



CoGrammar

K-means

**SKILLS
FOR LIFE**

SKILLS BOOTCAMPS



Department
for Education

Data Science Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
(FBV: Mutual Respect.)
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Open Classes.
You can submit these questions here: [Open Class Questions](#)

Data Science Lecture Housekeeping cont.

- For all **non-academic questions**, please submit a query:
www.hyperiondev.com/support
- Report a **safeguarding** incident:
www.hyperiondev.com/safeguardreporting
- We would love your **feedback** on lectures: [Feedback on Lectures](#)



Prestigious Co-Certification Opportunities


New Partnerships!

- **University of Manchester & Imperial College London** join our circle along with The University of Nottingham Online.

Exclusive Opportunity:

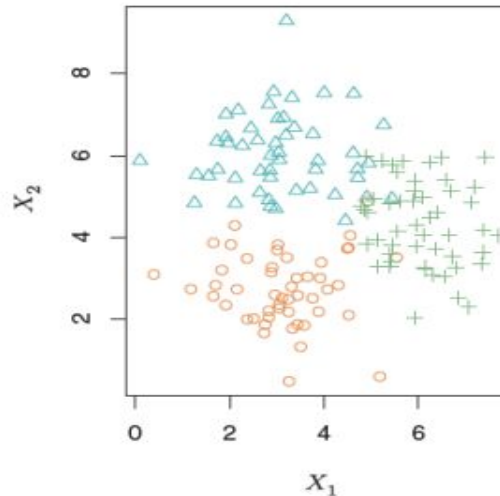
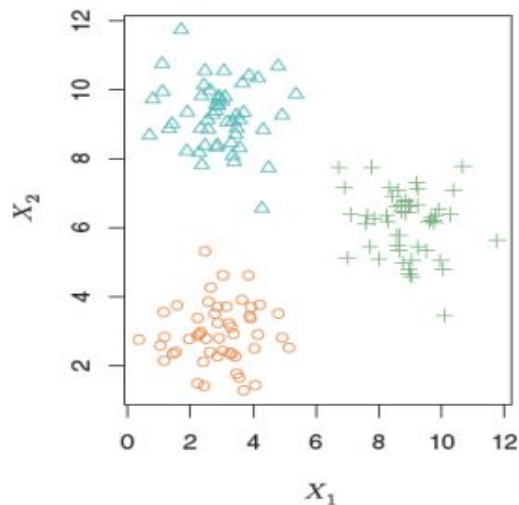
- Co-certification spots awarded on a first-come basis.
- Meet the criteria early to gain eligibility for the co-certification.

New Deadlines:

- **11 March 2024:** 112 GLH & BYB tasks completion.
 - **18 March 2024:** Record interview invitation or self-employment.
 - **15 July 2024:** Submit verified job offer or new contract.
- 

Introduction to Clustering

- Previously, when supervised learning was covered, it involved datasets with both input and output variables.
- However, we will take a look at another class of problems, unsupervised learning problems, where only the input variables are observed.
- Here we will look at the unsupervised method : clustering.



The graphs above show two datasets that are good candidates for applying clustering. The data on the left shows a clear grouping that a clustering algorithm could readily identify for us. The right hand side data groups with more overlap and will be harder to identify, but still suitable for a clustering approach, rather than using linear regression for example.

K-Means Clustering

K-Means clustering is the most well-known clustering algorithm. It is a *simple* and elegant approach for partitioning a dataset into K distinct clusters. To perform K-Means clustering, we first specify the desired number of clusters, K , and then assign each observation to exactly one of the K clusters.

Feature Space

- There are a number of different distance metrics that are used in algorithms to decide how similar observations are.
- The most common one is the Euclidean distance.

$$(x_i, y_i) \text{ and } (x_j, y_j) \text{ is } \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}.$$

- To compute the mean of a number of observations, we divide the sum of the observations by the number of observations
- To compute the mean of a certain number of (x,y) points, we compute the mean of all x values and the mean of all y values

The K-means Algorithm

The K-means algorithm follows the following steps :

- Select number of clusters , K
- Select random points from the data as starting values and initialise the mean of each cluster.
- For n number of iterations :
 - Assign each point to the cluster whose mean (or “centroid”) is the nearest.
 - Re-compute the means for each cluster based on its current members.
- Repeat steps until convergence.

Validating the clusters

- It's possible to find clusters in any data, but it is important to determine if these clusters actually represent underlying subgroups in the data or are merely grouping with similar noise.
- This is a very hard question to answer. There exist a number of techniques for assigning a significance value to a cluster in order to assess whether there is more evidence for the cluster than one would expect due to chance. However, there has been no consensus on a single best approach. The Silhouette Coefficient (`sklearn.metrics.silhouette_score`) is an example of an evaluation metric which indicates how similar samples within a cluster are, compared to other clusters. A higher Silhouette Coefficient score relates to a model with better-defined clusters

CoGrammar

Q & A SECTION

**Please use this time to ask
any questions relating to the
topic, should you have any.**



CoGrammar

Thank you for joining!