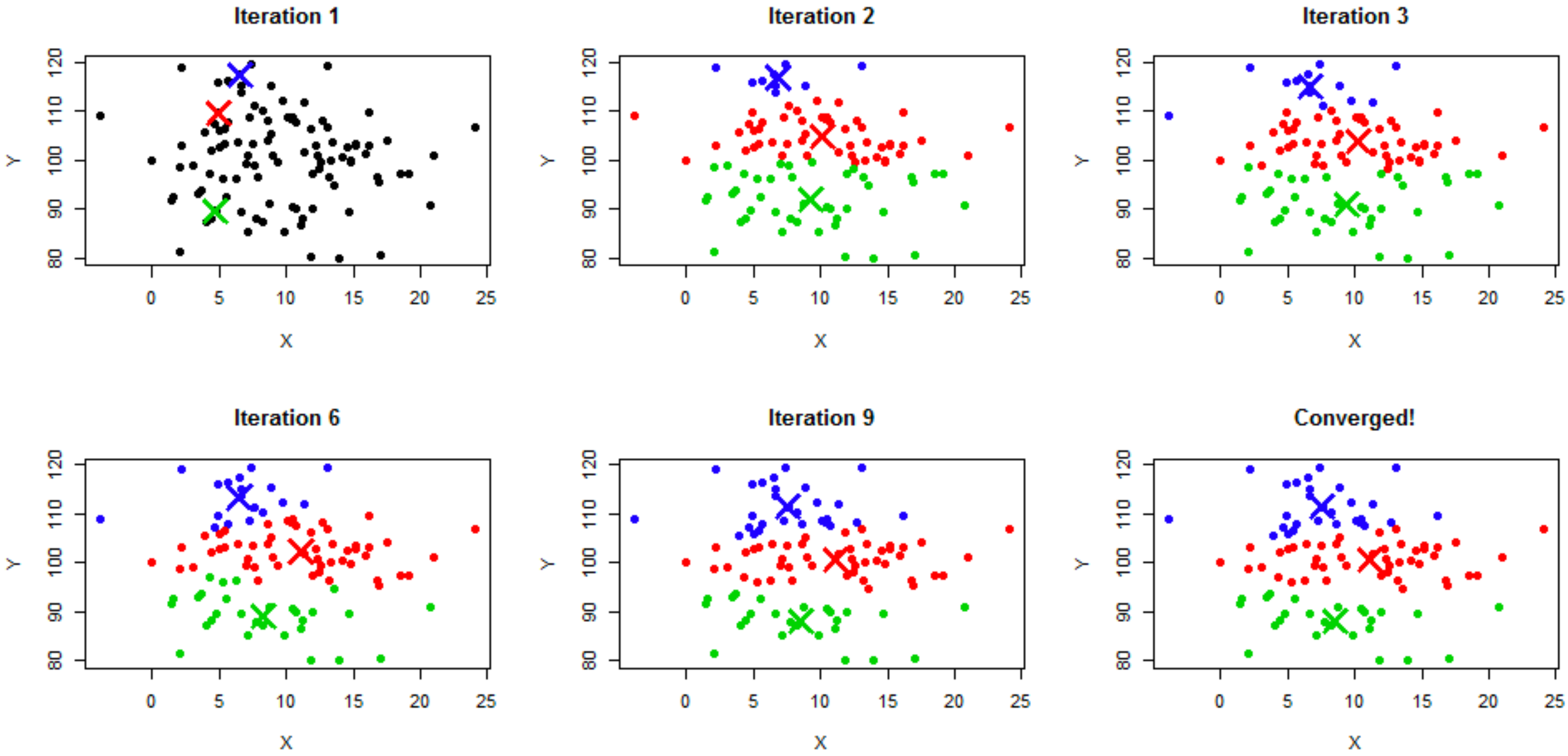




(a)
(b)
(c)
(d)
(e)
(f)

Example

- The example in the previous slide is done in 2 dimensions with K=2
- This means it was using only 2 features from the data and we assumed that there were 2 clusters, or classes, in the data.
- Obviously it is quite easy to see how this can become a bit more complicated with more features and classes!


Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y
0 5 10 15 20 25
80 90 100 110 120

Summary

- Today:
  - learn the basic intuition behind the k-means algorithm
  - learn what it means for a k-means algorithm to converge
  - understand the meaning of K
  - how K-means ”works” in R using a demo dataset

- Next class:
  - pros and cons of K-means
  - How to choose K
  - how k-means ‘works’ using more complex data

Related book

- Practical Guide to Cluster Analysis in R

- “Clustering is one of the important data mining methods for discovering knowledge in multidimensional data. The goal of clustering is to identify pattern or groups of similar objects within a data set of interest.”