Data Science

Data science is the practice of using computational methods to derive valuable and actionable insights from raw datasets. Data scientists must have extensive and diverse quantitative expertise to be able to solve these types of problems. A data scientist must also have **subject matter expertise** in the particular area in which they work.

Data engineering, on the other hand, is an engineering domain that's dedicated to overcoming data-processing bottlenecks and data-handling problems for applications that utilize large volumes, varieties, and velocities of data. With respect to data science, the purpose of data engineering is to engineer big data solutions by building coherent, modular, and scalable data processing platforms from which data scientists can subsequently derive insights. Data engineers need solid skills in computer science, database design, and software engineering to be able to perform this type of work.

Hire a data engineer to process your data and a data scientist to make sense of it for you.

Software as a Service (SaaS) is a term that describes cloud-hosted software services that are made available to users via the Internet.

✓ Structured data: Data that's stored, processed, and manipulated in a traditional relational database management system.

✓ Unstructured data: Data that's commonly generated from human activities and that doesn't fit into a structured database format. Such data could be derived from blog posts, emails, and Word documents.

✓ Semi-structured data: Data that doesn't fit into a structured database system, but is nonetheless structured by tags that are useful for creating a form of order and hierarchy in the data. Semi-structured data is commonly found in database and file systems. It can be stored as log files, XML files, or JSON data files.

Mathematics uses deterministic numerical methods and deductive reasoning to form a quantitative description of the world

statistics is a form of science that's derived from mathematics, but that focuses on using a stochastic approach — an approach based on probabilities — and inductive reasoning to form a quantitative description of the world.

The main point of distinction between statistics and data science is the need for subject-matter expertise.

Data scientists, are required to have a strong subject-matter expertise in the area in which they're working. Data scientists generate deep insights and then use their domain-specific expertise to understand exactly what those insights mean with respect to the area in which they're working.

As a data scientist, you must have sharp oral and written communication skills.

Cloud applications such as IBM's Watson Analytics automated data services — from cleanup and statistical modeling to analysis and data visualization.

Big data is data that exceeds the processing capacity of conventional database systems because it's too big, it moves too fast, or it doesn't fit the structural requirements of traditional database architectures.

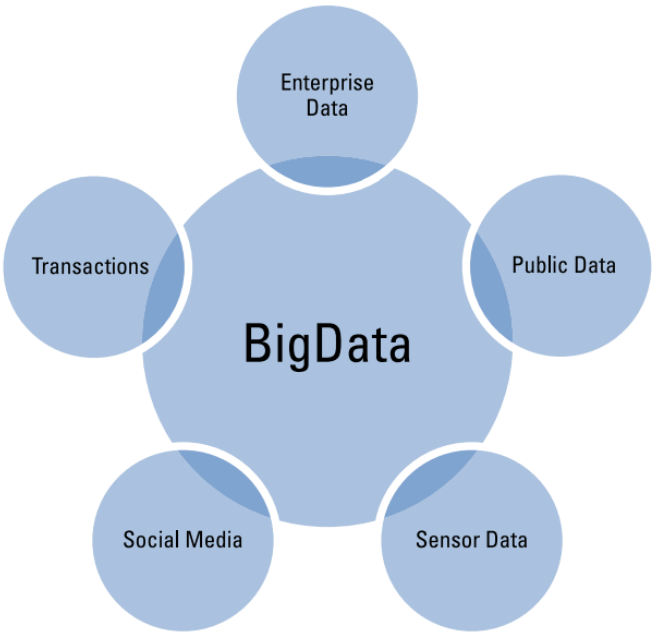
Four characteristics (the four Vs) define big data: volume, velocity, variety, and value.

data velocity is data volume per unit time. Big data velocities range anywhere between 30 kilobytes (K) per second up to even 30 gigabytes (GB) per second. The capabilities of data-handling and data-processing technologies often limit data velocities.

High-variety data sources can be derived from data streams that are generated from social networks or from automated machinery.

In its raw form, most big data is low value — in other words, the value-to-data quantity ratio is low in raw big data. Big data is comprised of huge numbers of very small transactions that come in a variety of formats.

Popular big data sources:



MapReduce is a programming paradigm that was designed to allow **parallel distributed processing** of large sets of data, converting them to sets of tuples, and then combining and reducing those tuples into smaller sets of tuples. MapReduce was designed to take big data and use parallel distributed computing to turn big data into little- or regular- sized data.

Parallel distributed processing refers to a powerful framework where mass volumes of data are processed very quickly by distributing processing tasks across clusters of commodity servers.

In the map task, you delegate your data to key-value pairs, transform it, and filter it. Then you assign the data to nodes for processing. In the reduce task, you aggregate that data down to smaller sized datasets. Data from the reduce step is transformed into a standard

key value format — where the key acts as the record identifier and the value is the value that's being identified by the key. Every reduce task has a fragment assigned to it. The reduce task simply processes the fragment and produces an output, which is also a key-value pair.

computing cluster: a group of nodes that are connected to each other and perform a shared computing task.

Hadoop Distributed File System (HDFS) is a file system that includes clusters of commodity servers that are used to store big data. HDFS makes big data handling and storage financially feasible by distributing storage tasks across clusters of cheap commodity servers.

Hadoop can offer you a great solution to handle, process, and group mass streams of structured, semi-structured, and unstructured data.

Distributed processing framework is a powerful framework where processing tasks are distributed across clusters of nodes so that large data volumes can be processed very quickly across the system as a whole.

Hadoop also supports hierarchical organization. Some of its nodes are classified as master nodes (JobTracker), and others are categorized as slaves (TaskTracker).

ACID: Atomicity, Consistency, Isolation, and Durability compliance.

In big data solutions, most database systems are not ACID compliant because they use Decision Support Systems (DSS) that batch process data before that data is read out.

Hadoop is a batch processor and can't process real-time, streaming data.

Real-time processing framework is a framework that is able to process data in real-time (or near real-time) as that data streams and flows into the system that can be classified into two categories:

✓ Frameworks that lower the overhead of MapReduce tasks to increase the overall time efficiency of the system: Solutions in this category include Apache Storm and Apache Spark for near–real-time stream processing.

✓ Frameworks that deploy innovative querying methods to facilitate real-time querying of big data: Some solutions in this category include Google's Dremel, Apache Drill, Shark for Apache Hive, and Cloudera's Impala.

Massively Parallel Processing (MPP) platforms can be used instead of MapReduce as an alternative approach for distributed data processing, if your goal is to deploy parallel processing on a traditional data warehouse.

MPP runs parallel computing tasks on costly, custom hardware, whereas MapReduce runs them on cheap commodity servers.

Well-known MPP products include the old-school Teradata platform, plus newer solutions like EMC2 's Greenplum DCA, HP's Vertica, IBM's Netezza, and Oracle's Exadata.

NoSQL databases, like MongoDB, are non-relational, distributed database systems that were designed to rise to the big data challenge and are able to handle the structured, semi-structured, and unstructured data sources that are common in big data systems.

NoSQL offers four categories of non-relational databases: graph databases, document databases, key-values stores, and column family stores.

data silos are data repositories that are disconnected and isolated from other data storage systems used across the organization.
Linked data format: a format that facilitates a joining of the different datasets in the Hadoop clusters.
Automatic failover configuration: is a configuration that facilitates an automatic switch to redundant, backup data handling systems in instances where the primary system might fail.

Types of data analytics

Descriptive analytics: This type of analytics answers the question, "What happened?" Descriptive analytics are based on historical and current data.**چی شد**
Diagnostic analytics: finds answers to the question, "why did this particular something happen?" or "what went wrong?" Diagnostic analytics are useful for deducing and inferring the success or failure of sub-components of any data-driven initiative.**چرا اینطوری شد**
Predictive analytics involve complex model building and analysis in order to predict a future event or trend.**حالا چطور میشه**
Prescriptive analytics: This type of analytics aims to optimize processes, structures, and systems through informed action that's based on predictive analytics (essentially telling you what you should do based on an informed estimation of what will happen).**حالا که قراره اینجوری شه چکار کنیم**

Three main requirements of any business-centric data scientist:

- **strong business knowledge**
- **strong coding acumen**
- **strong quantitative analysis skills via math and statistical modeling**

data wrangling: The processes and procedures that you use to clean and convert data from one format and structure to another so that the data is accurate and in the format analytics tools and scripts require for consumption, including

- Data extraction: to identify what datasets are relevant to the problem at hand, and then extract sufficient quantities of the data that's required to solve the problem. (This extraction process is commonly referred to as data mining)
- Data munging: involves cleaning the raw data extracted through data mining, then converting it into a format that allows for a more convenient consumption of the data.
- Data governance standards: are standards that are used as a quality control measure to ensure that manual and automated data sources conform to the data standards of the model at hand. Data governance standards must be applied so that the data is at the right granularity when it's stored and made ready for use.
    - Granularity is a measure of a dataset's level of detail.
- Data architecture: to be sure that the data is stored in a central data warehouse and not in separate silos.

insight-to-action arc: the process of taking decisive actions based on data insights.
Business Intelligence (BI): is to convert raw data into business insights that business leaders and managers can use to make data-informed decisions comprised of:

- **internal datasets**: supplied by your organization's own managers and stakeholders.
- **Tools, technologies, and skillsets**: online analytical processing, ETL (extracting, transforming, and loading data from one database into another) and data warehousing.

BI solutions are mostly built off of transactional data.

BI types of insights**:**

- **Customer service data** (what areas of business are causing the largest customer wait times?)
- **Sales and marketing data** (which marketing tactics are most effective and why?)
- **Operational data** (how efficiently is the help desk operating? Are there any immediate actions that must be taken to remedy a problem there?)
- **Employee performance data** (which employees are the most productive? Which are the least?)

Multidimensional databases: organize data into cubes that are stored as multi-dimensional arrays. This cubic data structure enables Online Analytical Processing (OLAP) which is a technology through which you can quickly and easily access and use your data for all sorts of different operations and analyses.

OLAP is just one type of data warehousing system: a centralized data repository that you can use to store and access your data.

Data mart: a data storage system that you can use to store one particular focus area of data, belonging to only one line of business in the enterprise.

Extract, transform, and load (ETL) is the process that you'd use to extract data, transform it, and load it into your database or data warehouse.

Business-centric data science: acts if you have large sets of structured and unstructured data sources that may or may not be complete and you want to convert those sources into valuable insights for decision support across the enterprise and has these elements:

- **Quantitative analysis**: mathematical modeling, multivariate statistical analysis, forecasting, and/or simulations.
- **Programming skills**
- **Business knowledge**

Data scientists often:

- employ the scientific method for data exploration, hypotheses formation, and hypothesis testing (through simulation and statistical modeling).
- create business data mash-ups from internal and external sources of structured and unstructured data fairly easily. Data mash-up is combination of two or more data sources that are then analyzed together in order to provide users with a more complete view of the situation at hand.
- Tools, technologies, and skillsets: using cloud-based platforms, statistical and mathematical programming, machine learning, data analysis using Python and R, and advanced data visualization.

Data scientists create valuable insights from all available data sources such as:

- **Transactional business data**

- **Social data related to the brand or business**
- **Machine data from business operations**: like SCADA data, machine data, or sensor data. SCADA refers to Supervisory Control and Data Acquisition, systems are used to control remotely operating mechanical systems and equipment.
- **Audio, video, image, and PDF file data**

Business Intelligence Vs. business-centric data science

- **Data sources**: BI uses only structured data from relational databases, whereas business-centric data science may use structured data and unstructured data, like that generated by machines or in social media conversations.
- **Outputs**: BI products include reports, data tables, and decision-support dashboards, whereas business-centric data science products either involve dashboard analytics or another type of advanced data visualization, but rarely tabular data reports. Data scientists generally communicate their findings through words or data visualizations, but not tables and reports. That's because the source datasets from which data scientists work are generally more complex than a typical business manager would be able to understand.
- **Technology**: BI runs off of relational databases, data warehouses, OLAP, and ETL technologies, whereas business-centric data science often runs off of data from data-engineered systems that use Hadoop, MapReduce, or Massively Parallel Processing.
- **Expertise**: BI relies heavily on IT and business technology expertise, whereas business-centric data science relies on expertise in statistics, math, programming, and business.

Basics of statistical probability

Statistics: a result that's derived from performing a mathematical operation on numerical data.

Descriptive Statistics: providing you with a description that illuminates some characteristic of your numerical dataset.

Inferential statistics: extract a smaller section of the dataset (samples) and attempt to deduce something significant about the larger dataset (population). For an inference to be valid, you must select your sample carefully so that you get a true representation of the population. However, the sample statistic is not exactly identical to its corresponding population statistic.

Random variable: is a measure of a trait or value associated with an object, a person, or a place — something in the real world — that is unpredictable.

Expectation is a weighted average of some measure associated with a random variable.

Discrete distribution: A random variable of which values can be counted.

Continuous distribution: A random variable that assigns probabilities to ranges of values.

Eight distinct classes of probability distributions:

- **Uniform distributions**: Used to distribute probability equally over all possible outcomes (discrete) or equal ranges of outcomes (continuous).
- **Binomial distributions**: Model the number of successes that can occur in a certain number of attempts when only two outcomes are possible

- **Geometric distributions**: is the count of the number of failures before the first success. (with 3 preconditions, see WIKI)
- **Hypergeometric distributions**: gives the probability that some number from a sample will be of a particular value. (without replacement of the samples) (each draw decrease the population)
- **Poisson and exponential distributions**: The Poisson (discrete) and exponential (continuous) distributions complement one another. Say that there is an intersection in your town that has a lot of accidents. A Poisson distribution answers the question, "What is the probability that such-and-such number of accidents will occur there within a week?" And an exponential distribution answers the question, "What is the probability that the time until the next accident is such-and-such length of time?
- **Normal distributions (continuous)**: Gaussian Distribution, models phenomena that tend toward some most-likely value (mean) with values at the two extremes becoming less likely
- **Gamma distributions**: allows non symmetric distributions and plans for all possibilities — even the possibility that no pattern exists. A method that's useful for predicting when a device is most likely to fail, based on observational data and probability distributions.
- **Beta distributions (continuous)**: The most versatile player on the distributions team, beta distributions are structured so that they can be made to fit almost any pattern, as needed.

Linear regression: a method for modeling the relationships between a dependent variable and one or several independent variables.

Simple linear regression: a mathematical modeling tool that you can use to predict the value of a dependent variable (DV) based on a single independent variable (IV) . (Linear relationships are represented mathematically by a straight line).

$$y=ax+b+\varepsilon$$

If you have some data on a real-world phenomenon and you pair this data as two variables in a linear relationship, you're not saying that one variable necessarily causes the other. You're simply using a linear relationship to show that, while one value goes up, the other also increases or decreases proportionally. To test evidence for guesses you form about the world, you can use linear regression.

The regression line will not fit the data exactly because each observation is assumed to contain another random noise variable, ε. Fitting the regression line to the data involves finding the values of a and b such that the sum of ε for all observations in the dataset is minimized.

Noise, or random noise, is unexplainable variation that is present in almost all datasets. A major assumption in performing hypothesis testing is that noise terms are symmetrically distributed. If they're not, then a linear model is not appropriate for your data.

Ordinary Least Squares (OLS) is a statistical method that fits a linear regression line to an observational dataset: squaring the vertical distance values that describe the distances

between the observational data points and the best-fit line, adding up those squared distances, and then adjusting the placement of the best-fit line so that this summed squared distance value is minimized. OLS is particularly useful for fitting a regression line to models containing more than one independent variable (IV).

When two or more IVs are strongly correlated with each other, this is called multicollinearity.

The ε 's regression errors for each of the observations should have following properties:

- normally distributed
- the regression errors should average out to zero
- the regression errors should have approximately the same amount of variability.

If your data does not meet these specification, then a linear model might not be appropriate.

Monte Carlo method is a simulation technique that randomly generates values for independent variables, and then simulates a range of potential process outcomes based on this set of randomly generated IVs.

The purpose of using Monte Carlo methods is to sample values randomly from a probability distribution and then use those values to make a prediction or validate a hypothesis. If you want to save money that you'd otherwise need to spend on real-world experimentations, then you can use computational simulations to model a natural process mathematically instead.

Cumulative Probability Distribution (CPD): is a distribution function that uses a cumulative probability to estimate the probability of a random variable falling within a particular range of numbers.

With Monte Carlo sampling, you assume that the null hypothesis is correct and then determine how probable your real-world observations are, given its assumptions.

An estimator is an educated guess about some measure associated with a real-world process.

در این تکنیک ما در واقع یک معادله یا مدل فرضی را برای تشریح یک واقعه حقیقی ایجاد می نماییم. سپس از رویه های Monte Carlo Simulation برای تولید متغیرهای تصادفی با ویژگی های آماری منطبق با آن واقعه حقیقی استفاده می کنیم تا در نهایت با تغذیه مدل پیشنهادی مان برای آن واقعه با داده های تصادفی ایجاد شده، نسبت به اثبات صحت فرضیه خود اقدام نماییم.

The level of confidence for a confidence interval is a measure of the likelihood that the unknown parameter is represented within the interval.

Time Series Analysis:

A time series is just a collection of data on attribute values over time.

Time series analysis is performed in order to predict future instances of the measure based on the past observational data.

Constant time series remain at roughly the same level over time, but are subject to some random error.

Trended series show a stable linear movement up or down.

Seasonality predictable, cyclical fluctuations that reoccur seasonally throughout a year.

Time series may show non-stationary processes that is unpredictable cyclical behavior that's not related to seasonality as a result of economic or industry-wide conditions and

cannot be forecasted, so you must transform non-stationary data to stationary data before moving forward with an evaluation.
Normally, the actual data contains some random error (noise), thus making it impossible to forecast perfectly.

Time Series Type	Mathematical Model
Constant	$y_t = b + random\ error$
Trended	$y_t = at + b + random\ error$
Constant Seasonal	$y_t = b + S^{(1)}I^{(1)} + S^{(2)}I^{(2)} + S^{(3)}I^{(3)} + S^{(4)}I^{(4)} + random\ error$
Trended Seasonal	$y_t = at + b + S^{(1)}I^{(1)} + S^{(2)}I^{(2)} + S^{(3)}I^{(3)} + S^{(4)}I^{(4)} + random\ error$
Cyclical	Here cycles are unpredictable and hence impossible to model

Explanation of Notation:

- y_t is the observation at the tth time period;
- a, b are constants; $t = 1, 2, 3, \ldots$ is time period;
- $S^{(i)}$ is seasonal constant for season $i = 1,\ 2,\ 3,\ 4^*$ and $\sum_{i=1}^{4} I^{(i)} = 1$ with $I^{(i)} = 0$ or 1

Autoregressive Moving Average (ARMA) is a class of forecasting methods that you can use to predict future values from current and historical data and combine these two techniques:

1. Autoregression techniques assume that previous observations are good predictors for future values and perform an autoregression analysis to forecast for those future values.
2. Moving average techniques measure the level of the constant time series and then update the forecast model if any changes are detected.

If you're looking for a simple model or a model that will work for only a small dataset, then the ARMA model is **not** a good fit for your needs. An alternative in this case might be to just stick with simple linear regression.

In order to use the **ARMA model** for reliable results, you need to have at least 50 observations and a trained analyst who can fit and interpret the model for you.

Clustering and Classification

Learning refers to an algorithm that's repeated over and over until a certain set of predetermined conditions are met.
In unsupervised learning algorithms must use inferential methods to discover patterns, relationships, and correlations within the raw data set.
If you have a dataset that describes multiple attributes about a particular feature and want to group your data points according to their attribute similarities, then use clustering algorithms.
Eyeballing: making a visual estimate of clusters in a dataset.

You can use clustering algorithms to generate clustering for **multiple dimensions** of data within your dataset.
Other unsupervised learning approaches include Markov methods and methods for dimension reduction.

Clustering pre-conditions:

1. You know and understand the dataset you're analyzing.
2. Before running the clustering algorithm, you don't have an exact idea as to the nature of the subsets (clusters). Often, you won't even know how many subsets there are in the dataset before you run the algorithm.
3. The subsets (clusters) are determined by only the one dataset you're analyzing.
4. Your goal is to determine a model that describes the subsets in a single dataset and only this dataset.

When implementing supervised classification, you should already know your data's subsets called categories.

Model overfitting: situations in which a model is so tightly fit to its underlying dataset, as well as the noise or random error inherent in that dataset, that the model performs poorly as a predictor for new data points. (low bias, high variance)
Overgeneralization is the opposite of overfitting: It happens when a data scientist tries to avoid misclassification due to overfitting by making a model extremely general. (high bias, low variance)

Classification pre-conditions:

1. You know and understand the dataset you're analyzing.
2. The subsets (categories) of your dataset are defined ahead of time and aren't determined by the data.
3. You want to build a model that correlates the data within its predefined categories so that the model can help predict the categorization of future data points.

Similarity Metrics:

Some popular geometric metrics used for calculating distances between data points include Euclidean, Manhattan, or Minkowski distance metrics.

If your data is numeric but non-plottable (such as curves instead of points), you can generate similarity scores based on differences between data, instead of the actual values of the data itself.

For non-numeric data, you can use metrics like the Jaccard distance metric, which is an index that compares the number of features that two data points have in common. For example Jaccard metric generates a numerical index value that quantifies the similarity between text strings.

Clustering Algorithms:

Speed and robustness of the k-means algorithm makes it a very popular choice among experienced data scientists.

As alternatives, Kernel Density Estimation methods, hierarchical algorithms, and neighborhood algorithms are also available to help you identify clusters in your dataset.

K-mean Clustering:
Subdivides data points of a dataset into clusters based on nearest mean values.

If your dataset has more than three dimensions, however, you can use computational methods to generate a good value for k. For example:

Silhouette coefficient: a method that calculates the average distance of each point from all other points in a cluster, and then compares that value with the average distance to every point in every other cluster.

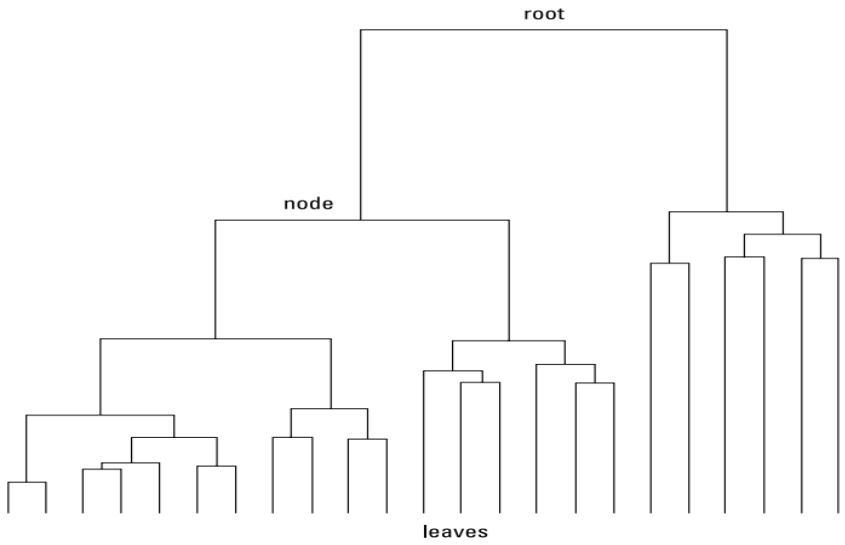
During k-mean clustering, The centers are moved from regions of lower density to regions of higher density until all centers are within a region of local maximum density.

One **weakness** of the **k-means** algorithm is that it may produce incorrect results by placing cluster centers in areas of local minimum density. This happens when centers get lost in low-density regions. Ironically, this happens most often when the underlying data is very well clustered, with tight, dense regions that are separated by wide, sparse areas.

Estimating clusters with Kernel Density Estimation (KDE)
(KDE) is just such a smoothing method; it works by placing a kernel — a weighting function that is useful for quantifying density — on each data point in the data set and then summing the kernels to generate a kernel density estimate for the overall region and generates a plot of gradual density change between data points.

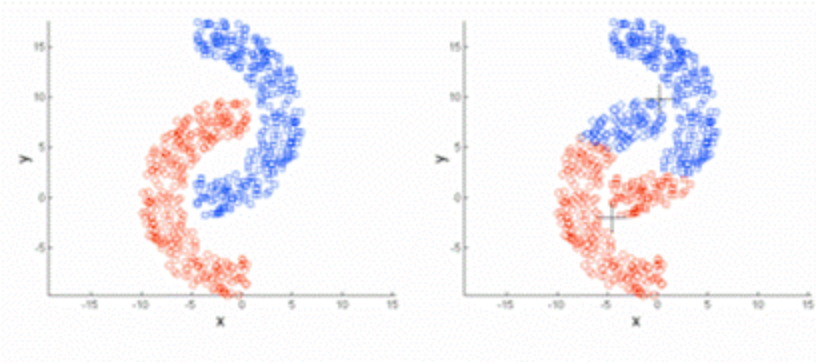
Clustering with hierarchical and neighborhood algorithms

A hierarchical clustering algorithm results in a dataset that's called a dendrogram — the top, or root, of a dendrogram is the entire dataset, each level down is a node where the data is split into two sets (usually of unequal size), and finally at the bottom are leaves that each correspond to a single data point:



Hierarchical clustering algorithms are more computationally expensive than k-means algorithms because with each iteration of hierarchical clustering, many points must be compared to many other points.

Non-globular Cluster — a configuration where some points in a cluster are closer to points in a different cluster than they are to points in the center of their own cluster.



If your dataset shows non-globular clustering, then you can use neighborhood clustering algorithms, like DBSCAN, to determine whether each point is closer to its neighbors in the same cluster, or whether it is closer to its neighboring data points in other clusters. But they have these two weaknesses:

- They can be computationally expensive.
- They have some parameters that If you guess either of them incorrectly, the algorithm misidentifies clusters.

Decision Trees and Random Forests

A decision tree algorithm works by developing a set of yes or no rules that you can follow for new data in order to see exactly how it will be characterized by the model. DTs are prone to error propagation. This occurs when one of the model rules is incorrect.

Random forest algorithms are a slower but more powerful alternative. Instead of building a tree from the data, the algorithm creates random trees and then determines which one best classifies the testing data. This method eliminates the risk of error propagation that is inherent in decision tree models.

Nearest Neighbor Algorithms

Since the nearest neighbor technique is a classification method, you can use it to do things as scientific as deducing the molecular structure of a vital human protein or uncovering key biological evolutionary relationships, and as business-driven as designing recommendation engines for e-commerce sites or building predictive models for consumer transactions. The applications are limitless.

Note that, in this technique similarity comparisons can be based on any quantitative attribute, whether that be distance, age, income, weight, or anything else that can describe the data point you're investigating.

The nearest neighbor algorithm is known as a single-link algorithm — an algorithm that merges clusters if the clusters share at least one connective edge (a shared boundary line).

There are different clustering methods, depending on how you want your dataset to be divided. The two main types of clustering algorithms are

- **Hierarchical**: Algorithms create separate sets of nested clusters, each in their own hierarchal level. Use hierarchical clustering algorithms only if you already know the separation distance between the data points in your dataset. The k-nearest neighbor algorithm belongs to the hierarchical class of clustering algorithms.
- **Partitional**: Algorithms create just a single set of clusters.

Average Nearest Neighbor Algorithms for classification

kNN

In this type of algorithms, the attributes you use as a basis to compare your data points must be quantitative, but the data points you choose to compare can represent either a quantitative or qualitative category.

In kNN, the classifier classifies the query point P according to the classification labels found in a majority of k-nearest points surrounding the query point. kNN is a good classification method for you to use if you know very little about the distribution of your dataset.

When using kNN, it's crucial that you choose a k value that minimizes noise — unexplainable random variation. In other words, it's vital that you choose a k value that includes sufficient data points in the selection process. **The problem with kNN is that it takes a lot longer than other classification methods to classify a sample.** . As with other hierarchical algorithms, you want to use kNN only on datasets that have a few

thousand data points or less. In this capacity, kNN is useful for website categorization, web-page ranking, and other user dynamics across the web. kNN classification techniques are also quite beneficial in customer relationship management (CRM).

Most of the time, you probably wouldn't know that your data points are influencing one another known as second order effects, but you can test for this kind of influence by using nearest neighbor distances to draw inferences from point patterns.

Average nearest neighbor algorithms calculate a descriptive index value that represents the average distance between a data point and its nearest neighbor. If the calculated index value is less than 1, then the data is said to show clustered patterning. If the index value is greater than 1, then the data is said to show dispersion patterning.

Clustered patterning indicates that there is some sort of interaction going on between the data points and that this interaction is causing an increase in average similarity values. In **dispersion patterning**, on the other hand, interaction between the data points causes a decrease in average similarity values.

kNN use cases

1. K-nearest neighbor techniques for pattern recognition are often used for **theft prevention** in the modern retail business. The modern systems are now able to use k-nearest neighbor for visual pattern recognition to scan and detect hidden packages in the bottom bin of a shopping cart at checkout. If an object is detected that's an exact match for an object listed in the database, then the price of the spotted product could even automatically be added to the customer's bill.
2. Average nearest neighbor algorithm classification and point pattern detection can be used in grocery retail to **identify key patterns in customer purchasing behavior**, and subsequently increase sales and customer satisfaction by anticipating customer behavior.

Mathematical Modeling in DS

Multi-Criteria Decision Making (MCDM)

They can be used anywhere you have several criteria on which you need to base your decision, you can use MCDM methods to help you evaluate alternatives.

A set is a group of numbers that shares some similar characteristic in which membership is a binary function (0 or 1)

MCDM techniques are in no way limited to just land use issues or investment portfolio theory. Some other real-world applications where MCDM is useful include sustainable energy development, resource management, scheduling, nuclear medicine, and oilfield

development. Since MCDM is now a mature field of decision science, it's been applied to almost every area of human endeavor.

You can use fuzzy multi-criteria decision making (FMCDM) to evaluate all the same types of problems as you would with MCDM. Evaluations based on fuzzy criteria lead to a range of potential outcomes, each with its own **level of suitability** as a solution. Possible solutions comprise a **spectrum of suitability** that's plotted on a chart as something called an efficient frontier. Fuzzy-set membership can be represented by the numbers 0 and 1, it can also be represented by any number that lies between 0 and 1. Encoding and decoding fuzzy-set membership allows a decision model to prioritize or rank individual items within that model.

Gut feeling: basic feeling or reaction without a logical rationale.

You can use **weighting factors** to quantify the relative significance of criteria in your decision model.

Zero-sum system: Optimizing with respect to one criterion must come at the sacrifice of at least one other criterion.
یعنی برآورده کردن یک معیار موجب کاهش برآورده سازی حداقل یک معیار دیگر شود
in other words, for each solution where you improve with respect to one criterion, you must always give up or lose something with respect to another criterion known as **efficient frontier** which is the central concept of MCDM.

A MCDM evaluation doesn't have just one optimal solution. Instead, you end up with many efficient solutions all caught up in a zero-sum system.

Using Numerical Methods in Data Science (Numerical approximation)

Taylor polynomials

A **Taylor series** is a power series representation of a function that serves as an approximation method for finding a function's value when that value is generated by summing its derivatives at any given degree, n. **Use a Taylor series to approximate the value of a function simply by looking at its first few terms. The higher you go, the more accurate your approximation value will be (with the cost of more computations).**

Taylor polynomial is the polynomial function that forms the basis of a Taylor series.

Figure 7-5: The standard form of a Taylor polynomial.

$$P_n(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}}{n!}(x-a)^n$$

In data science, it's common for a practitioner to have some rough knowledge about what part of the solution space holds the optimal solution. Taylor polynomials offer a standardized method by which you can use your knowledge of the solution space to generate a quick approximation for the value of the function in that region.

bisection search algorithm (simplest root finding method that's available)

A **bisecting search algorithm** is a method for bisecting intervals and searching for input values of a continuous function. Data scientists use a bisection search algorithm as a numerical approach to find a quick approximation of a solution. The algorithm does this by searching and finding the roots of any continuous mathematical function.

You can use the bisect method of the SciPy library to get the job done.

Markov Chains and Stochastic Methods

A **stochastic model** is a tool that you can use to estimate probable outcomes when one or more model variables is changed randomly.

A **Markov chain** — also called a **discrete time Markov chain** — is a stochastic process that acts as a mathematical method to chain together a series of randomly generated variables representing the present state in order to model how changes in those present state variables affect future states.

In Markov methods, future states must depend on the value of the present state and be conditionally independent from all past states known as **Markov Property**.

Markov chains are extremely useful in modeling a variety of real-world processes. They're commonly used in

- **stockmarket exchange models**
- **financial asset-pricing models**
- **speech-to-text recognition systems**
- **webpage search and rank systems**
- **thermodynamic systems**
- **gene-regulation systems**

- **state-estimation models**
- **pattern recognition**
- **population modeling**

A Markov chain will eventually reach a **steady state (equilibrium)** — a long-term set of probabilities for the chain's states.

Modeling Spatial Data with Statistics