Queen Mary
University of London

DR ADRIAN BEVAN

MULTIVARIATE ANALYSIS AND ITS USE IN HIGH ENERGY PHYSICS

ADDITIONAL NOTES: UNSUPERVISED LEARNING

LECTURE PLAN

- Introduction
- K-means algorithm
- Anomaly detection
- Summary
- Suggested reading

INTRODUCTION

- We have been using labeled sets of data with a loss function to compare labels, Y, against model predictions for the data X.
  - This is supervised learning.
  - Can be thought of as computing a conditional probability P(Y|X).
- For unsupervised learning we don't have the luxury of labels and we want to learn the model in the absence of that information.
  - The goal of unsupervised learning is to infer the probability distribution P(X) from the data without using labels.

INTRODUCTION

- Many methods exist as it can be difficult to determine the accuracy of unsupervised methods.
  - Clustering methods such as K nearest neighbour (or K-means)
- Heuristic arguments are used to motivate models and justify the quality of outcomes.

K-MEANS ALGORITHM

- The aim is to determine the centroid positions C of K clusters in the data containing N examples using a Euclidean distance from the cluster mean to some data example.
- The variance of the clusters is minimised in order to determine the corresponding means of the cluster.

K-MEANS ALGORITHM

This follows Section 14.3.6 of Hastie et al.

- Step 1:
  - Given C compute the total cluster variance and minimise this with respect to the means of the clusters.

$$\min_{c,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2$$

  - This gives the current mean positions of the clusters.
- Step 2:
  - Given a set of means m, minimise these by assigning elements to the closest current cluster mean. i.e.

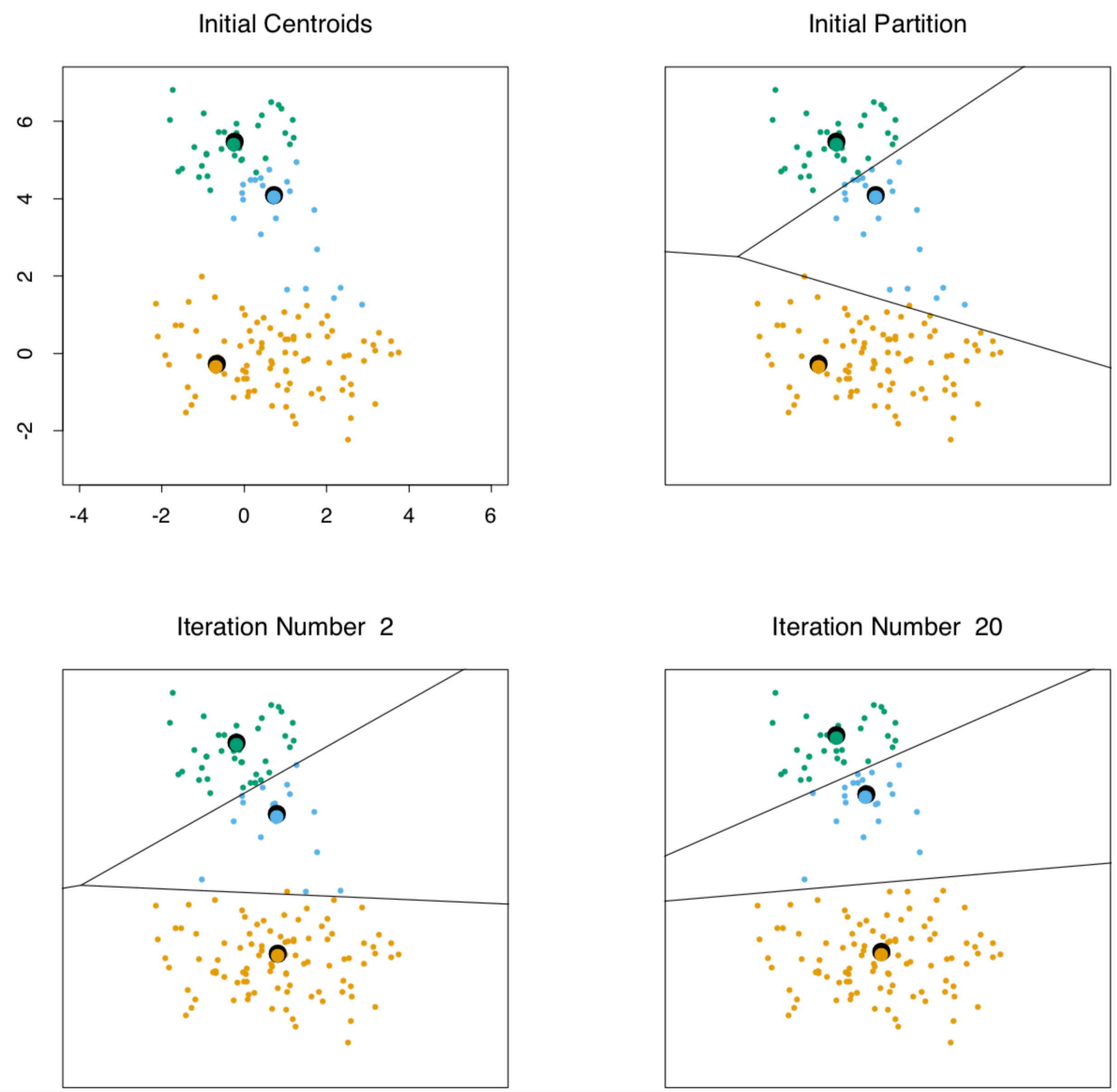
$$C(i) = argmin_{1 \leq k \leq K} ||x_i - m_k||^2$$

- Step 3:
  - Iterate until the assignments stabilise.

K-MEANS ALGORITHM

This follows Section 14.3.6 of Hastie et al.

- This example shows successive iterations of the K-means algorithm to a set of data with K=3.



Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 14

This algorithm has the number of clusters, K, as a parameter.

Clustering results will depend on the choice of K.

ANOMALY DETECTION

- There are a number of anomaly detection algorithms that are available. Outlier detection is a common problem, but this is not something that has received much attention in the community compared with other problems.
- Some references that may be of interest as a starting point in this area:
  - Goldstein and Uchida, A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173
  - Ahmad et al., Unsupervised real-time anomaly detection for streaming data, Neurocomputing 262 (2017) 134-147.
  - Schlegl et al., Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, proceedings of IPMI 2017.
  - Chalapathy et al., Anomaly Detection using One-Class Neural Networks, https://arxiv.org/abs/1802.06360.

SUMMARY

- Unsupervised learning complements the techniques of supervised learning that we have discussed in this short course.
  - We discussed only one algorithm, a simple K-means clustering approach.
  - More complicated algorithms exist.
- Unsupervised learning can be applied to a wide range of situations.
  - The clustering example shown here is something that lends it self for discriminating between signals and noise in a detector, for example.

SUGGESTED READING (NON-HEP)

- e.g. see Chapter 14 of Hastie, Tibshirani and Friedman, Elements of Statistical Learning and references therein.