Do we Know what is Collab?

Do we Know what is Collab?

Colab (colab.research.google.com):
Google Colab provides a ready-to-use python environment, with most of the common packages for data science pre-installed. It also provides a free GPU useful for doing deep learning. However, at this stage, you will not need the GPU option and should refrain from using it. There is a hard limit of 12 hours on using Colab. After this time expires, your notebooks will be reset and you will have to run your code again. Also, note that colab notebooks might automatically disconnect if you do not do any activity for some time.

Guide: https://www.kdnuggets.com/2020/06/google-colab-deep-learning.html
Tips and tricks: https://www.google-colab.com/google-colab-tips-and-tricks/

Fundamentals of Big Data Analytics

Lecture 3- Exploratory Data Analysis

Dr. Iqra Safder
Assistant Professor
FAST NUCES, Lahore

Data Object and Attribute

DATA OBJECT AND ATTRIBUTE

Data object

- represents an entity in the data set
- also called data item, point, instance, example, sample, row, observation
- e.g. a patient, movie, student, customer, product, book, tweet
- described by a set of attributes

Attribute

- is a data field, representing a feature/characteristic of data objects
- also called variable, feature, dimension, column, coordinate, field
- e.g. reaction to a test, genre/director, course, address, price/category, author, publisher, word

Size and dimensions of data

Size of Data refers to number of data objects

Dimension of Data refers to number of attributes

Sparsity in Data

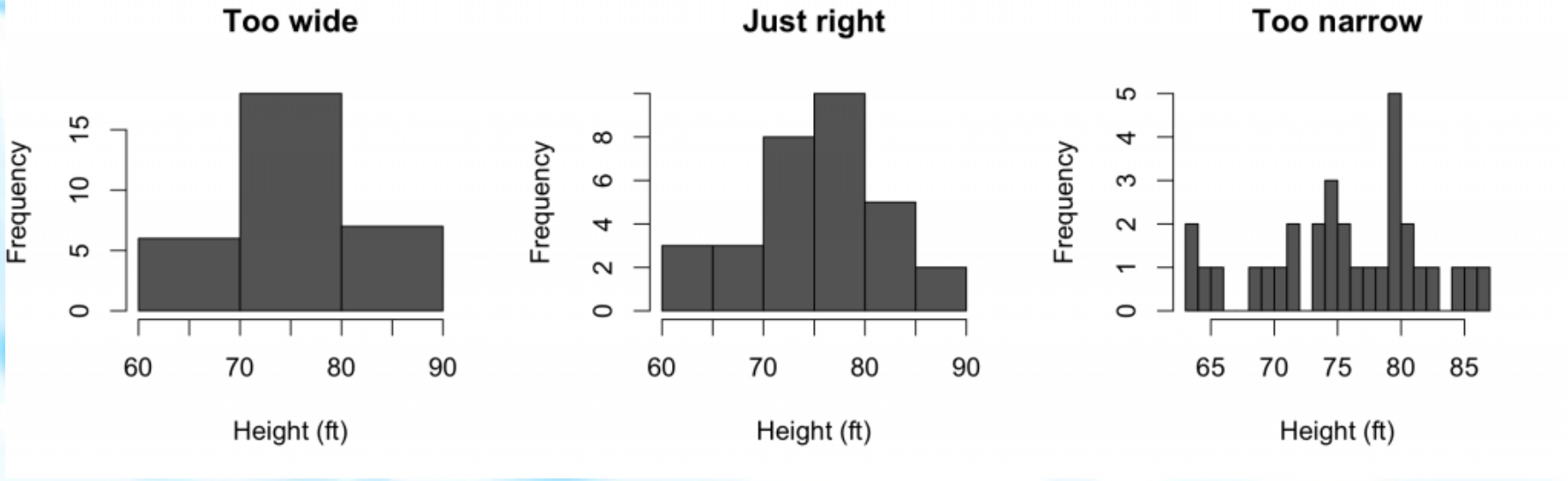
If most of the feature values are missing, then the data is called sparse

- Missing values could be represented as NaN, blank, -, 0
- This could be a problem for many statistical methods
- For efficient computation, can use libraries for sparse data
  - e.g. sparse matrix multiplication, sparse storage schemes, scipy.sparse

Resolution of Data

Different resolution reveal Different patterns

- If resolution is too fine, a pattern may be buried in noise
- If the resolution is too coarse pattern may disappear
- See number of bins in histograms below



Types of Data

Types of data based on number of attributes

- Univariate Data
- Bivariate Data
- Multivariate Data

Types of Data

- Univariate: Consists of only one feature. Analysis deals with only one quantity that changes

Heights (cm)
164
167.3
170
174.2
178
180
186

  - What is average height?
  - How much the values deviate form the average height?

Types of Data

- Bivariate: Involves two different features

  Analysis of this type of data deals with comparisons, relationships, causes and explanations

Temperature (°C)	Ice Cream Sales
20	2000
25	2500
35	5000
43	7800

  - Are the temperature and ice cream sales related/dependent?
  - As temperature increases, sales also increases

Types of Data

- Multivariate: Objects are described by more than 2 features

  To see if one or more of them are predictive of a certain outcome

  The predictive variables are independent variables and the outcome is the dependent variable

Roll Num	CS100	SS101	MT200	MGMT240	Major
19100115	A	B	B	C	CS
19100120	B	A	B	C	PHY
19100122	B	B	C	A	CS
19100126	C	A	C	A	EE
19100127	B	A	C	C	CS
19100133	C	B	A	B	PHY
19100135	C	C	A	C	Maths

Types of Attributes

Types of Attributes

Roll Num	Gender	Grade	Age	Major
19100115	Male	B	23	CS
19100120	Male	A	22	PHY
19100122	Female	B	21	CS
19100126	Male	C	19	EE
19100127	Female	A	21	CS
19100133	Female	B	20	PHY
19100135	Male	C	22	Maths

- Nominal/Categorical Attributes
- Ordinal Attributes
- Numeric Attributes

Types of Attributes: Nominal/Categorical

- Possible values are symbols, labels or names of things, categories
  - gender, major, state, color
- Describe a feature qualitatively and values have no order
- Not quantitative, arithmetic operations can't be performed on them
  - male — female = ?? green + blue = ??
- Can code by numbers (numeric symbols) e.g. postal codes, roll numb

Can compute

- frequency of values and the most frequent value
- ~~middle value~~
- ~~average value of an attribute~~

Binary Attribute: - special case of nominal TRUE/FALSE, Pass/Fail, 0/1

- Symmetric: Both symbols carry the same weight e.g. gender
- Asymmetric: Both symbols are not equally important, e.g. Pass/Fail

Types of Attributes: Ordinal Attributes

- Possible values have meaningful order, Type of categorical with an order
  - Grades : A,B,C,D
  - Serving Sizes : Small, Medium, Large
  - Ratings : poor, average, excellent
- No quantified difference between two levels
  - A is higher/better than B but
  - Cannot quantify how much higher is A than B, or
  - if the difference between A and B the same as the difference between B and C
- Can be obtained by discretizing numeric quantities (data reduction)

Can compute

- frequency of values and the most frequent value
- middle value
- ~~average value of an attribute~~

Types of Attributes: Numeric Attributes

- Quantitative and measurable
- can quantify the difference between two values
  - temperature, age, number of courses, height, years of experience

Can compute

- frequency of values and the most frequent value
- middle value
- average value of an attribute

Useful Resources for datasets:

- https://archive.ics.uci.edu/ml/datasets.php
- https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Find me @

Dr. Iqra Safder
Assistant Professor
iqra.safder@nu.edu.pk
Office: Ground floor, Civil blook,
NUCES, Lahore

National University
Of Computer and Emerging Sciences