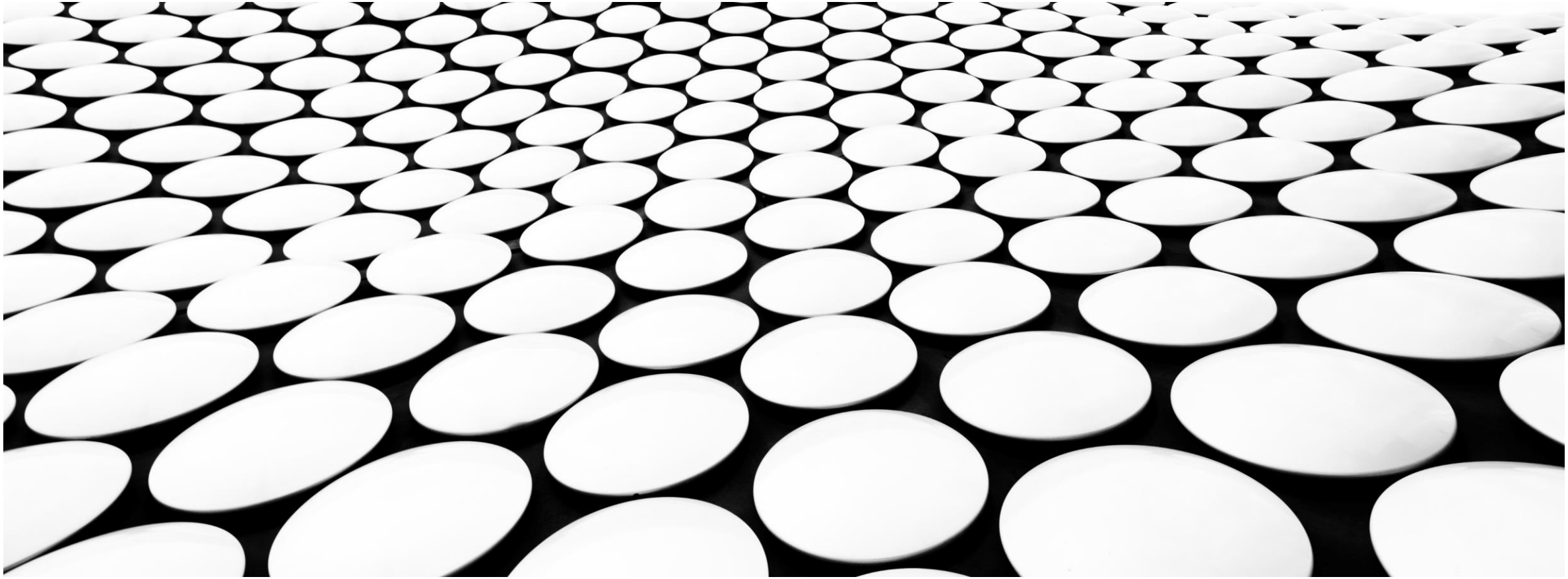
K-NEAREST NEIGHBORS ALGORITHM



1. KYEYUNE TADEO
2. KAYONDO UMAR

INTRODUCTORY CONCEPTS

- Regression
- Classification
- Parameters, Parametric and Non parametric classification
- Normalizing
- Over fitting/Under fitting

REGRESSION ANALYSIS

- In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables.
  - In an experiment, values of dependent variables are studied under the supposition or hypothesis that they depend on the values of other variables.
  - Independent variables, in turn, are not seen as depending on any other variable in the scope of the experiment in question. In this sense, some common independent variables are time, space, density, mass, fluid flow rate, and previous values of some observed value of interest (e.g. human population size) to predict future values (the dependent variable).

REGRESSION ANALYSIS CONTINUED…

- It is always the dependent variable whose variation is being studied, by altering inputs, also known as regressors in a statistical context. In mathematical modeling, the dependent variable is studied to see if and how much it varies as the independent variables vary.
- In the simple linear model:

$$y_i = \mathrm{a} + \mathrm{b}x_i + e_i$$

- the term yi is the i-th value of the dependent variable and xi is the i-th value of the independent variable. The term ei is known as the "error" and contains the variability of the dependent variable not explained by the independent variable. With multiple independent variables, the model is

$$y_i = \mathrm{a} + \mathrm{b}x_{i,1} + \mathrm{b}x_{i,2} + \ldots + \mathrm{b}x_{i,n} + e_{\mathrm{i}}$$

where n is the number of independent variables.

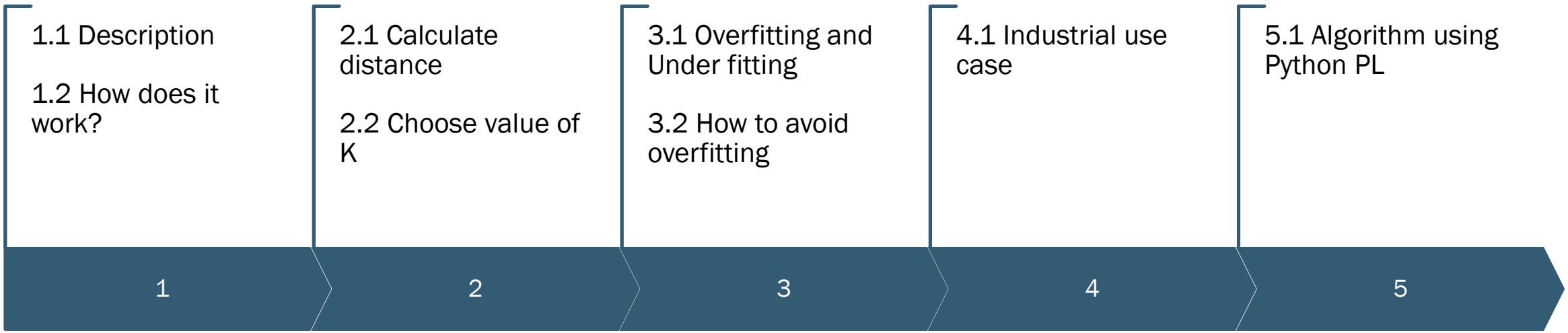
CLASSIFICATION

- Classification is the process of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.).
  Classification is an example of pattern recognition.
- Classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance (todays topic of discussion).

PARAMETERS, PARAMETRIC AND NON PARAMETRIC STATISTICS

- A parameter is any measured quantity of a statistical population that summarizes or describes an aspect of the population, such as a mean or a standard deviation. If a population exactly follows a known and defined distribution, for example the normal distribution, then a small set of parameters can be measured which completely describes the population, and can be considered to define a probability distribution for the purposes of extracting samples from this population.
- Nonparametric statistics is the branch of statistics that is not based solely on parametrized families of probability distributions (common examples of parameters are the mean and variance). Nonparametric statistics is based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified.
    - Includes techniques that do not rely on data belonging to any particular parametric family of probability distributions.
    - Includes techniques that do not assume that the *structure* of a model is fixed. Typically, the model grows in size to accommodate the complexity of the data. In these techniques, individual variables *are* typically assumed to belong to parametric distributions, and assumptions about the types of connections among variables are also made.

K-NEAREST NEIGHBORS ALGORITHM



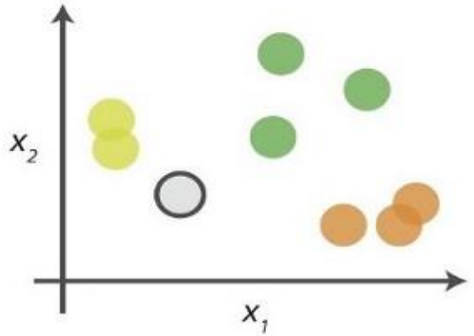
K-NEAREST NEIGHBORS ALGORITHM

- k-nearest neighbors algorithm (k-NN or KNN) is a *non-parametric* classification method used for `classification` and `regression`.

  The input consists of the k closest training examples in data set.

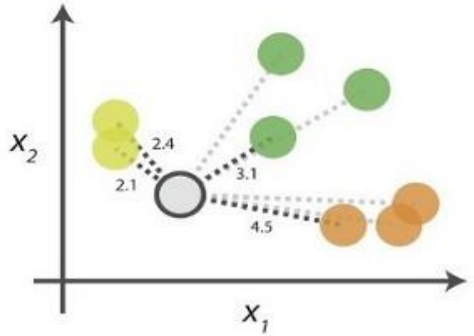
  - In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
  - In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.
  - Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy
  - a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

0. Look at the data



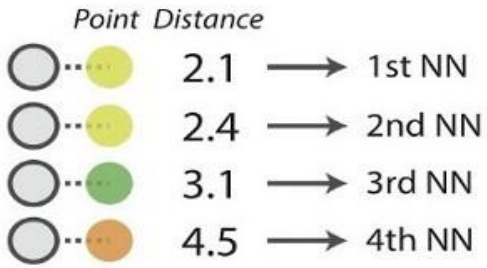
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances



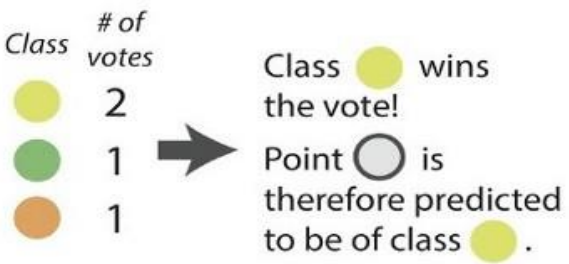
Start by calculating the distances between the grey point and all other points.

2. Find neighbours



Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels



Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

DISTANCE BETWEEN POINTS - EUCLIDEAN DISTANCE

- The distance between any two points on the real line is the absolute value of the numerical difference of their coordinates. Thus if p and q are two points on the real line, then the distance between them is given by:

$$d(p, q) = |p - q|.$$

- In the Euclidean plane, let point p have Cartesian coordinates (p 1 , p 2) and let point q have coordinates (q 1 , q 2). Then the distance between p and q is given by:

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

EUCLIDEAN DISTANCE C'TD

- A more complicated formula, giving the same value, but generalizing more readily to higher dimensions, is:

$$d(p, q) = \sqrt{(p - q)^2}.$$

In this formula, squaring and then taking the square root leaves any positive number unchanged, but replaces any negative number by its absolute value.

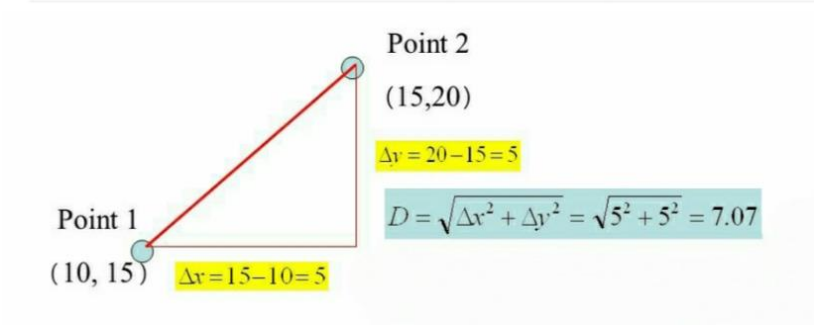
EUCLIDEAN DISTANCE C'TD …

- In general, for points given by Cartesian coordinates in n-dimensional Euclidean space, the distance is:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

This is summarized as:

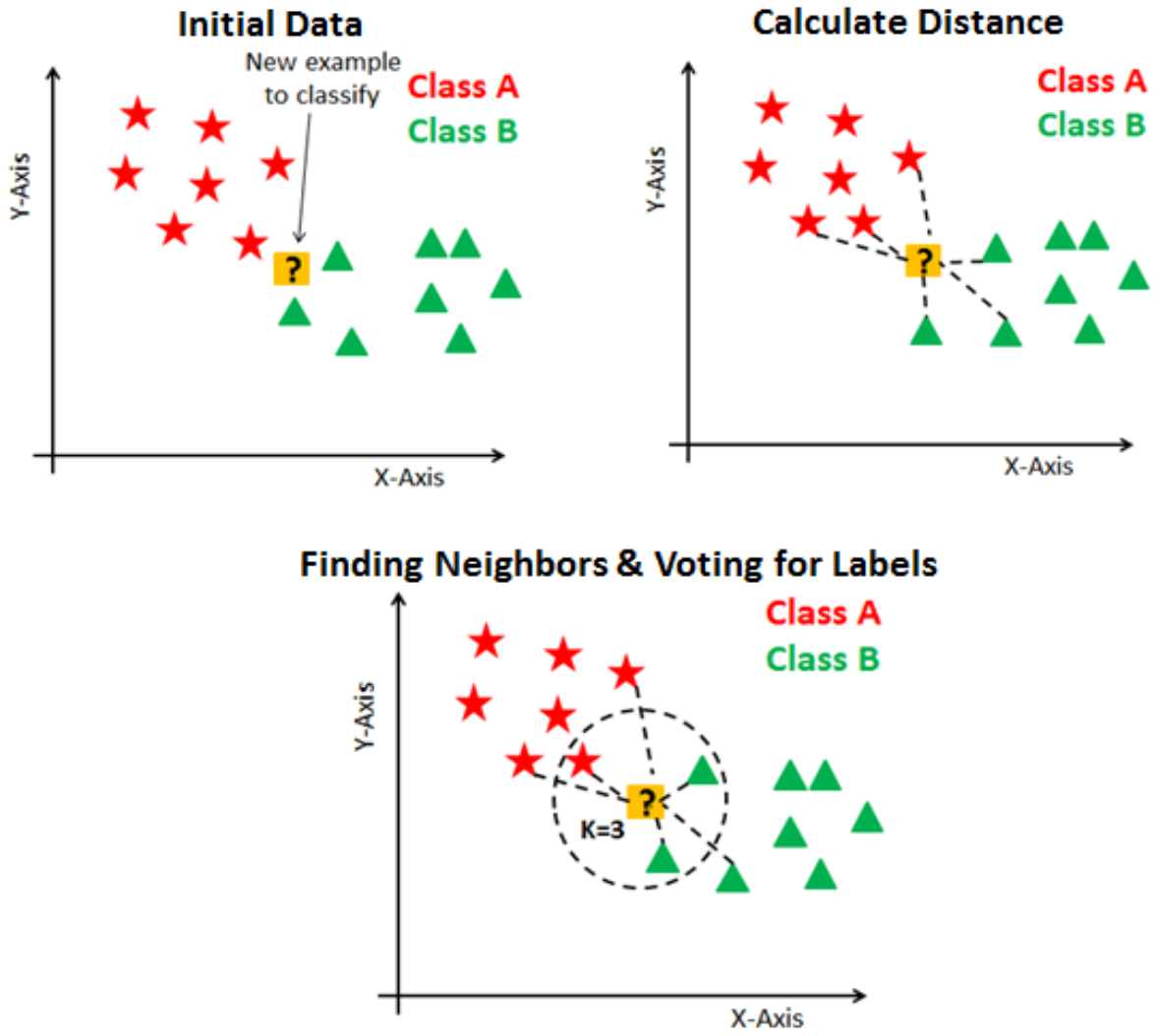
$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$



KNN - CLASSIFICATION

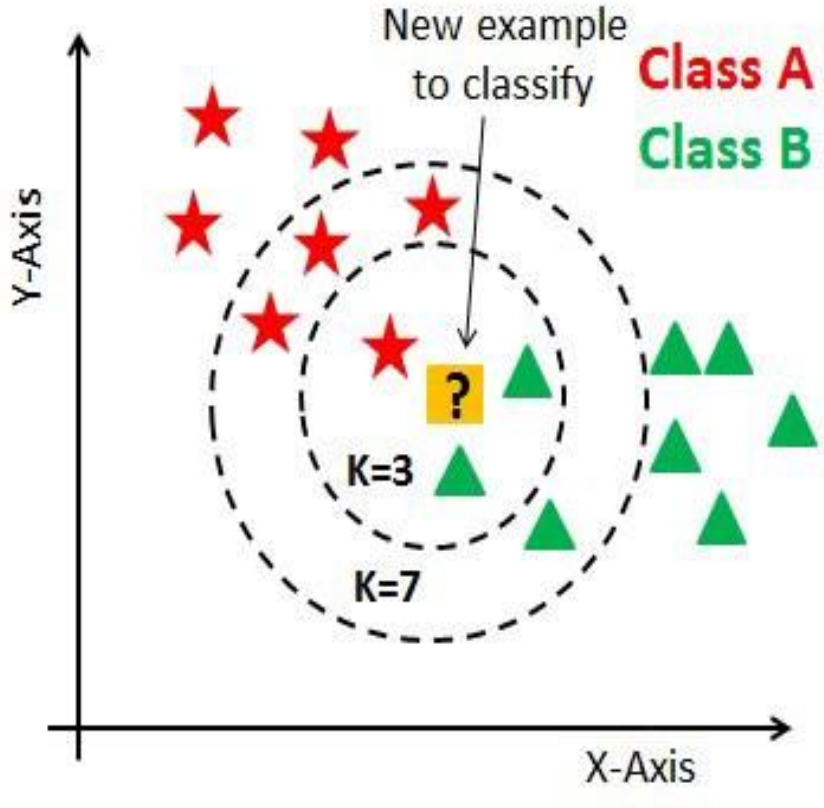
- To classify data, let's follow the procedure below;
  - Initialize the K value.
  - Calculate the distance between test input and K trained nearest neighbors.
  - Check class categories of nearest neighbors and determine the type in which test input falls.
  - Classification will be done by taking the majority of votes.
  - Return the class category.

KNN – CLASSIFICATION C'TD …



CHOOSING THE VALUE OF K

- **K** value indicates the count of the nearest neighbors. We have to compute distances between test points and trained labels points. Updating distance metrics with every iteration is computationally expensive, and that's why KNN is a lazy learning algorithm



- As you can verify from the above image, if we proceed with K=3, then we predict that test input belongs to class B, and if we continue with K=7, then we predict that test input belongs to class A.
- That's how you can imagine that the K value has a powerful effect on KNN performance.

CHOOSING THE VALUE OF K C'TD ...

- If you randomly select the K value and get into this situation, this means that your K value is not optimized so you need to optimize it based on your dataset. How do you choose the optimal value of K?
  - There are no pre-defined statistical methods to find the most favorable value of K.
  - Initialize a random K value and start computing.
  - Choosing a small value of K leads to unstable decision boundaries.
  - The substantial K value is better for classification as it leads to smoothening the decision boundaries.
  - Derive a plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate.

KNN OVERFITTING AND UNDERFITTING

- The value of k in the KNN algorithm is related to the error rate of the model.
- A small value of k could lead to overfitting as well as a big value of k can lead to underfitting.
- Overfitting imply that the model is well on the training data but has poor performance when new data is coming.
- Underfitting refers to a model that is not good on the training data and also cannot be generalized to predict new data.
- To avoid overfitting,
  - To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set X_test, y_test.

QUESTION

- A survey was conducted to find out sports activities at given age brackets in both male and female. Assume the data set on the next slide was obtained after a survey.

- Given a new female person named Brenda aged 18 years. Predict which kind of sports activity Brenda is most likely to be in.

Name	Age	Gender	Sport
Conrad	61	M	Football
Deo	75	M	Football
Solomon	63	M	Tennis
Miriam	59	F	Football
Freedom	48	M	Netball
Jerry	53	M	Basket ball
Allan	81	M	Tennis
Brian	92	M	Golf
Bernard	57	M	Golf
George	60	M	Golf
Eric	77	M	Tennis
Faith	82	F	Tennis
Gideon	93	M	Football
Mark	76	M	Football
Nicholas	60	M	Football
Paul	51	M	Cricket
Kelmo	82	M	Cricket
Umar	73	M	Golf
Brenda	18	F	?

SOLUTION

- **Assumptions**

  Let K=3

- We are going to convert string data sets in the gender column to numeric;

  Male = 0

  Female =1

- Now KNN is going to find the distance between each data set in the reference data and the new data set (Brenda).
- Using Euclidean distance; Brenda | Age = 18, Gender = 1, Data point = (18, 1);

SOLUTION C'TD – CALCULATING EUCLIDEAN DISTANCE

- Let's use Conrad for illustrating;

$$= \sqrt{(x1 - x2)2 + (y1 - y2)2}$$

$$= \sqrt{(18 - 61)2 + (1 - 0)2}$$

$$= \sqrt{(-43)2 + (1)2}$$

$$= \sqrt{1{,}849 + 1}$$

$$= \sqrt{1{,}850}$$

$$= 43.0116$$

Name	Age	Gender	Euclidean Distance	Sport
Conrad	61	0	43.0116	Football
Deo	75	0	57.0087	Football
Solomon	63	0	45.0111	Tennis
Miriam	59	1	41	Football
Freedom	48	0	30.0167	Cricket
Jerry	53	0	35.0143	Basket ball
Allan	81	0	63.0079	Tennis
Brian	92	0	74.0068	Golf
Bernard	57	0	39.0128	Golf
George	60	0	42.0119	Golf
Eric	77	0	59.0085	Tennis
Faith	82	1	64	Tennis
Gideon	93	0	75.0067	Football
Mark	76	0	58.0086	Football
Nicholas	60	0	42.0119	Football
Paul	51	0	33.0151	Cricket
Kelmo	82	0	64.0078	Cricket
Umar	73	0	55.0091	Golf
Brenda	18	1		?

SOLUTION C'TD

- Let's use Conrad for illustrating;
  - After calculating the Euclidean distance of all the data sets, we get the K nearest neighbors. For our example K=3, we get the 3 more nearest data sets

Freedom	48	0	30.0167	Cricket
Paul	51	0	33.0151	Cricket
Jerry	53	0	35.0143	Basket ball

  - Since cricket is the most common sport (Voting), then we can predict that Brenda is most likely to like Cricket.

KNN IMPLEMENTATION USING PYTHON

- **We're going to follow the steps below;**
  - Handle data: open the dataset from csv & split it into training and test data
  - Calculate the Euclidean distance between the data sets
  - Locate K most similar/nearest data instances
  - Generate a response from a set of data instances and determine where it belong
  - Summarize the accuracy of the prediction
  - Combine the functions into one executable function

INDUSTRIAL USE CASES OF THE KNN ALGORITHM

- Recommender systems
  - Suggest you which item you are likely to buy based on the items in your shopping cart, or the item you're currently viewing.
- Concept search
  - Extra concepts from a set of documents available in data sets on the internet