Data Mining and Machine Learning

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Peter Jančovič

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

Data Mining

- Objective of Data Mining is to find structure and patterns in large, abstract data sets
  - Is the data homogeneous or does it consist of several separately i
  - Are there p
  - If so, do these patterns have _____ interpretation?
  - Are there correlations in the data?
  - Is there redundancy in the data?

Data Mining and Machine Learning

UNIVERSITY^{OF}
BIRMINGHAM

Partitioning data into "clusters"

- In this lecture we will start to develop tools to understand the structure of data that can be partitioned into (more or less) distinct subsets

- Can think of Assignment Project Exam Help g from distinct "sources"

  https://eduassistpro.github.io/

- We will consider three dif Add WeChat edu_assist_pro niques:

  - Clustering

  - Multi-modal statistical modelling (Gaussian Mixture Models – GMMs)

  - Decision trees

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Clustering - Objectives

- To explain the motivation for clustering

- To introduce the ideas of distance and distortion

- To describe   isive clustering

- To explain t https://eduassistpro.github.io/ h clustering and decision trees

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITYOF
BIRMINGHAM

What does the data look like?

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM
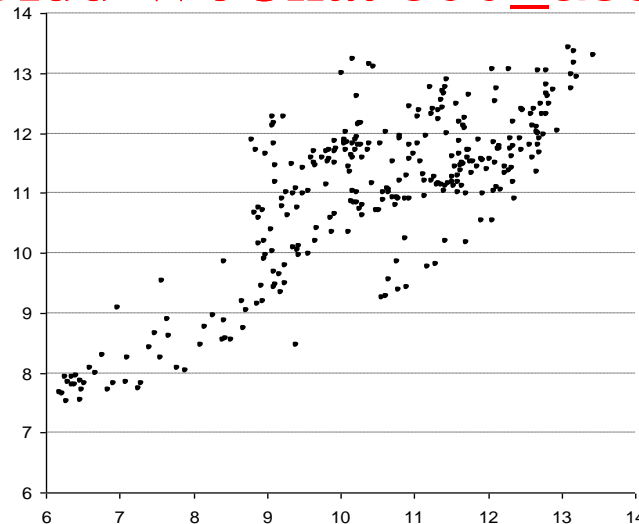
Structure of data

- Typical real data is <u>not</u> uniformly distrubuted

- It has <u>structure</u>

- Variables mi

- The data mi   ural 'clusters' – it
  may have been generated   different
  "sources"

- The purpose of cluster analysis is to find this
  underlying structure <u>automatically</u>

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

Clusters and centroids

- Assume clusters are spherical - determined by <u>centres</u>

- Cluster centres are called <u>centroids</u>

- Questions: How many centroids do we need? Where should we put them?

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Distance

- A function $d(x,y)$ defined on pairs of points x and y is called a <u>distance</u> or <u>metric</u> if it satisfies:
  - $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$
  - $d(x,y) = d(y,x)$ (<u>symmetry</u>)
  - $d(x,z) \leq d(x,y) + d(y,z)$ for all points x, y and z (<u>triangle inequality</u>)

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Example metrics

- The most common metric is the Euclidean metric

- If $x = [x_1, x_2, ..., x_N]$ and $y = [y_1, y_2, ..., y_N]$ then:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_N - y_N)^2}$$

- This is normal distance in space

- There are lots of others, but focus on this one

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

The L^p Metrics

- Euclidean distance is sometimes called the L^2-metric

$$d_2(x, y) = \left[\sum_{n=1}^{N} (x_n - y_n)^2 \right]^{\frac{1}{2}}$$

- It is one of a family of metrics -metrics

$$d_p(x, y) = \left[\sum_{n=1}^{N} (x_n - y_n)^p \right]^{\frac{1}{p}}$$

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

Special L^p metrics

- $p=1$ – the 'City Block' metric

$$d_1(x, y) = \left[\sum_{n=1}^{N} |x_n - y_n| \right]$$

- $p=\infty$

$$d_\infty(x, y) = \max_{n=1,\ldots,N} |x_n - y_n|$$

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

Unit sphere

- For a metric d defined on N dimensional space, the <u>unit sphere</u> is the set of vectors x such that $d(x,0) = 1$

- What do the unit spheres in 2D look like for these metrics?

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

Example Unit Spheres (2D)

L^2 unit sphere, $x^2 + y^2 = 1$

L^1 unit sphere, $|x|+|y|=1$

L^∞ unit sphere, $\max\{x,y\} = 1$

1

-1

-1

1

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM
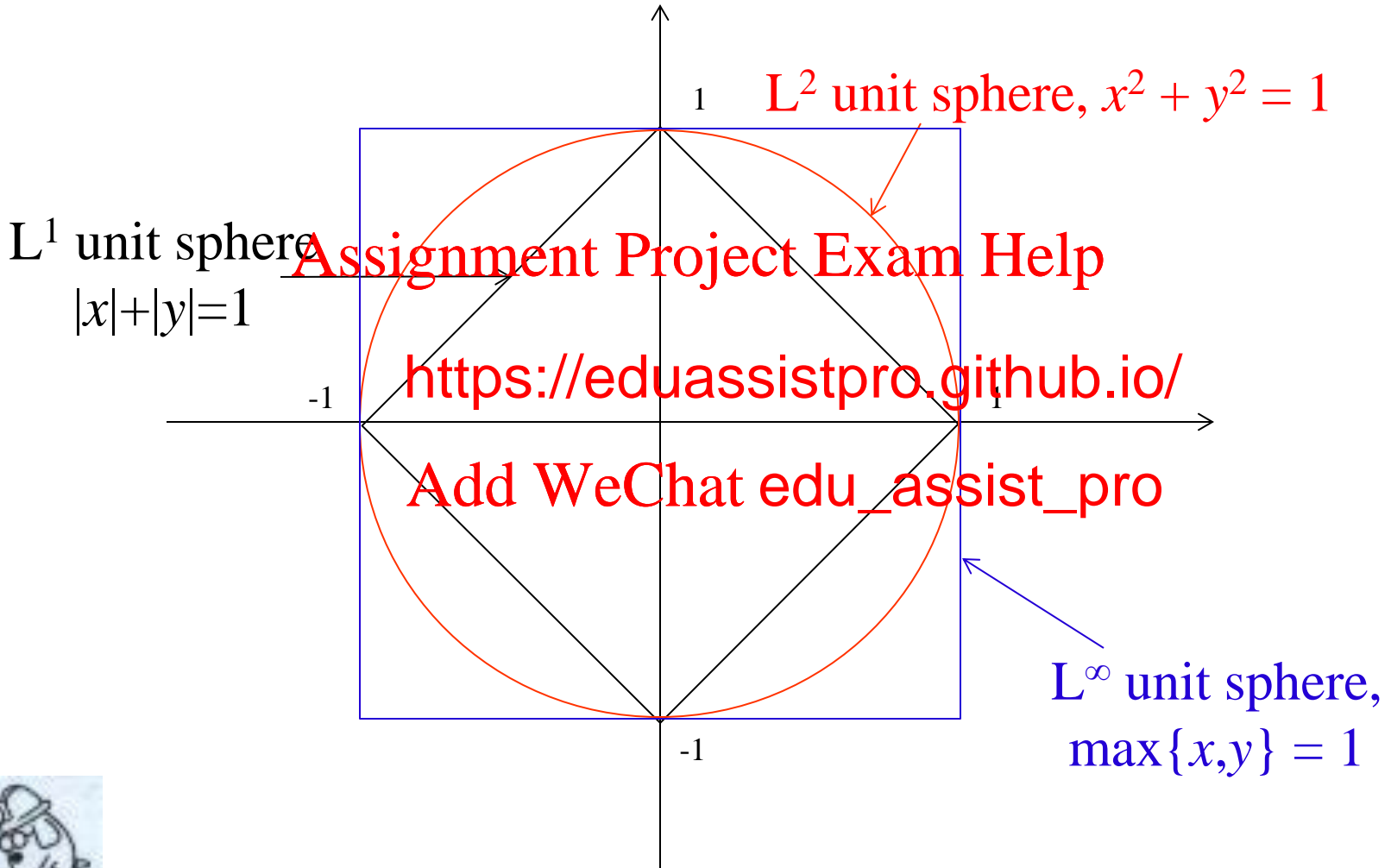
Distortion

- <u>Distortion</u> is a measure of how well a set of centroids models a set of data

- Suppose we have:
  - data point
  - centroids $c_1, \ldots, c_M$

- For each data point y_t let $c_{i(t)}$ be its <u>closest centroid</u>

- In other words: $d(y_t, c_{i(t)}) = \min_m d(y_t, c_m)$

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

Distortion

- The <u>distortion</u> for the centroid set $C = c_1, \ldots, c_M$ is defined by:

$$Dist(C) = \sum_{t=1}^{T} d\left(y_t, c_{i(t)}\right)$$

- In other words, the distortion is the sum of distances

between each data point a st centroid

- The task of clustering is to find a centroid set C such that the distortion $Dist(C)$ is <u>minimised</u>

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Types of Clustering

- We will start with two types of cluster analysis:

  – Agglomerative clustering, or 'bottom-up' hierarchical clustering

  – Divisive clustering

- In the next lecture we will sophisticated clustering m   d k-means clustering
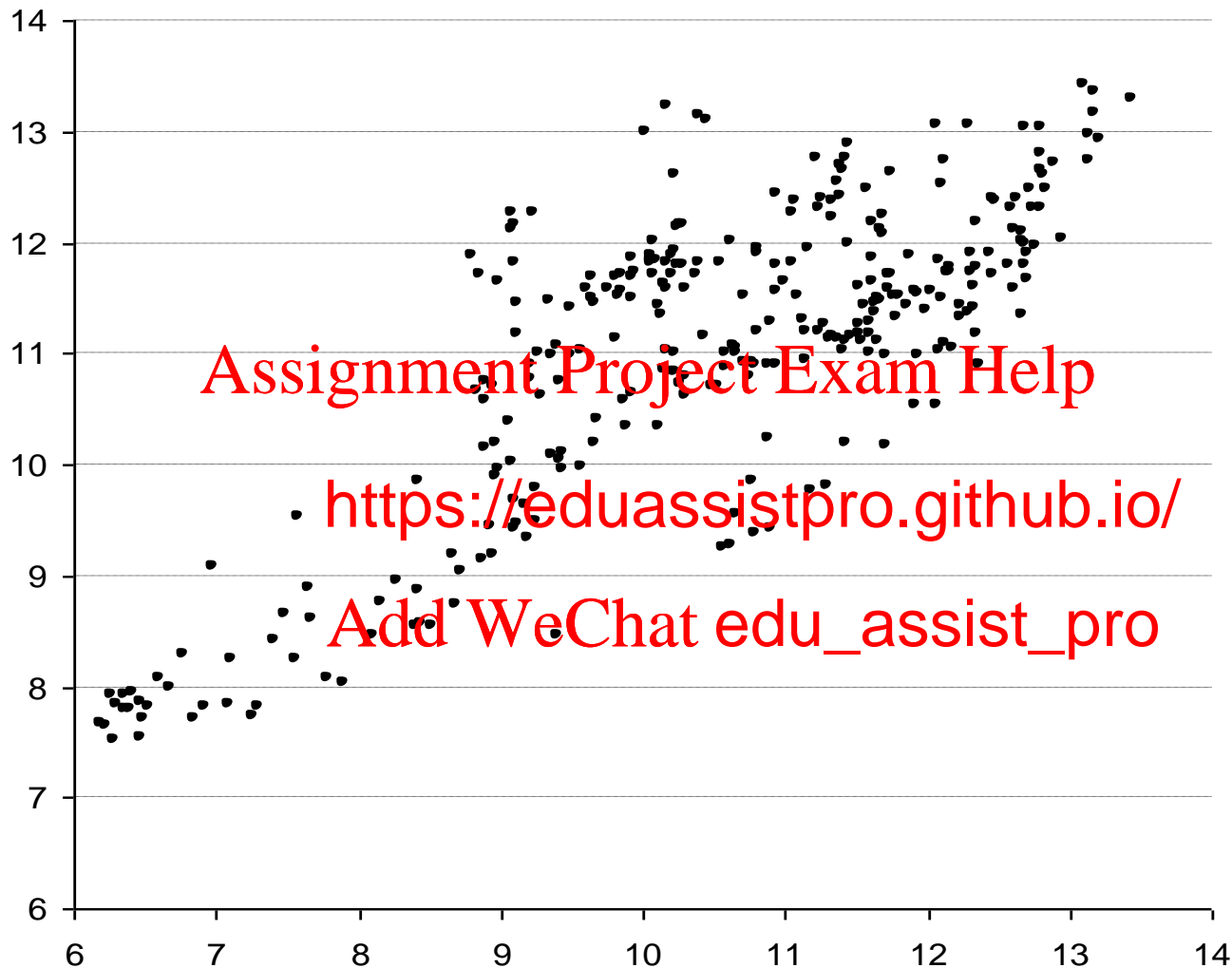
Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM
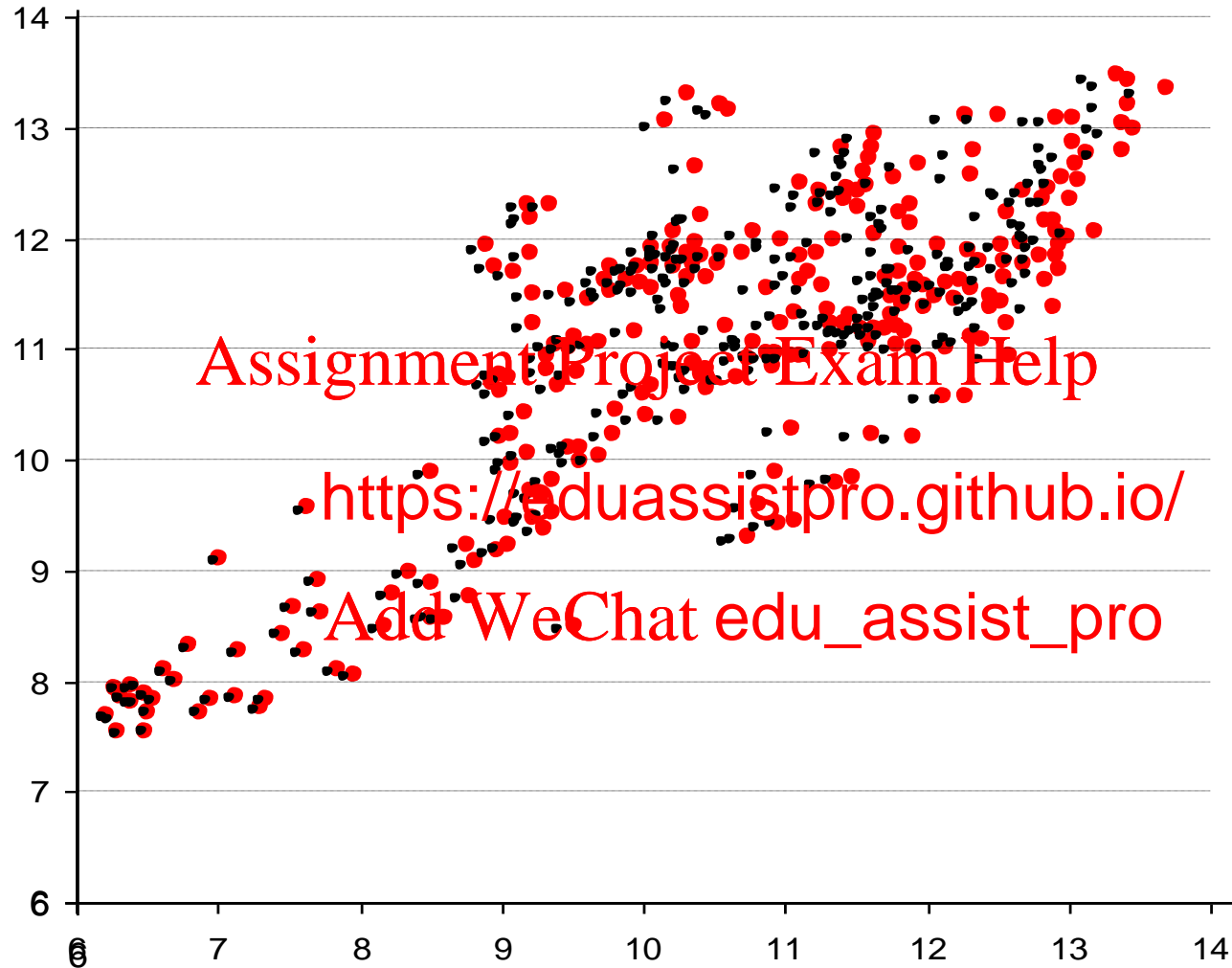
Agglomerative clustering

- Agglomerative clustering begins by assuming that each data point belongs to its own, unique, 1 point cluster – each point is a centroid

- Clusters are required number of c

- The simplest agglomerativ g algorithm is one which, at each stage, combines the two closest centroids into a single centroid

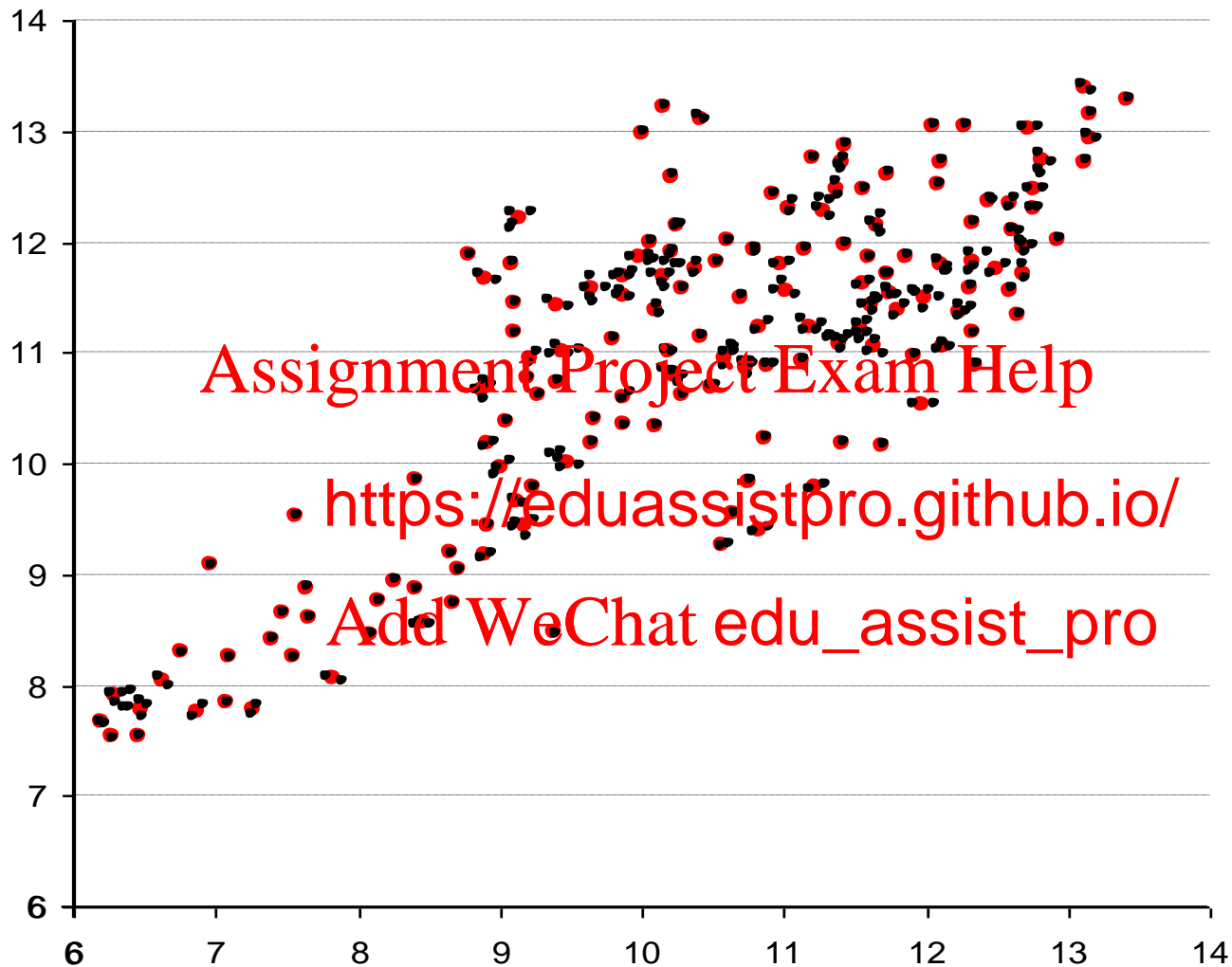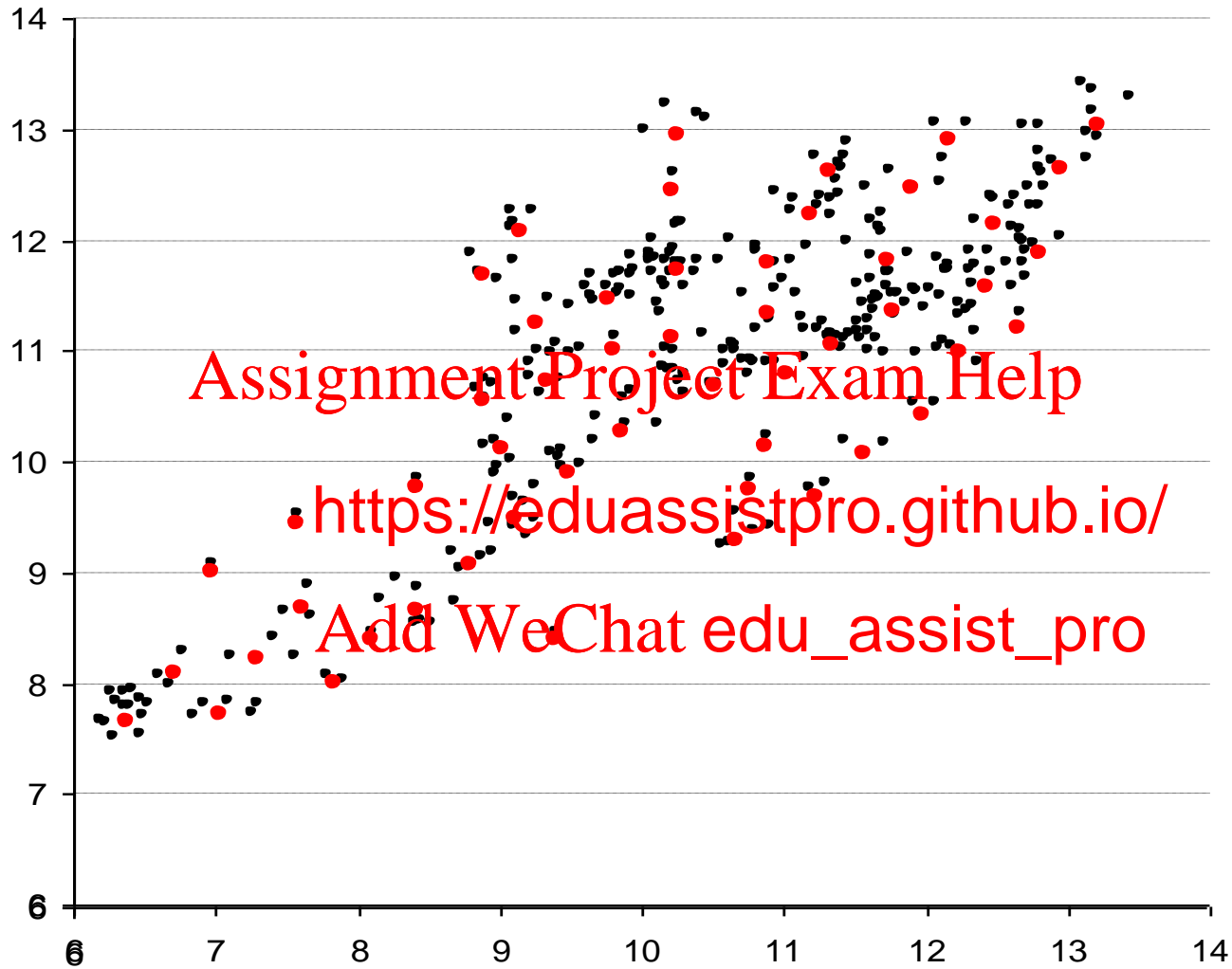Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Original data (302 points)

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

252 centroids

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

152 centroids

Data Mining and Machine Learning

52 centroids

Data Mining and Machine Learning

UNIVERSITY OF
BIRMINGHAM

12 centroids

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM
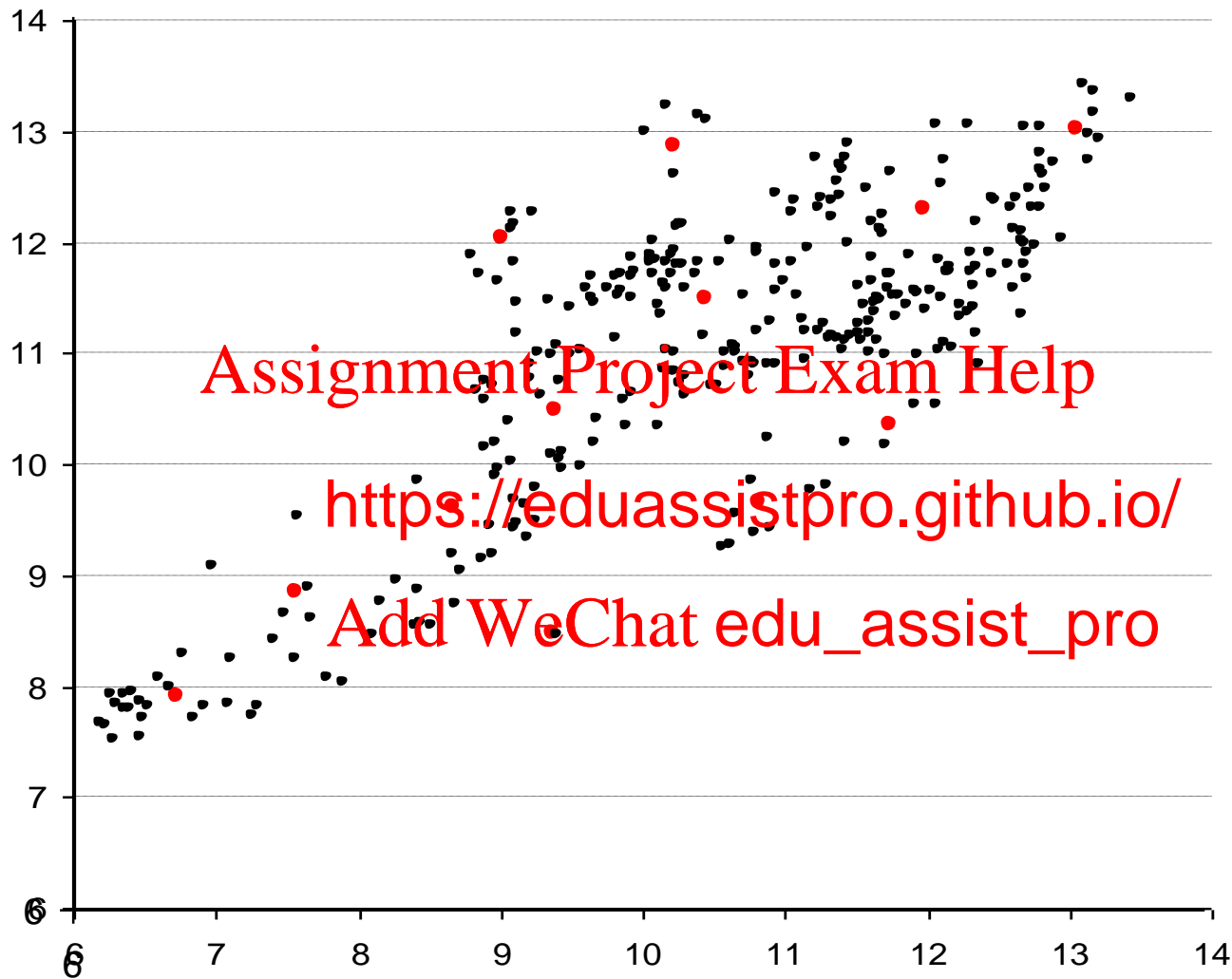
Optimality of agglomerative clustering

- The result of agglomerative clustering is not optimal

- Generally it does not result in a set of centroids C such that

- For example,

    – Outliers may be given their own centroids

    – Dense clusters may be given too few centroids
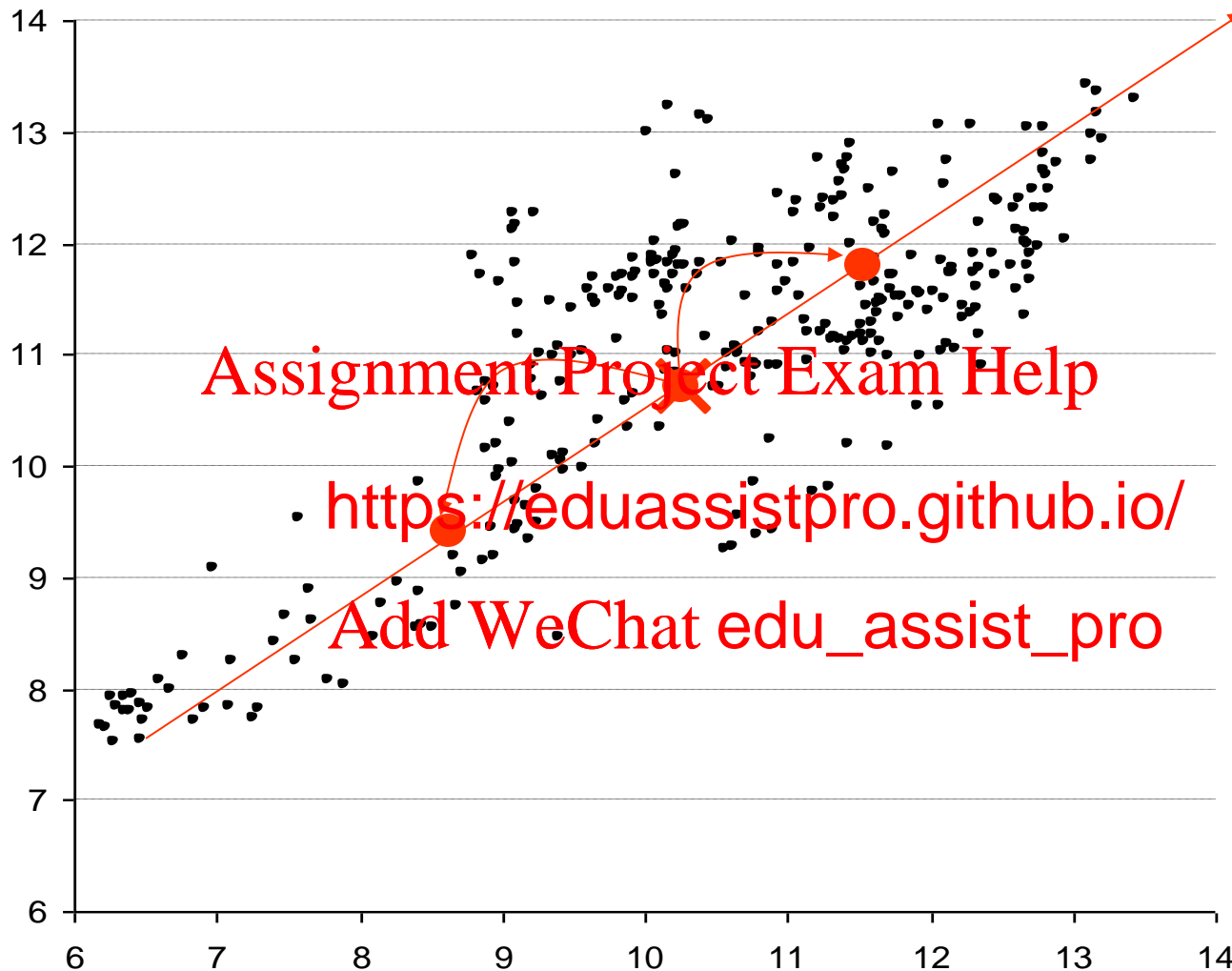
UNIVERSITY OF BIRMINGHAM

Divisive Clustering
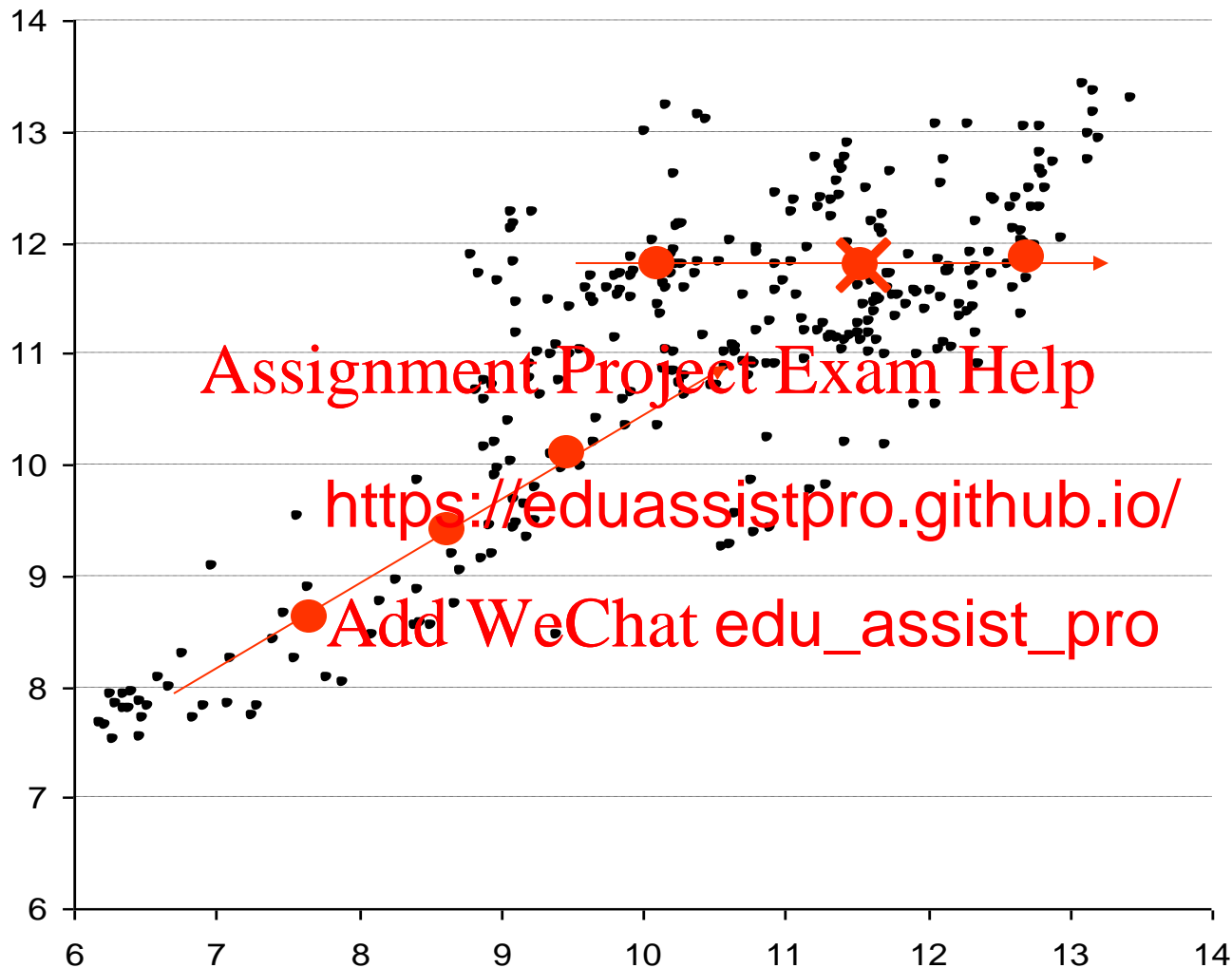
- Divisive clustering begins by assuming that there is just one centroid – typically in the centre of the set of data points

- That point is Assignment Project Exam Help centroids

- Then each of these is repla https://eduassistpro.github.io/ new centroids

- … Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Original data (302 points)

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Original data (302 points)

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Optimality of divisive clustering

- The result of agglomerative clustering is not optimal

- Generally it does not result in a set of centroids C such that

- Sequential decision makin lly suboptimal

  - Decisions are not reversible

  - If a point goes to a particular half of a partition it will never be re-allocated to the other half

  - Probably not how a human would do it

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Decision tree interpretation

Single centroid - whole set

Top down clustering - divisive

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Multiple centroids – one per data point

Bottom up clustering - agglomerative

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Optimality

- An 'optimal' set of centroids is one which minimises the distortion

- In general, neither method gives optimal sets of centroids

- A more principled approac   e to think of distortion as a function of   d set and minimize it

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Notation and method

- *N* dimensional space

- *T* data points $X = \{x_1,...,x_T\}$

- *K* centroids

- Calculate

$$\frac{d}{dc_k^n} Dist(C)$$

for each *k* and *n*, set to zero and solve

Data Mining and Machine Learning

UNIVERSITY OF BIRMINGHAM

Summary

- Distance metrics and distortion

- Agglomerative clustering

- Divisive clu

- Decision tre

Assignment Project Exam Help

https://eduassistpro.github.io/

Add WeChat edu_assist_pro

Data Mining and Machine Learning

UNIVERSITYOF
BIRMINGHAM