

**Machine Learning– CA2**

**B9DA104**

**Master of Science (MSc) in Data Analytics**

**Tim Browning**

**10543928**

**Dublin Business School**

Contents

[Q1) Define the following? 4](#_Toc56590497)

[Define Data sampling and its steps with examples? 4](#_Toc56590498)

[Define Decision tree, Information gain and Entropy? 5](#_Toc56590499)

[Define Chinese restaurant algorithm and Agglomerative Hierarchical Clustering with example? 6](#_Toc56590500)

[Q2 Program the following 10](#_Toc56590501)

# Q1) Define the following?

## Define Data sampling and its steps with examples?

It is a statistical technique used to analyse a subset of a dataset to discover patterns and trends in the larger data set. It allows data analysts to work with a more manageable amount of data to build and run analytical models more efficiently while still producing accurate results. They are particularly useful with data sets that are too large and analysing a sample is more cost effective then surveying the entire dataset.

(“What is data sampling?,” n.d.)

**Methods:** Probability sampling and non-probabilistic sampling

**Probability Sampling Methods:** Simple random sampling, Stratified sampling, Cluster sampling and Systematic sampling.

**Non-Probability Sampling Methods:** Convenience sampling, Quota sampling, Judgement (or Purposive) Sampling, Snowball sampling and Bias in sampling

Simple Random Sampling

Each member of the population has the same probability of being chosen

The code below generates 1000 sample points

Sample\_df = df.sample(1000)

Stratified Sampling

Say there is an election and we need to estimate the average number of votes for each candidate. There are our three towns:

Town 1: population = 100,000

Town 2: population = 300,000

Town 3: population = 400,000

We are looking for a sample size of 800

Instead of taking a sample size for the entire population, we can choose to take a sample of 100, 300 and 400, respectively. This will produce a smaller error in estimation because it removes the bias.

**Steps:**

1. Selecting the sampling technique to be executed
2. Define parameters
3. Execute a report

## Define Decision tree, Information gain and Entropy?

**Decision Tree**: partitions the data into smaller subsets where each subset contains responses of one class. Decision Tree algorithms are also known as Classification and Regression Trees(CART). Cart predictors can be either categorical or numerical. It is a supervised learning method used to solve regression and classification problems. The decision nodes and leaf nodes contain the results to the problem being solved.

Decision Trees are built up recursively by increasing the information (reducing the uncertainty) contained in the reduced data set following each split

A decision tree is used to visually represent decisions and decision making. (Gupta, 2017)

**Information Gain:**  It is often used in decision trees. It decides the best and most significant predictor that should be used at each decision tree step. Information gain can also be used for feature selection. (Brownlee, 2019)

**Entropy:**is the measures of impurity or uncertainty in the information being processed. Entropy controls how a Decision Tree splits the data. It affects how a Decision Tree draws its boundaries.

## Define Chinese restaurant algorithm and Agglomerative Hierarchical Clustering with example?

**Chinese restaurant algorithm:**  is an unsupervised learning method used for clustering of data. It is a random process, the process is like a Chinese restaurant that has a finite number of tables, and a customer chooses to sit at any of the tables(occupied or empty). The algorithm will calculate the probability to select an empty or occupied table based on the total of customers already at each table. (“processing by chinese manu,” n.d.).

Example

Customers 1, 3, 4, and 7 are sat at empty tables, and customers 2, 5, 6, and 8 are sat at existing tables.

The number of tables in the restaurant is infinite.

Customer 1 enters and he can sit anywhere he likes. Customer 2 can sit in any empty seat, with the following probabilities:

* Table 1: 1 / (1 + α)
* Any empty table/New table: α / (1 + α)

The probability where the 9th customer will sit is:

Table 1: 3 / (8 + α)

Table 2: 1 / (8 + α)

Table 3: 3 / (8 + α)

Table 4: 1 / (8 + α)s

New Table: α / (8 + α)

The number of people that have already sat at a particular table is the numerator and the number of customers in the restaurant (i – 1) plus α is the denominator

The probability of the ith customer sitting at an existing table is nk / (α + i – 1),

The probability of the ith customer sitting at a new table is α / (α + i – 1).

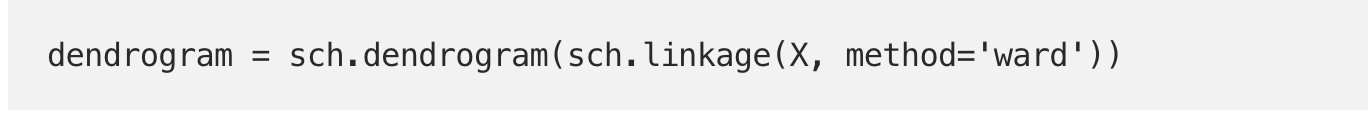
As more people sit at a table, those tables increase in popularity so new customers are less likely to sit at empty tables(“Chinese Restaurant Process: Simple Definition & Example - Statistics How To,” n.d.)

**Agglomerative Hierarchical Clustering** is used to group objects in clusters based on their similarity. The algorithm treats each object as a singleton cluster. The pairs of clusters are then merged until all clusters have merged into one big cluster. This results in a tree based representation of the objects called a dendrogram (“Agglomerative Hierarchical Clustering,” n.d.)

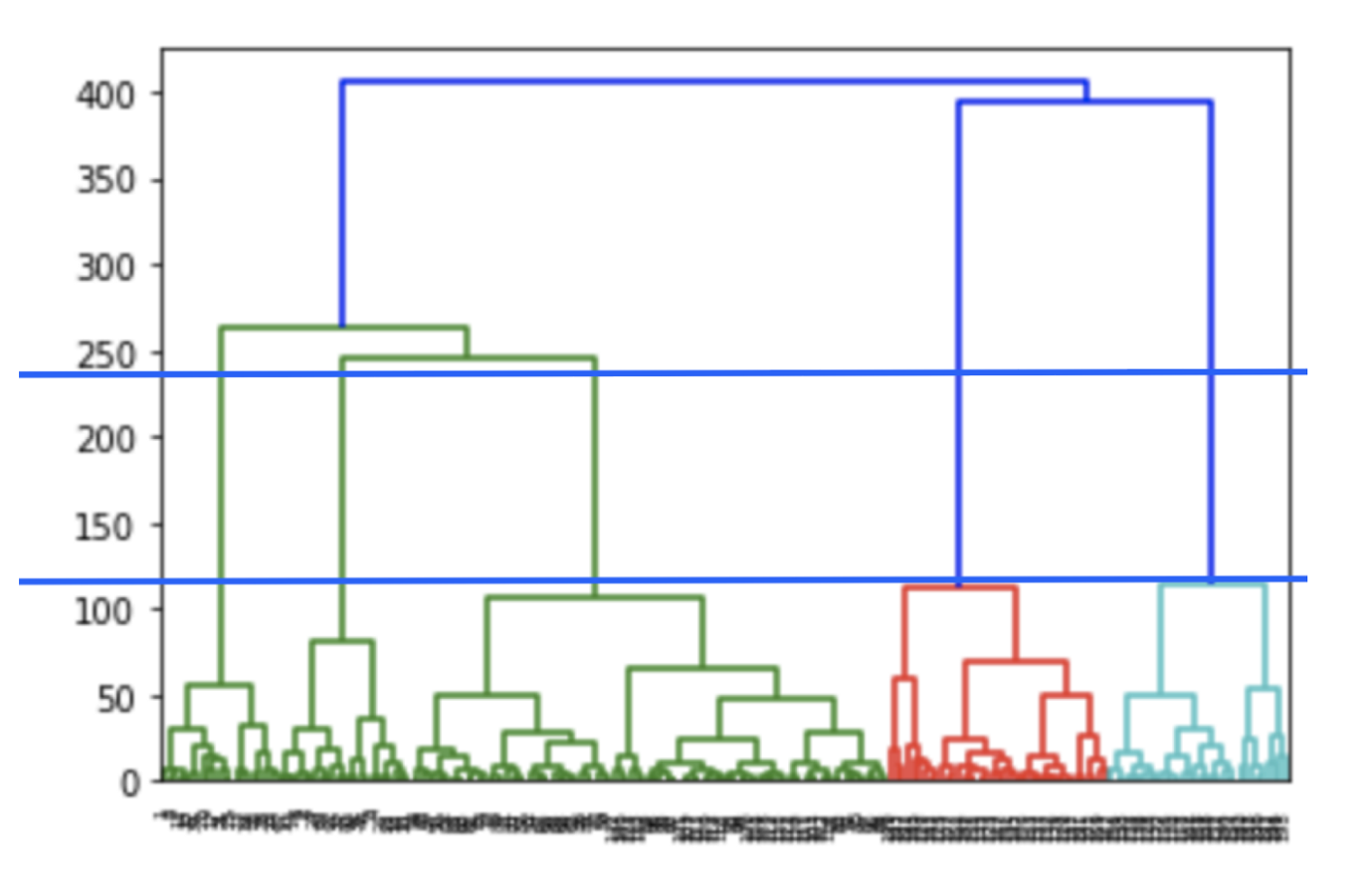
Example

In this example, the dataset contains a list of customers with their gender, age, annual income and spending score

The code below will display a dendrogram graph



In the dendrogram, The middle green section at its highest point does not intersect with any clusters. Five of the vertical lines do not cross the threshold. Therefore the optimal number of clusters is 5.



To calculate the proximity of clusters, we can create a Agglomerative Clustering instance using the Euclidean distance as the measure between the points and ward linkage.

.

Graphical user interface, application

Description automatically generated

labels\_ returns an array of integers. These values correspond to the distinct categories

A picture containing text

Description automatically generated

Using a scatterplot we can display all the samples belonging to a category as a specific colour (“Agglomerative Hierarchical Clustering,” n.d.)

Chart, scatter chart

Description automatically generated

# Q2 Program the following

Link - https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/

Data - https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

HTM file & Jupyter notebook attached

https://htmtopdf.herokuapp.com/ipynbviewer/temp/42061c4337387f07b6eae38faf1cbdc4/CA2\_Final%20(7).html?t=1605697953293

Agglomerative Hierarchical Clustering, n.d. . Datanovia. URL https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/ (accessed 11.12.20).

Brownlee, J., 2019. Information Gain and Mutual Information for Machine Learning. Mach. Learn. Mastery. URL https://machinelearningmastery.com/information-gain-and-mutual-information/ (accessed 11.7.20).

Chinese Restaurant Process: Simple Definition & Example - Statistics How To [WWW Document], n.d. URL https://www.statisticshowto.com/chinese-restaurant-process/ (accessed 11.12.20).

Gupta, P., 2017. Decision Trees in Machine Learning [WWW Document]. Medium. URL https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 (accessed 11.7.20).

processing by chinese manu [WWW Document], n.d. URL https://www.tcgtenis.pl/Pink/ovj0i7/processing-by-chinese-manu.html (accessed 11.7.20).

What is data sampling? - Definition from WhatIs.com [WWW Document], n.d. . SearchBusinessAnalytics. URL https://searchbusinessanalytics.techtarget.com/definition/data-sampling (accessed 11.3.20).