Data Science Methods for Clean Energy Research

Week 10 L1: Unsupervised Learning
March 6, 2017

UNIVERSITY *of* WASHINGTON

Outline

- Quick review from last time
- A brief note on the support vector machine
- Comparison of supervised vs. unsupervised learning
- Principal components analysis (PCA)
- Clustering
- Wrap up

Topics last time

> Decision trees
  - Classification and regression trees
  - Tree depth and relationship to bias/variance concepts

> Ensemble methods
  - Bagging
  - Random forest

> Visualization of DTs

<u>Big picture concepts:</u>

- DT can outperform regression or classification, especially when relationships are complex or nonlinear
- DT are highly dependent on training data used, so ensemble methods are strongly advised

Visualization of the DT in Python

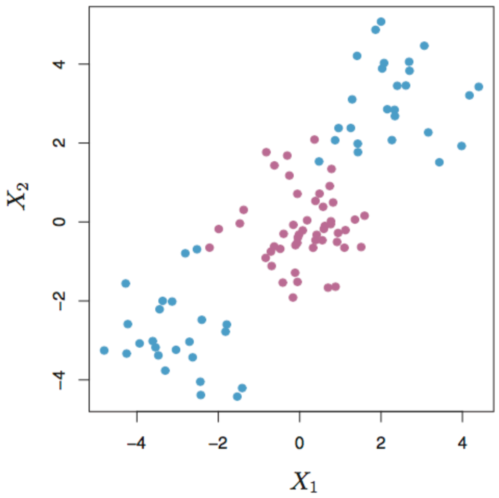
Support vector machine (SVM)

> **I had hoped for time to discuss an advanced classification technique called SVM: 1 slide instead ☹**

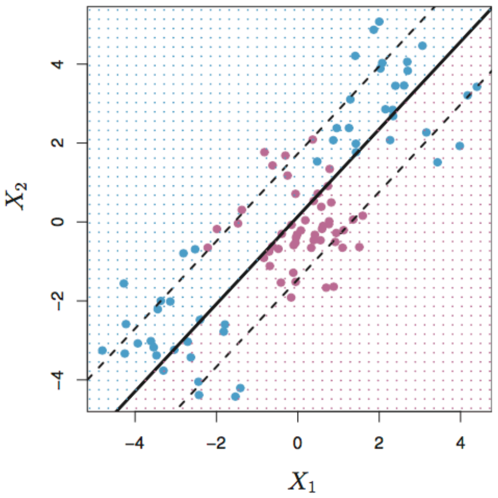
> **Qualitative idea:** construct decision boundaries when there is a nonlinear relationship between features

Data from figs 9.8/9.9

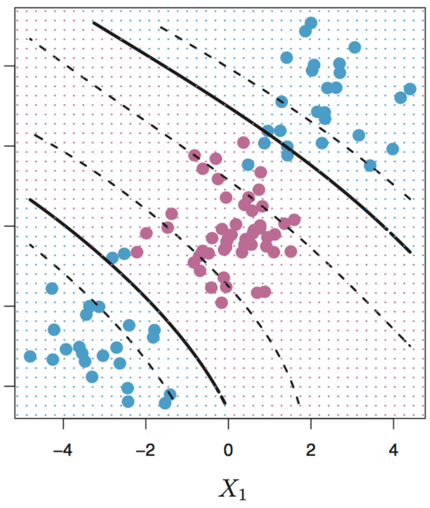
2 features, 2 class



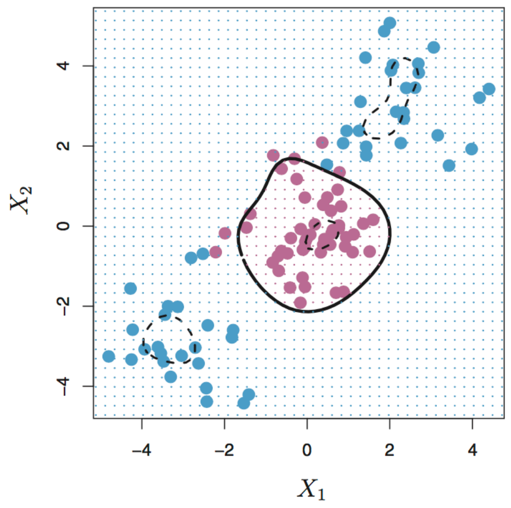
Linear boundary



SVM:
curved boundary

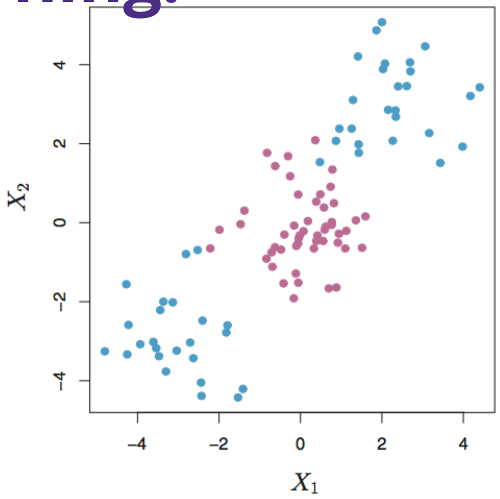


SVM:
radial boundary



Supervised vs. unsupervised learning

- **Consider the X1/X2 data below. Imagine the scenario in which we did not know in advance the response Y (**the color of the circle or class**)**
- **Main goal:**
  - **Use a large set of features (X) [**there are no longer any responses, Y!**] and determine how the data may be grouped together**
- **Central challenge in unsupervised learning:**
  - **How to validate the data?**
  - **Or… Is there any "answer" ?**



Two concepts in unsupervised learning

1. **Find a way to group the data in reduced dimensionality so that sub-groups describe most of the variance:** principal components analysis (PCA)

2. **Find similar sub-groups of data within our total data set:** clustering

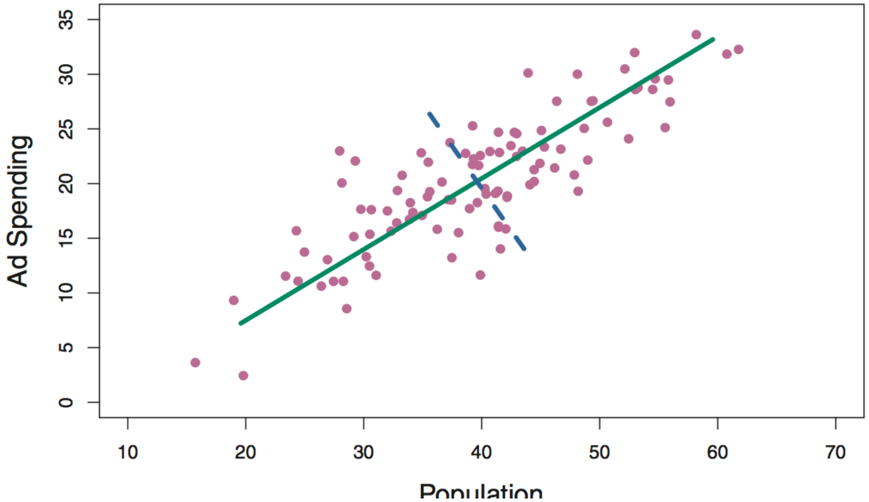
PCA analysis

- > The concept of a principal component
  - Suppose we have two descriptors (Fig 6.14), which are related somehow
  - We seek the relationship that captures most of the variance in a linear summation of all of our descriptors
    - > What are the coefficients (ϕ_{11}, ϕ_{21}) that maximize $Var(\phi_{11}X_1 + \phi_{21}X_2)$ given $\phi_{11}{}^2 + \phi_{21}{}^2 = 1$
  - If we are successful, then the coefficients tell us something interesting about how the variables are related…

The **first** principal component for this data is shown by the green line and given by :

$$Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}}).$$

$$(\phi_{11} = 0.839\,, \phi_{21} = 0.544)$$



Eq 6.19 / Fig 6.14

PCA definitions and concepts

- **One way to look for the relationship between variables that maximizes the variance is to look at the matrix of scatter plots (**not feasible in high dimensionality**)**
- **The first principal component (Z_1) is determined by solving**

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1. \qquad (10.3)$$

- **The coefficients ϕ are known as** loadings **for each of the responses**
- **The PC loadings are often the most informative outcome of our PCA**

The 2nd principal component

- **But what if there are other variables that help describe the variance in our observables?**
- **The 2nd PC (Z_2) is the linear combination of $\phi_{j2}X_j$ that has maximum variance:** must be orthogonal, or totally uncorrelated to Z_1

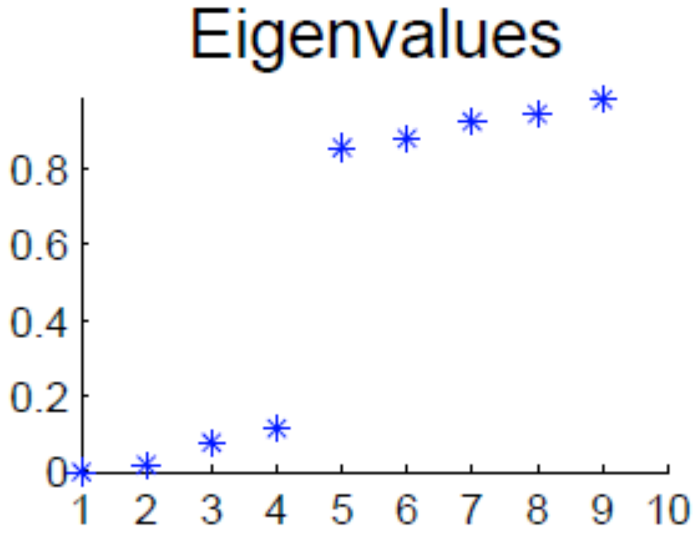
$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip}, \qquad (10.4)$$

- **The orthogonality of the 2nd PC comes in as an additional constraint**

PCA summary

- **There are many principal components, they are usually determined through eigen decomposition of the covariance matrix of X**
- **The eigenvalues, when ordered, often display a** spectral gap, **which can be useful in determining which set are the more useful to focus on**

Random internet picture:



PCA analysis results may depend on the scale of X$_i$!

- As in Ridge and LASSO regression, the ultimate answer of your PCA analysis is not scale invariant!
- Prior to conducting PCA you should:
  1. **Set all the means of each X$_i$ equal to zero:** transformation $\acute{X}_i = X_i - \bar{X}$
  2. **Set the variance of each X$_i$ equal to one:** transformation $\acute{X}_i = X_i/\sigma_i$
- We are looking for variables that explain the variance and don't want the order of magnitude (or choice of units!) to numerically swamp out an important effect

How to use principal components

> PCA is usually an “exploratory” method

> Proportion of variance explained and what this means in practice

> You could potentially bootstrap your PCA if you have enough data, but take care

Python implementation and tips

- PCA can become expensive to calculate, especially as the data set size grows
  - As a result there are often many methods to pick from
- Implementation is easy and many tutorials online, sklearn PCA is well supported
- Advice
  - Go slow and use a subset of your data (if you have many points)
  - Use PCA as a guide and as an exploratory tool
  - Constantly interrogate the results and ask if they make sense! You don't have "test set error" to fall back on, so you need to use your brain!

PCA questions?

Clustering

- **Our other main tool in unsupervised learning is** clustering
- **Clustering seeks to group items by minimizing a** distance metric **between groups of observations or groups of features**
- **K means**
  - **Algorithm**
  - **How to use it**
  - **What the results mean**
  - **Warnings: size K , many trials**
- **Hierarchical clustering**
- **Lots of other approaches // clustering vs PCA**
- **How to implement it in Python**

K-means clustering, a simple algorithm

- **One of the most common clustering methods**
- **Requires, as an input, specification of the final number of clusters you want** (K)
- **Rules:**
  - **Each observation must be placed in at least one of the clusters**
  - **No clusters may overlap, each observation can only be placed in a single cluster**
  - **The goal is to minimize the** variance of observations within each of the clusters

Same data, different values of K

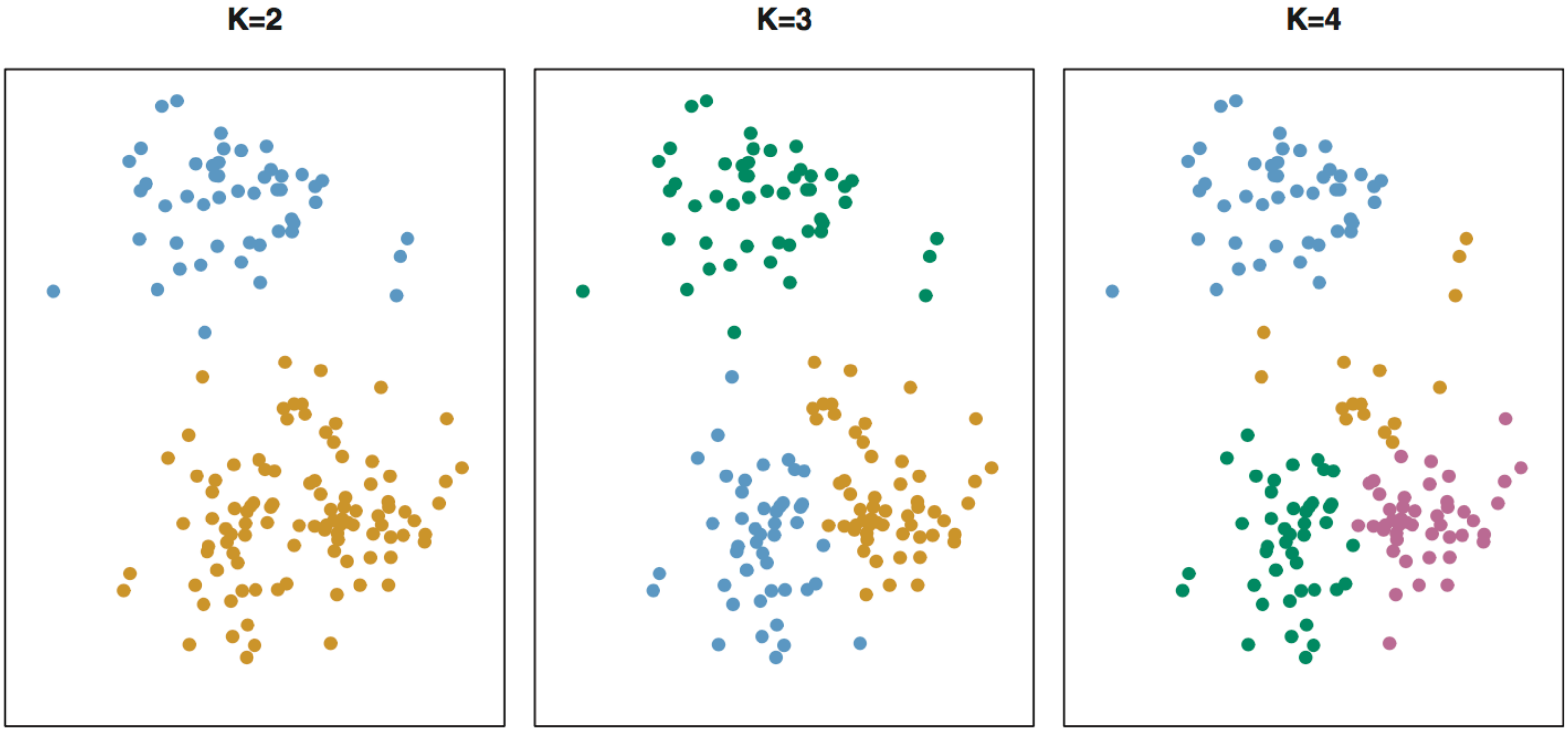
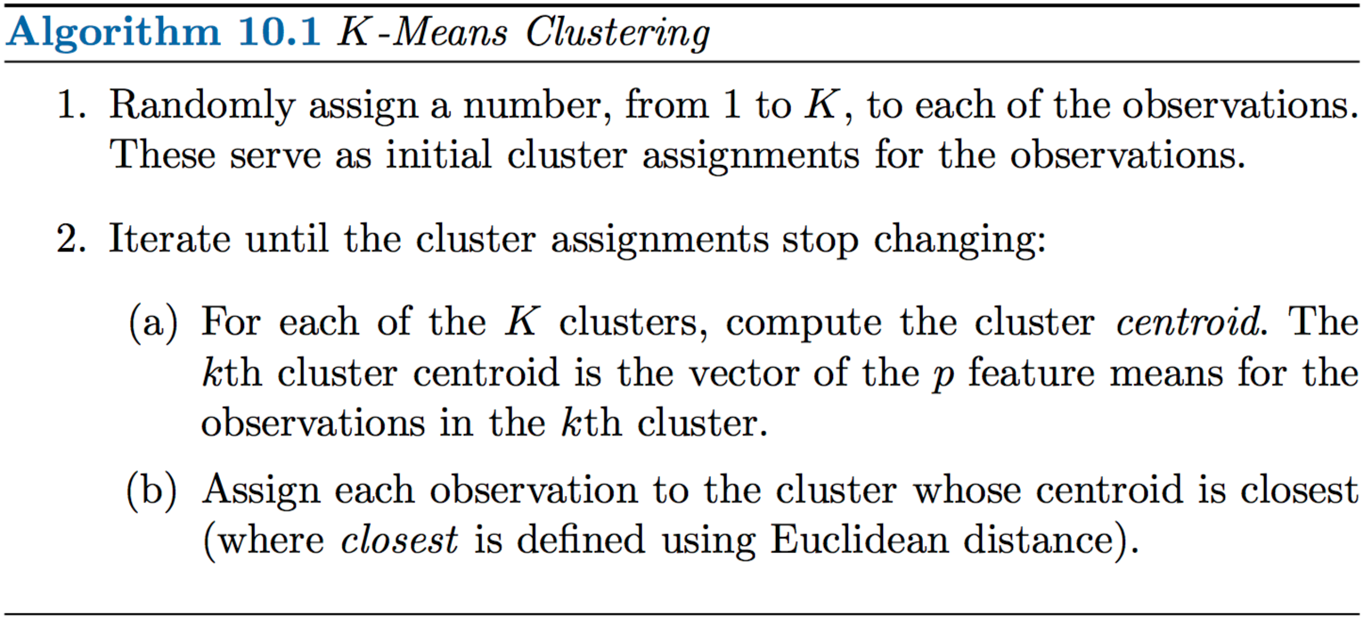


FIGURE 10.5. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.*

Algorithm for K-means



Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2, \qquad (10.12)$$

Algorithm for K-means (e.g., K=3)

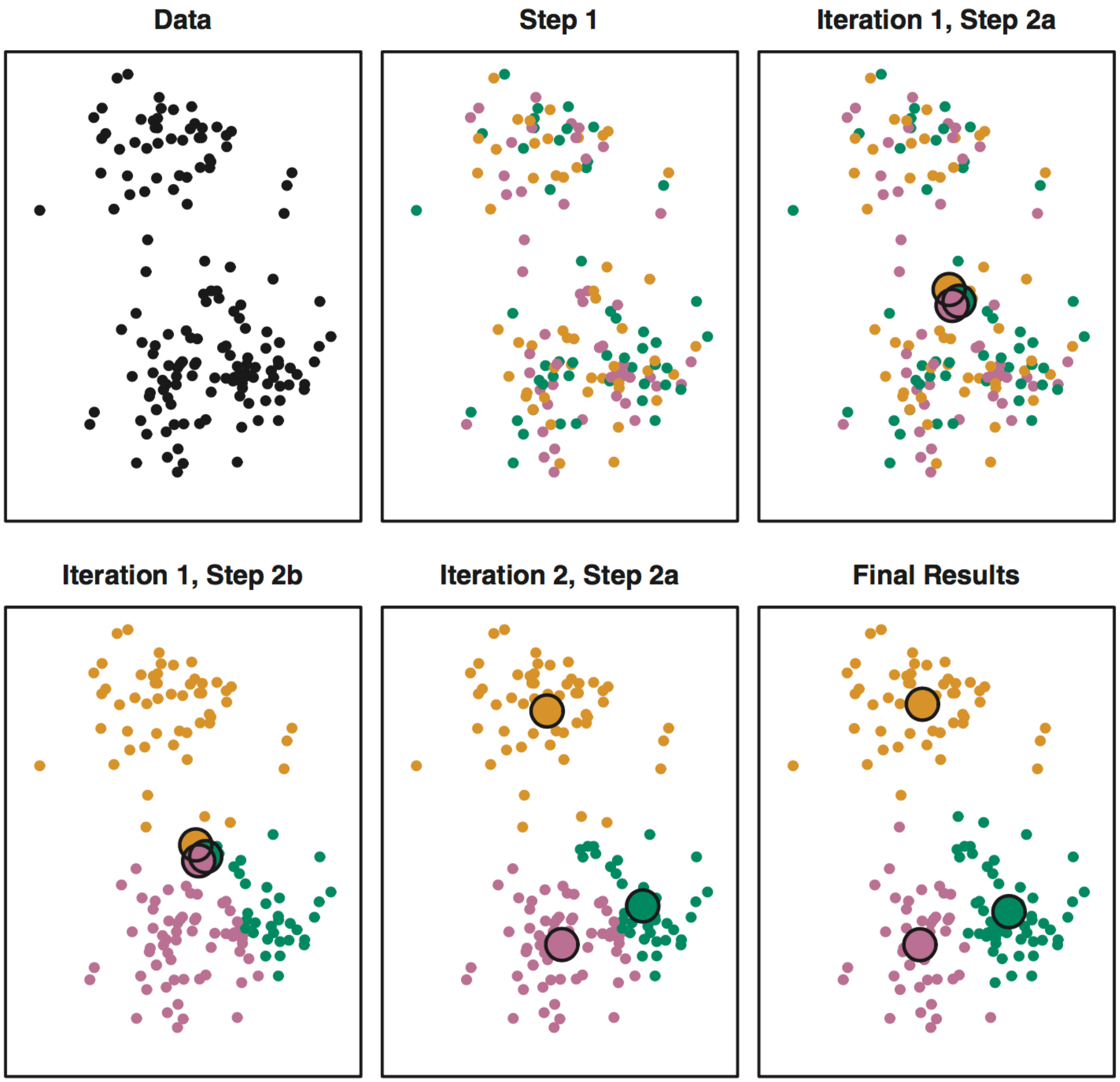


FIGURE 10.6. *The progress of the K-means algorithm on the example of Fig-*

Importance of sampling

- **K-means clustering is a stochastic process**
  - **Initial random assignment of data to classes**
  - **The optimization scheme leads to a** local optimization **, it is not guaranteed to find the global minimum**
- **Process**
  - **Re-seed different initial clusters and repeat optimization of cluster centers / assignments**
  - **Monitor the sum of distances (**each point's distance from each cluster center**) as your error metric**
  - **Choose clustering arrangement with lowest error**

Importance of sampling

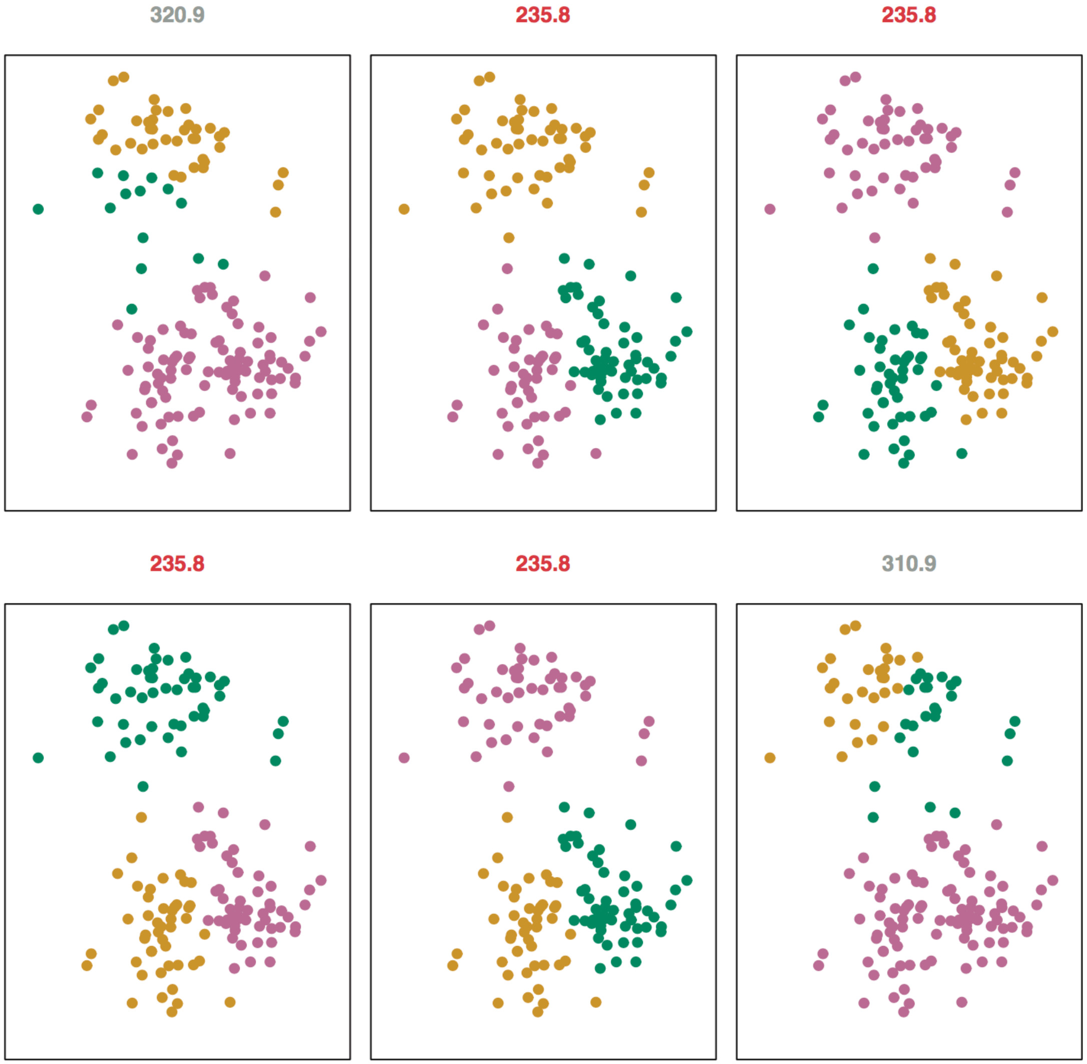


FIGURE 10.7. *K-means clustering performed six times on the data from Fig-*

What to do with your clusters

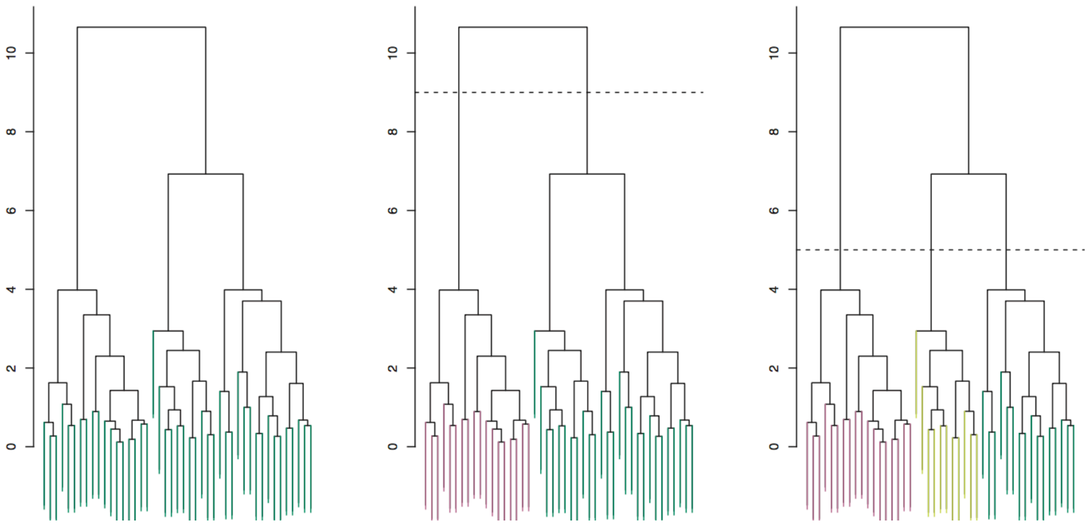
- **As in PCA, clustering is an** exploratory **analysis tool**
- **If you have clustered your** observations **(most common): you can interrogate the different clusters to see if they have common features**
  - **For observations with many features, you can also cluster the features and look @ common observations…**
- **In K-means, you have to choose the number of clusters:** you must assess the degree to which this choice makes an impact on your scientific conclusions!

Performing K-means clustering

- The module `sklearn.clustering` can perform K-means and has many variants
- Quick Q: what is a good strategy to learn how to implement a new method like K-means?
- Take care as with PCA if the size of your data set grows, the computational cost to complete the clustering can become prohibitive…

Hierarchical clustering vs K–means

- **The main limitation of K-means,**
- **Section 10.3.2 discusses** hierarchical clustering**, an approach that clusters all the data using a tree type structure**
  - **Choice of how many clusters can be made after the clustering is completed…**
- **See ISL for more detail**



Clustering vs PCA

- **Simple definition in ISL (p385):**
  - “*PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance*”
  - “Clustering looks to find homogeneous subgroups among the observations”
- **Unsupervised learning, especially use of results, can be a bit of an art…**
  - **In some cases, it might be appropriate to look at both approaches**

(if time) Feedback on class...