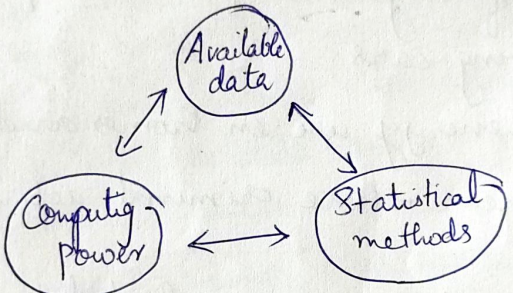
Module II

Module 2

Machine Learning

- The field of machine learning provides a set of algorithms that transform data into actionable knowledge.
- The field of study interested in the development of computer algorithms to transform data into intelligent action is known as Machine Learning
- This field originated in an environment where available data, statistical methods, and computing power rapidly and simultaneously evolved.



A closely related sibling of ML → data mining

eg: Recommendation Engines (Netflix)

- Machine learning is used in internet search engines, email filters to sort out spam, websites to make personalised recommendations, banking software to detect unusual transactions, and lots of apps on our phones such as voice recognition.

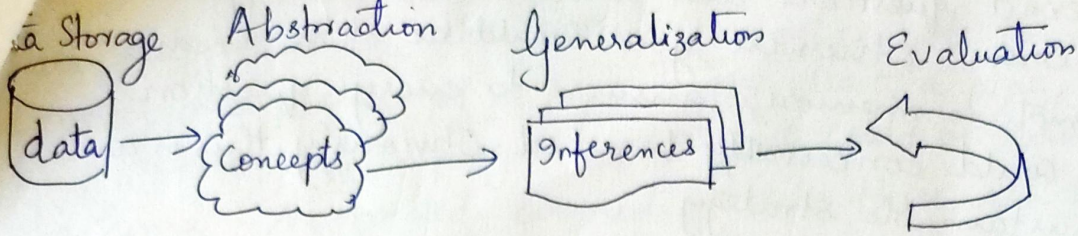
applications of Machine learning includes.

- Identification of unwanted spam messages in e-ma
- Segmentation of customer behavior for targeted advertising
- Forecasts of weather behaviour and long-term climate changes.
- Reduction of fraudulent credit card transactions.
- Actuarial estimates of financial damage of storms and natural disasters.
- prediction of popular election outcomes
- Development of algorithms for auto-piloting drones and self-driving cars.
- Optimization of energy use in homes and office buildings
- projection of areas where criminal activity is most likely
- Discovery of genetic sequences linked to diseases.

How machines learn

Regardless of whether the learner is a human or machine, the basic learning process is similar. It can be divided into Four interrelated components:

1. Data storage
2. Abstraction
3. Generalization
4. Evaluation



The fig illustrates the steps in the learning process.

With computers these processes are explicit, and because of the entire process is transparent, the learned knowledge can be examined, transferred, and utilized for future action.

① Data Storage

- Utilizes Observation, memory, and recall to provide a factual basis for further reasoning.
- All learning must begin with data. Humans and computers alike utilize data storage as a foundation for more advanced reasoning.
- In a human being, this consists of a brain that uses electrochemical signals in a network of biological cells to store and process observations for short and long-term recall. Computers have similar capabilities of short- and long-term recall using hard disk drives, flash memory and Random access memory (RAM) in combination with a central processing unit (CPU).

To better understand the nuances of this idea, it may help to think about the last time you studied for a difficult test, perhaps for a university final exam or a career certification. Did you wish for an eidetic (photographic) memory? If so, you may be disappointed to learn that perfect recall is unlikely to be of much assistance. Even if you could memorize material perfectly, your rote learning is of no use, unless you know in advance

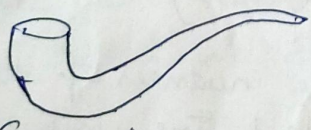
the exact questions and answers that will appear in the exam. otherwise, you would be stuck in an attempt to memorize answers to every question that could conceivably be asked. Obviously this is an unsustainable strategy.

Instead, a better approach is to spend time selectively, memorizing a small set of representative ideas while developing strategies on how the ideas relate and how to use the stored information. In this way, large ideas can be understood without needing to memorize them by rote.

his suggests connect the to a

② Abstraction

- involves the translation of stored data into broader representations and concepts.
- This works of assigning meaning to stored data occurs during the abstraction process, in which raw data comes to have a more abstract meaning.
- This type of connection say between an object and its representation, is exemplified by the famous Rene Magritte painting 'The Treachery of Images':



Ceci n'est pas une pipe. ⇒ caption

The painting depicts a tobacco pipe with the caption ("this is not a pipe"). The point Magritte was illustrating is that a representation of a pipe is not truly a pipe. Yet, in spite of the fact that the pipe is not real, anybody viewing the painting easily recognizes it as a pipe.

This suggests that the observer's mind is able to connect the picture of a pipe to the idea of a pipe, to a memory of a physical pipe that could be held in the hand. Abstracted connections like these are the basis of knowledge representation, the formation of logical structures that assist in turning raw sensory information into a meaningful insight.

- During a machine's process of knowledge representation the computer summarizes stored raw data using a model, an explicit description of the patterns within the data.

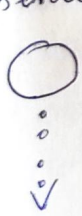
There are many different types of models.

- Mathematical equations
- Relational diagrams such as trees and graphs
- Logical if/else rules
- Groupings of data known as clusters

- The process of fitting a model to a dataset is known as training. When the model has been trained, the data is transformed into an abstract form that summarizes the original information.

For instance :- Discovery of gravity — By fitting equations to observational data, Sir Issac Newton inferred the concept of gravity.

Observations ⟶ Data ⟶ Model



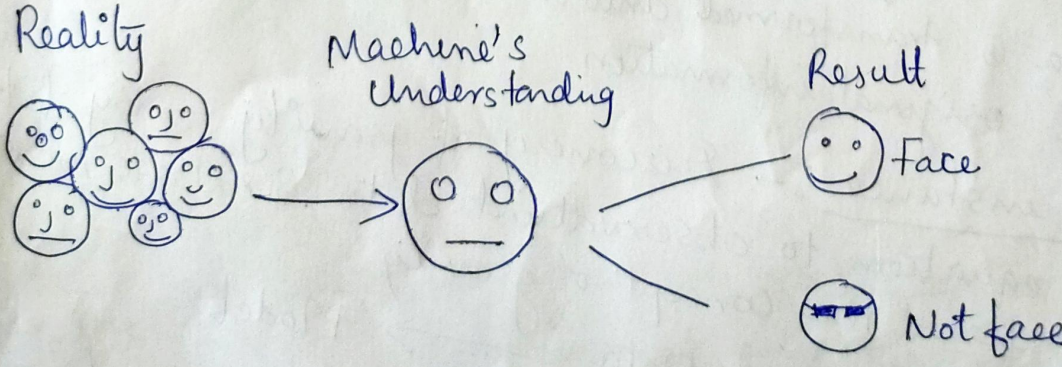
Distance	Time
4.9 m	1s
19.6 m	2s
44.1 m	3s
78.5 m	4s

$g = 9.8\ m/s^2$

③ Generalization : uses abstracted data to create knowledge and inferences that drive action in new contexts.

Evaluation

- The term generalization describes the process of turning abstracted knowledge into a form that can be utilized for future action, on tasks that are similar, but not identical, to those it has seen before.
- Traditionally, it has been imagineed as a search through the entire set of models (i.es theories or inferences) that could be abstracted during training
- In other words, if you can imagenie a hypothetical set containing every possible theory that could be established from the data, generalization involves the reduction of this set into a manageable number of important findings.
- Machine learning algorithms generally employ shortcuts that reduce the search space more quickly. Toward this end, the algorithm will employ heuristics, which are educated guesses about where to find the most useful inferences.



Evaluation

- provides a feedback mechanism to measure the utility of learned knowledge and inform potential improvements
- Generally, evaluation occurs after a model has been trained on an initial training dataset.
- In parts, ~~models~~ for then, the model is evaluated on a new test dataset in order to judge how well its characterization of the training data generalizes to new, unseen data.
- In parts, models fail to perfectly generalize due to the problem of noise, a term that describes unexplained or unexplainable variations in data. Noisy data is caused by seeming random events.

(1) Measurement error due to imprecise sensors that sometimes add or subtract a bit from the readings.

(2) Issues with human subjects, such as survey respondents reporting random answers to survey questions, in order to finish more quickly.

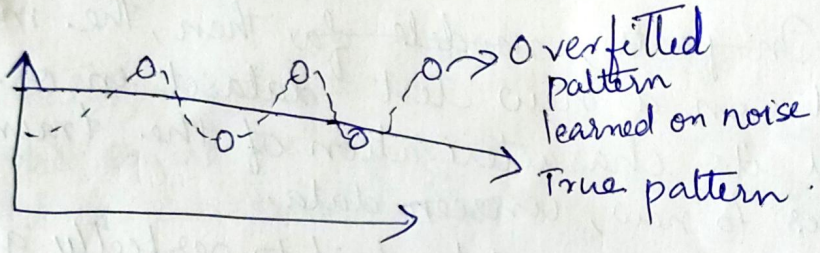
(3) Data quality problems, including missing, null, truncated, incorrectly coded or corrupted values.

(4) Phenomena that are so complex or so little understood that they impact the data in ways that appear to be unsystematic.

Trying to model noise is the basis of a problem called overfitting. Because most noisy data is unexplainable by definition, attempting to explain the noise will result in erroneous conclusions that do not generalize well to new cases.

Efforts to explain the noise will also result in more complex models that will miss the true pattern that the learner tries to identify.

A model that seems to perform well during training, but does poorly during evaluation, is said to be overfitted to the training dataset, as it does not generalize well to the test dataset.



Machine Learning in practice

(1) Data collection : The data collection step involves gathering the learning material, an algorithm will use to generate actionable knowledge.

In most cases, the data will need to be combined into a single source like a textfile, spreadsheet, or database.

(2) Data exploration and preparation — The quality of any ML project is based largely on the quality of its i/p data. Thus, it is important to learn more about the data and its nuances (variations) during a practice called data exploration.

Additional work is needed to prepare the data for the learning process. This involves.

- fixing or cleaning so-called "messy" data,
- eliminating Unnecessary data.
- recoding the data to conform to the learner's expected i/Ps.

(3) Model training : By the time the data has been prepared for analysis, you are likely to have a sense of what you are capable of learning from the data.

The specific ML task chosen will inform the selection of an appropriate algorithm, and the algorithm will represent the data in the form of model.

(4) Model evaluation : - it is important to evaluate how well the algorithm learns from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset

(5) Model improvement :- If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the performance of the model.

Types of input data

Before applying ML to real-world problems, it is important to understand the terminology that distinguishes among i/p datasets.

(1) Unit of observation — used to describe the smallest entity with measured properties of interest for a study.
unit of observation in the form of persons, objects or things, transactions, time points, geographic regions or measurements.

unit of analysis — which is the smallest unit from which the inference is made.
for eg: data observed from people might be used to analyze trends across countries

Datasets that store the units of observation and their properties can be imagined as collections of data consisting of.

(1) Examples : Instances of the unit of observation for which properties have been recorded.

(2) Features : Recorded properties or attributes of examples that may be useful for learning

11:40
12:35

Types of Machine learning algorithms

Machine Learning algorithms are divided into categories according to their purpose.

① predictive model — is used for ~~fo~~ tasks that involve, as the name implies, the prediction of one value using other values in the dataset.

The learning algorithm attempts to discover and model the relationship between the target feature (the feature being predicted) and the other features.

For instance, a predictive model could be used to predict past events, such as the date of a baby's conception using the mother's present-day hormone levels.

• predictive models can also be used in real time to control traffic lights during rush hours.

Because predictive model are given clear instruction on what they need to learn and how they are intended to learn it, the process of training a predictive model is known as supervised learning

② the often used supervised ML task of predicting which category an example belongs to is known as classification. It is easy to think of potential uses for a classifier.

For instance, you could predict whether

- an email msg is spam.
- a person has cancer.
- A football team will win or lose.
- An applicant will default on a loan.

In classification, the target feature to be predicted is a categorical ~~an example belongs to is known as classification~~

It is easiest to understand features and examples through real-world cases. To build a learning algorithm to identify spam e-mail, the unit of observation could be e-mail messages, the examples would be specific messages, and the features might consist of the words used in the messages.

examples and features are commonly collected in matrix format, which means the each example has exactly the same features.

Following spreadsheet shows a dataset in matrix format. In matrix data, each row → example and each column → feature.

↓

it records various automobile features such as price, mileage, color, transmission type.

row indicates examples of automobiles,

features.

Year	Model	Price	mileage	Color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	7351	Gray	AUTO
2011	SEL	19995	11613	Silver	AUTO
2012	SEL	17809	8367	Gray	Manual
2010	SEL	17500	25125	Blue	AUTO
2011	SES	17000	21026	Silver	AUTO
2010	SEL	16995	32655	Silver	AUTO

examples.

Features comes in various forms.

↙

if a feature represents a characteristic measured in numbers, it is called numeric.

↘

Alternatively, if a feature is an attribute that consists of a set of categories, the feature is called categorical or nominal

→ Special case of categorical variables is called ordinal. (categorised in an ordered

feature known as the class, and is divided into categories called levels. A class can have two or more levels, and the levels may or may not be ordinal.

There are many types of classification algorithms, with strengths and weaknesses suited for different types of i/p data.

- Supervised learners can also be used to predict numeric data such as income, laboratory values, test scores, or counts of items. To predict such numeric values, a common form of numeric prediction fits linear regression models to the input data.
- Regression methods are widely used for forecasting, as they quantify in exact ~~items~~ terms the association between inputs and the target, including both, the magnitude and uncertainity of the relationship.
- A descriptive model is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. As opposed to predictive models that predict a target of interest, in a descriptive model, no single feature is more important than any other.

In fact, because there is no target to learn, the process of training a descriptive model is called Unsupervised learning.

For eg: the descriptive modeling task called pattern discovery is used to identify useful associations within data. Pattern discovery is often used for market basket analysis on retailers' transaction purchase data. Here, the goal is to identify items that are frequently purchased together, such that the learned information can be used to refine marketing tactics.

The descriptive modeling task of dividing a into homogeneous group is called clustering. this is sometimes used for segmentation analysis that identifies groups of individuals with similar behavior or demographic information, so that advertising campaigns could be tailored for particular audiences.

Although the machine is capable of identifying the clusters, human intervention is required to interpret them.

for eg: given five different clusters of shoppers at a grocery store, the marketing team will need to understand the differences among the groups in order to create a promotion that best suits each group.

Lastly, a class of machine learning algorithms known as meta-learners is not tied to a specific learning task, but is rather focused on learning how to learn more effectively.
A meta-learning algorithm uses the result of some learnings to inform additional learnig.

Date of announcement :-
conduct :-
Date of return :-

Lazy learning : classification using k-Nearest Neighbour algorithm - Measuring similarity with distance, choice of k, preparing data for use with k-NN.

- ~~Machine~~ Understanding nearest neighbor classification

- Nearest neighbor classifiers are defined by their characteristic of classifying unlabeled examples by assigning them the class of similar labeled examples.

- Despite the simplicity of this idea, nearest neighbor methods are extremely powerful. They have been used successfully for.
  - computer vision applications, including optical character recognition and facial recognition in both still images and video.
  - predicting whether a person will enjoy a movie or music recommendation
  - Identifying patterns in genetic data, perhaps to use them in detecting specific proteins or diseases.

In general, nearest neighbor classifiers are well-suited for classification tasks, where relationships among the features and the target classes are numerous, complicated, or extremely difficult to understand, yet the items of similar class type tend to be fairly homogeneous.

The K-NN algorithm

The nearest neighbors approach to classifica
is exemplified by the k-nearest neighbors algorithm
(K-NN).

One of the simplest machine learning algorithms,
it is still used widely.

- The K-NN algorithm gets its name from the fact that it uses information about an example's k-nearest neighbors to classify unlabeled examples.

letter k is a variable term ⟹ implying that any number of nearest neighbors could be used.

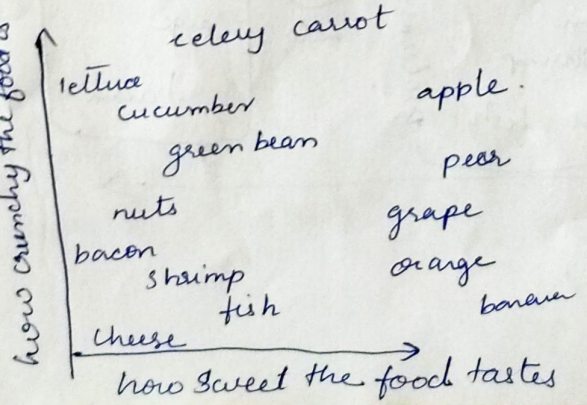
After choosing k, the algorithm requires a training dataset made up of examples that have been classified into several categories, as labeled by a nominal variable.
Then for each unlabeled record in the test dataset, K-NN identifies k records in the training data that are the "nearest" in similarity.
The unlabelled test instance is assigned the class of the majority of the k nearest neighbors.

~~Ingredient~~

Suppose that prior to eating the mystery meal we had created a dataset in which we recorded our impressions of a no: of ingredients we tasted previously. To keep things simple, we rated only two features of each ingredient. The first is a measure from 1 to 10 of how crunchy the ingredient is and the second is a 1 to 10 score of how sweet the ingredient tastes. We then labeled each ingredient as one of the three types of food: fruits, vegetables, or proteins.

Ingredient	Sweetness	Crunchiness	Food type.
apple	10	9	fruit
bacon	1	4	Protein
banana	10	1	fruit
carrot	7	10	Vegetable
celery	3	10	Vegetable
cheese	1	1	Protein

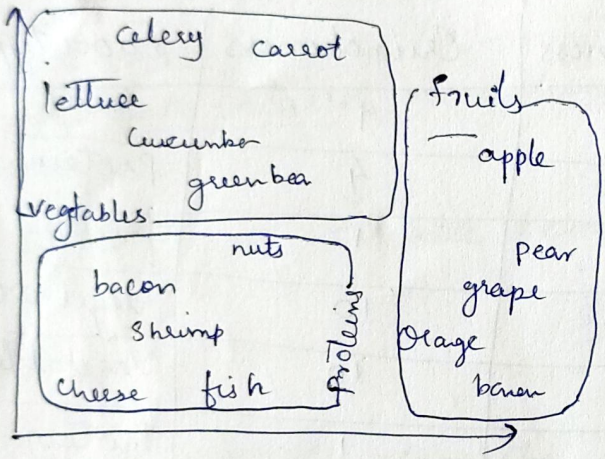
KNN algorithm treats the features as coordinates in a multidimensional feature space.
Dataset includes 2 features, the feature space is 2-Dimentional.
We can plot 2-dimensional data on a scatter plot with
x dimension → ingredients sweetness.
y dimension → the crunchiness.
After adding few more ingredients.



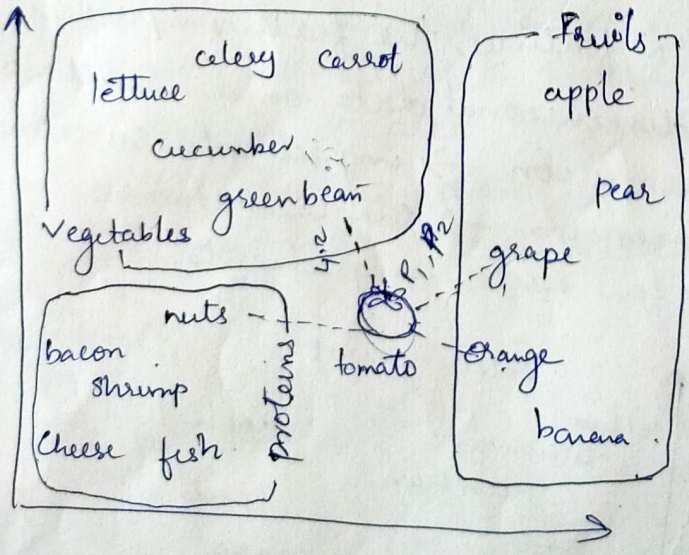
Similar types of food tend to be grouped closely together.

here, vegetables tend to be crunchy but not sweet,
fruits tend to be sweet and either crunchy or not crunchy,
while proteins tend to be neither crunchy nor sweet.

KNN.



Measuring similarity with distance



Locating the tomato's nearest neighbors requires a distance function, or a formula that measures the similarity between the two instances.

(1) Euclidean distance, which is the distance one would measure if it were possible to use a ruler to connect two points, using dotted lines connecting the tomato to its neighbors.
Here we are finding shortest direct route.

Euclidean formula

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

P & q are examples to be compared, each having n features.

p_1 – refers to the value of the first feature of example p, while

q_1 – refers to the values of the first feature of example q.

The distance formula involves comparing the values of each feature.

eg: to calculate the distance between the tomato (sweetness = 6, crunchiness = 4), and green bean (sweetness = 3, crunchiness = 7)

we can use formula as follows.

$$dist(\text{tomato}, \text{green bean}) = \sqrt{(6-3)^2 + (4-7)^2} = 4.2$$

Similarly, we can calculate the distance between the tomato and several of its closest neighbors as follows.

grape

Sweetness - 8

Crunchiness - 5

fruit type - fruit

Distance to the tomato = Sqrt((6-8)^2 + (4-5)^2) = 2.2

green bean

Sweetness - 3

Crunchiness - 7

fruit type - vegetable

Distance to the tomato = Sqrt((6-3)^2 + (4-7)^2) = 4.2

nuts

Sweetness - 3

Crunchiness - 6

protein

Distance = sqrt((6-3)^2 + (4-6)^2) = 3.6

Orange

Sweetness - 7.

Crunchiness - 3

fruit.

Distance = Sqrt((6-7)^2 + (4-3)^2) = 1.4

neighbors be

This is called 1-NN classification because $k = 1$

The orange is the nearest neighbor to the tomato, with a distance of 1.4.

As orange is a fruit, the 1-NN algorithm would classify tomato as a fruit

---

Choosing an appropriate k

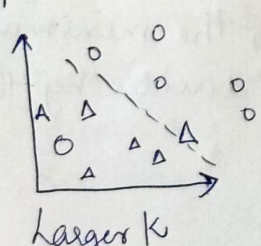
The decision of how many neighbors to use for k-NN determines how well the model will generalize to future data. The balance b/w overfitting and underfitting the training data is a problem known as bias-variance tradeoff.

Choosing a large k reduces the impact or variance caused by noisy data, but can bias the learner so that it runs the risk of ignoring small, but important patterns

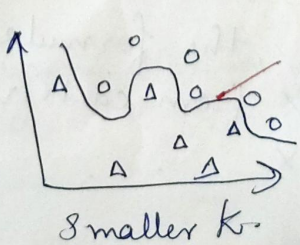
The following figure illustrates, more generally, how the decision boundary (depicted by a dashed line) is affected by larger or smaller k values.

Smaller values allow more complex decision boundaries that more carfully fit the training data.

The problem is that we do not know whether the straight boundary or the curved boundary better represents the true underlying concepts to be learned.



Larger k



Smaller k.

choosing k depends on the difficulty of the concept to be learned, and the number of records in the training data.

One common practice is to begin with $k = \sqrt{\text{no: of training examples}}$

In the food classifier we developed previously, we might set k = 4 because there were 15 example ingredients in the training data and the Sq root of 15 is 3.87

- An alternative approach is to test several k values on a variety of test dataset and choose the one that delivers the best classification performance.

That said, unless the data is very noisy, a ~~large~~ large training dataset can make the choice of k less important. This is because even subtle concepts will have a sufficiently large pool of egs (examples) to vote as nearest neighbors.

Pending

Preparing data for use with k-NN.

The traditional method of rescaling features for k-NN is min-max normalization. ✓

This process transforms a feature such that all of its values fall in a range between 0 and 1.

The formula for normalizing a feature is

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Essentially, the formula subtracts the minimum of feature X from each value and divides by the range of X

~~choosing k depends on the diffi~~

Normalized features values can be interpreted from 0 percent to 100 percent, the original values can be fell in along the range between the original minimum and maximums.

- Another common transformation is called Z-score standardization.
The following formula subtracts the mean value of feature X, and divides the outcome by the standard deviation of X

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - Mean(X)}{StdDev(X)}$$

formula, which is based on the properties of normal distribution.

- rescales each of the features values in terms of how many standard deviations they fall above or below the mean value. The resulting value is called a z-score.
The z-scores fall in an unbound range of negative and positive numbers.
Unlike the normalized values, they have no predefined minimu and maximum.

- The Euclidean distance formula is not defined for nominal data. Therefore to calculate the distance between nominal features, we need to convert them into a numeric format.
A typical solution utilizes dummy coding, where a value of
1, indicates one category and
0, the other.

for eg, dummy coding for a gender variable could be constructed as

$$\text{male} = \begin{cases} 1 & \text{if } x = \text{male} \\ 0 & \text{otherwise} \end{cases}$$

Notice how the dummy coding of the two category (binary) gender variable results in a single new feature named male. There is no need to construct a separate feature for female. since the two sexes are mutually exclusive, knowing one or the other is enough.

An n-category nominal feature can be dummy coded by creating the binary indicator variables for (n-1) levels of the feature.

for eg: the dummy coding for a three-category temperature variable (hot, medium, or cold) could be set up as (3-1) = 2 features, as shown here.

$$\text{hot} = \begin{cases} 1 & \text{if } x = \text{hot} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{medium} = \begin{cases} 1 & \text{if } x = \text{medium} \\ 0 & \text{otherwise} \end{cases}$$

Knowing that hot and medium are both 0 is enough to know that the temperature is cold. We therefore do not need a third feature for the cold category.

Why is the K-NN algorithm lazy?

A lazy learner is not really learning anything. Instead, it merely stores the training data verbatim. Verbatim

This allows the training phase, which is not actually training anything, to occur very rapidly.

The downside is that the process of making predictions tends to be relatively slow in comparison to training.

Due to the heavy reliance dependence on the training instances rather than an abstracted model, lazy learning is also known as instance-based learning or rote learning.

The instance based learners do not build a model, the method is said to be in a class of non-parametric learning methods - no parameters are learned about the data.

non parametric methods limit our ability to understand how the classifier is using the data. On the other hand this allows the learner to find natural patterns rather than trying to fit the data into a preconceived and potentially biased functional form.