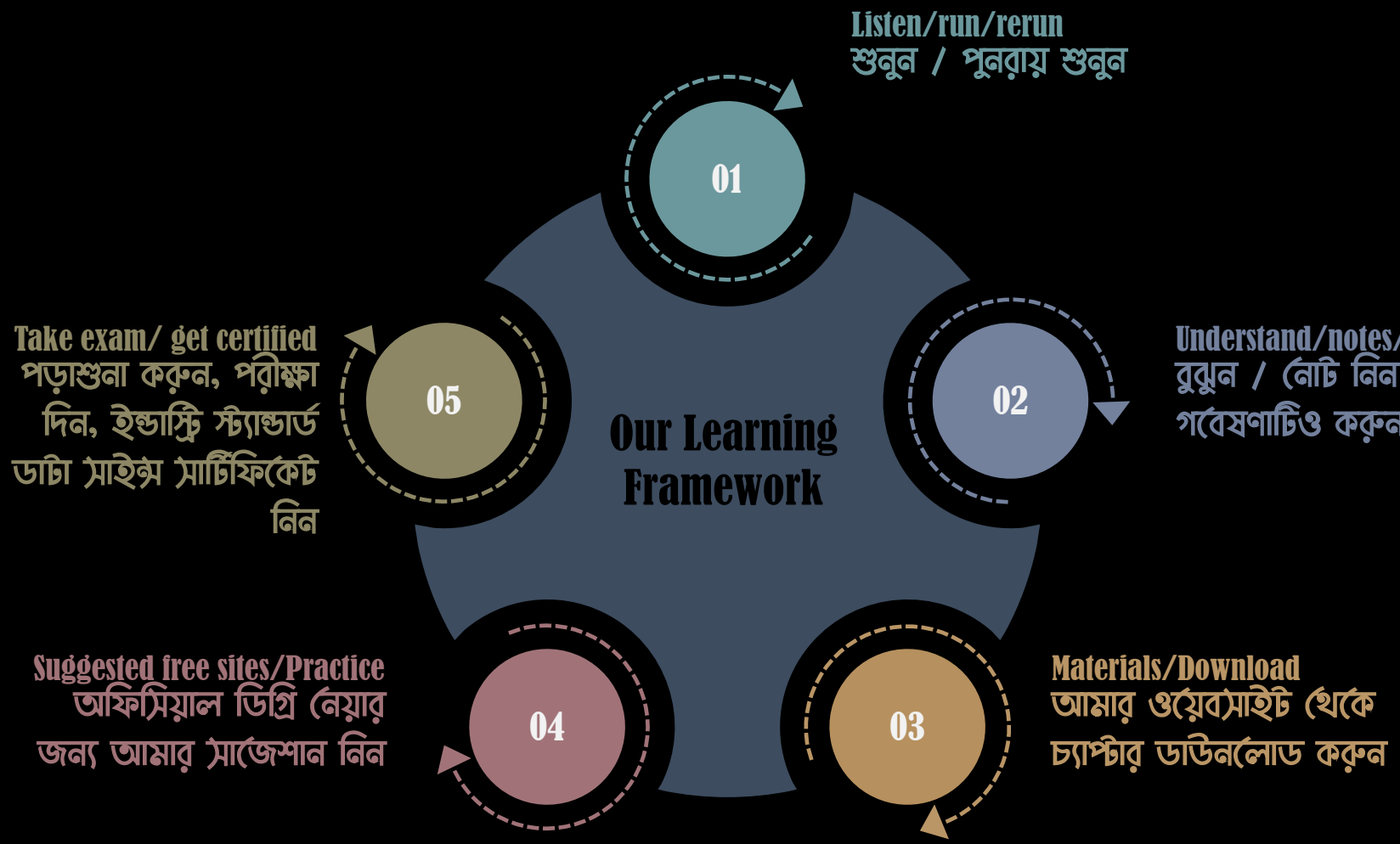
CHAPTER FOUR:
DATA SCIENCE APPLICATIONS

DATA SCIENCE FOUNDATION COUSE CHAPTER 4 (PART 11)

ENAMUL HAQUE

LEARNING FRAMEWORK

ডাটা সায়েন্স ফাউন্ডেশন কোর্স ।। ১১ তম পর্ব ।। ব্যাসিক্স অফ ডাটা সায়েন্স



CHAPTER 11 MAIN TOPICS

ডেটা সায়েন্সের জন্য বেসিক স্ট্যাটিস্টিক্যাল কনসেপ্ট

BASIC STATISTICAL CONCEPTS FOR DATA SCIENCE | **METHODS AND METRICS**

তথ্য/ বিশ্লেষণ / ডেটা পরিষ্কার বোঝা

UNDERSTANDING DATA ANALYSIS | **DATA CLEANUP**

ডেটা মাইনিং বনাম ডেটা সায়েন্স/ Self-Driving car ডেটা বিজ্ঞানেই ব্যবহার

DATA MINING VS DATA SCIENCE | **DATA SCIENCE IN SELF-DRIVING CARS**

BASIC STATISTICAL CONCEPTS FOR DATA SCIENCE - ডেটা সায়েন্সের জন্য প্রাথমিক পরিসংখ্যানগত ধারণা

Statistics - পরিসংখ্যান গণিতের একটি শাখা যা সংখ্যাসূচক তথ্য সংগ্রহ, বিশ্লেষণ, ব্যাখ্যা এবং উপস্থাপনের সাথে সম্পর্কিত। ডেটা বিজ্ঞানী রেবেকা ভিকারি ডেটা সায়েন্স শেখার সময় আপনার যেসব মৌলিক পরিসংখ্যান ধারণা বুঝতে হবে তা বর্ণনা করে। এগুলি বিশেষত উন্নত কৌশল নয়, তবে সেগুলি আরও জটিল পদ্ধতিগুলি শেখার আগে আপনার প্রয়োজনীয় মৌলিক প্রয়োজনীয়তাগুলি বেছে নিচ্ছে।

Statistical sample - পরিসংখ্যানগত নমুনা

Descriptive statistics - বর্ণনামূলক পরিসংখ্যান

Probability - সম্ভাব্যতা

Bias - পক্ষপাত

Compromise between preload and variance - প্রিলোড এবং বৈচিত্র্যের মধ্যে আপস

Correlation - পারস্পরিক সম্পর্ক

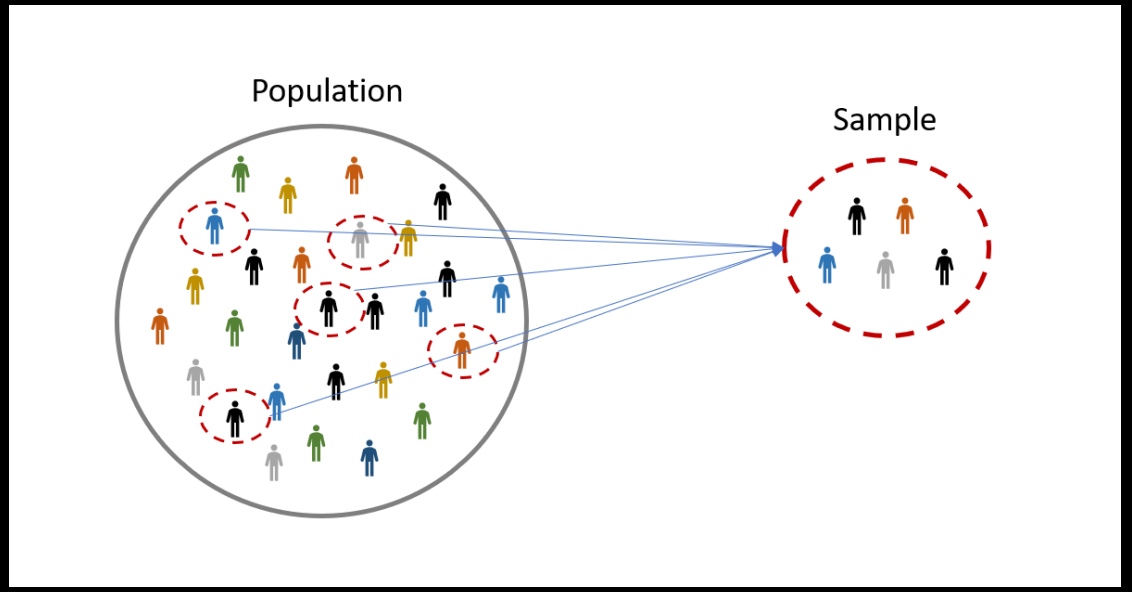
STATISTICAL SAMPLE - পরিসংখ্যানগত নমুনা

Statistics refer to the entire set of raw data that you might have available for a test or experiment as a population. For several reasons, you can't necessarily measure patterns and trends across the population. For this reason, we can use statistics to sample, perform some calculations for this data set, and, using the probability and some assumptions, to under-stand trends for the entire population or to predict future events.

Using statistics, we can take a sample of part of the population, perform some calculations for this data set, and use the probability and assumptions to understand trends with certainty that understand trends for the entire population.

Suppose, for example, that we want to understand the prevalence of a disease such as breast cancer in the whole population of the United Kingdom. For practical reasons, it is not possible to examine the entire population. Instead, we can take a random sample and measure the prevalence among them. Assuming that our sample is sufficiently randomised and representative of the whole population, we can estimate prevalence.

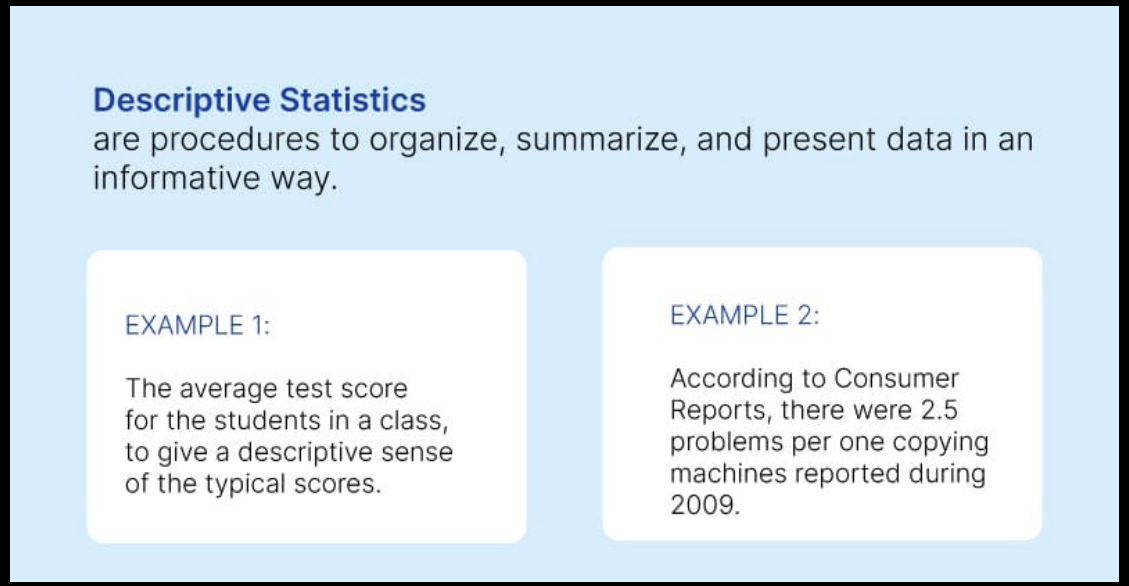
পরিসংখ্যান, মানের নিশ্চয়তা এবং জরিপ পদ্ধতিতে, নমুনা সম্পূর্ণ জনসংখ্যার বৈশিষ্ট্য অনুমান করার জন্য একটি পরিসংখ্যানগত জনগোষ্ঠীর মধ্যে থেকে ব্যক্তিদের একটি উপসেট (একটি পরিসংখ্যানের নমুনা) নির্বাচন করা হয়। পরিসংখ্যানবিদরা নমুনাগুলিকে প্রশ্নে জনগণের প্রতিনিধিত্ব করার চেষ্টা করেন। স্যাম্পলিংয়ের দুটি সুবিধা হ'ল সম্পূর্ণ জনসংখ্যার পরিমাপের চেয়ে কম ব্যয় এবং দ্রুত ডেটা সংগ্রহ। প্রতিটি পর্যবেক্ষণ স্বতন্ত্র বস্তু বা ব্যক্তি হিসাবে পৃথক পৃথক পর্যবেক্ষণকারী সংস্থার এক বা একাধিক বৈশিষ্ট্য (যেমন ওজন, অবস্থান, রঙ) পরিমাপ করে। জরিপের নমুনা গ্রহণের ক্ষেত্রে, নমুনা ডিজাইনের জন্য, বিশেষত স্তরযুক্ত নমুনার ক্ষেত্রে ওজন প্রয়োগ করা যেতে পারে। সম্ভাব্যতা তত্ত্ব এবং পরিসংখ্যানতত্ত্ব থেকে প্রাপ্ত ফলাফল অনুশীলনকে গাইড করার জন্য নিযুক্ত করা হয়। ব্যবসা এবং চিকিৎসা গবেষণায়, জনসংখ্যা সম্পর্কে তথ্য সংগ্রহের জন্য নমুনাটি ব্যাপকভাবে ব্যবহৃত হয়। গ্রহণের নমুনা নির্ধারণের জন্য ব্যবহৃত হয় যদি উৎপাদনের প্রচুর পরিমাণের উপাদান পরিচালনা সংক্রান্ত নির্দিষ্টকরণের সাথে মেলে কিনা। ধরুন, উদাহরণস্বরূপ, আমরা যুক্তরাজ্যের সমগ্র জনসংখ্যার মধ্যে স্তন ক্যান্সারের মতো একটি রোগের বিস্তার বুঝতে চাই। ব্যবহারিক কারণে, সমগ্র জনসংখ্যা পরীক্ষা করা সম্ভব নয়। পরিবর্তে, আমরা একটি এলোমেলো নমুনা নিতে পারি এবং তাদের মধ্যে ব্যাপকতা পরিমাপ করতে পারি। ধরে নিচ্ছি যে আমাদের নমুনা যথেষ্ট র‍্যান্ডমাইজড এবং পুরো জনসংখ্যার প্রতিনিধি, আমরা বিস্তারের অনুমান করতে পারি।



DESCRIPTIVE STATISTICS - বর্ণনামূলক পরিসংখ্যান

Descriptive statistics help us, as the name suggests, to describe the data. In other words, it allows us to understand the underlying characteristics. It does not predict anything, makes no assumptions or completes nothing. It just describes what the data sample we have looks like.
Descriptive statistics are derived from calculations, often referred to as parameters. These include things like:
Mean - the central value, commonly referred to as average.
Median - the average if we have ordered the data from low to high and divided precisely in half.
Mode - the most common value.
Descriptive statistics are helpful but can often hide important information about the record. For example, suppose a document contains multiple numbers that are much larger than the others. In that case, the mean may be distorted and does not accurately represent the data.
A distribution is a chart, often a histogram, that shows the number of times each value is displayed in a record. This type of chart gives us information about the distribution and skewness of the data.
One of the essential distributions is the normal distribution, usually referred to as a bell curve due to its shape. It has an asymmetric shape, with most values grouped around the central peak and the more distant values evenly distributed on each curve's side.

বর্ণনামূলক পরিসংখ্যান আমাদের সাহায্য করে, যেমন নাম প্রস্তাব করে, ডেটা বর্ণনা করতে। অন্য কথায়, এটি আমাদের অন্তর্নিহিত বৈশিষ্ট্যগুলি বুঝতে দেয়। এটি কোন কিছুরই পূর্বাভাস দেয় না, কোন অনুমান করে না বা কোন কিছুই সম্পূর্ণ করে না। এটি কেবল বর্ণনা করে যে আমাদের কাছে থাকা ডেটা নমুনা কেমন দেখাচ্ছে।
বর্ণনামূলক পরিসংখ্যান গণনা থেকে উদ্ভূত হয়, যা প্রায়ই প্যারামিটার হিসাবে উল্লেখ করা হয়। এর মধ্যে বিষয়গুলি অন্তর্ভুক্ত রয়েছে:
মানে - কেন্দ্রীয় মান, সাধারণত গড় হিসাবে উল্লেখ করা হয়।
মধ্যমা - যদি আমরা কম থেকে উচ্চ পর্যন্ত ডেটা অর্ডার করি এবং সঠিকভাবে অর্ধেক ভাগ করি।
মোড - সবচেয়ে সাধারণ মান।



PROBABILITY - সম্ভাব্যতা

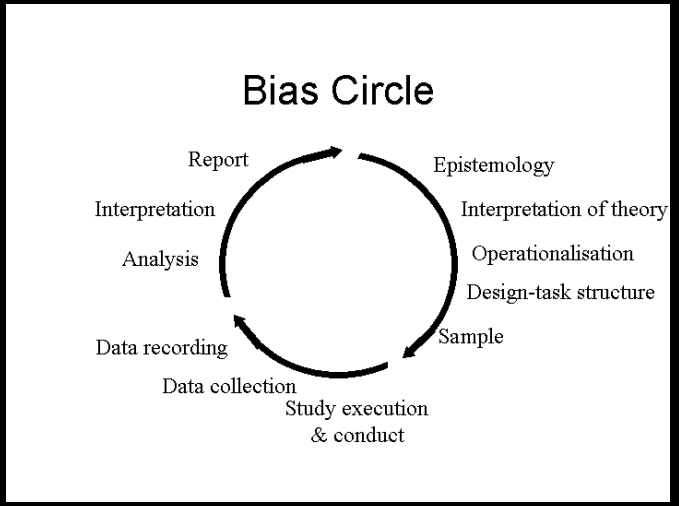
The probability, in simple terms, is the likelihood of an event occurring. In statistics, an event results from an experiment that can be something like dice or an AB test result.
The probability for a single event is calculated by dividing the number of events by the total number of possible results. For example, if you throw a six on a dice, there are 6 possible results. So, the chance of dicing a six is 1/6 = 0.167; sometimes, it is expressed as a percentage, i.e., 16.7%.
Events can be either independent or dependent. For dependent events, a previous event affects the subsequent event. Suppose we have a bag of M & Ms and wanted to determine the probability that a red M & M will be selected at random. Each time we remove the selected M & M from the bag, the likelihood of picking red changes due to previous events' effects.
Independent events are not affected by previous events. In the M & M bag case, we put it back in the bag every time we select one. The probability of choosing red remains the same each time.
Whether an event is independent or not is important because we calculate the probability of multiple events changes depending on the type.
The probability of multiple independent events is calculated by simply multiplying the probability of each event. Suppose we wanted to calculate the probability of dicing 6 three times in the example of the dice's roll. This would look like this:
1/6 = 0.167 1/6 = 0.167 1/6 = 0.167
0.167 * 0.167 * 0.167 = 0.005
The calculation is different for dependent events, also known as conditional probability. If we take the example of M & M, let's imagine we have a bag with only two colours, red and yellow, and we know that the pack contains 3 red and 2 yellow, and we want to calculate the probability of selecting two red ones in a row. In the first selection, the probability of making a red selection is 3/5 = 0.6. We removed an M & M that was randomly red in the second selection, so our second probability calculation is 2/4 = 0.5. Therefore, the probability of picking two reds in a row is 0.6 * 0.5 = 0.3.

সম্ভাব্যতা, সহজ ভাষায়, একটি ঘটনা ঘটার সম্ভাবনা। পরিসংখ্যানগুলিতে, একটি ইভেন্ট একটি পরীক্ষা থেকে আসে যা পাশা বা এবি পরীক্ষার ফলাফল হতে পারে। একটি ইভেন্টের সম্ভাব্যতা সম্ভাব্য ফলাফলের মোট সংখ্যা দ্বারা ইভেন্টের সংখ্যাকে ভাগ করে গণনা করা হয়। উদাহরণস্বরূপ, যদি আপনি একটি পাশা উপর একটি ছক্কা নিক্ষেপ, 6 সম্ভাব্য ফলাফল আছে। সুতরাং, ছক্কা মারার সুযোগ হল 1/6 = 0.167; কখনও কখনও, এটি শতাংশ হিসাবে প্রকাশ করা হয়, অর্থাৎ, 16.7%। ঘটনা স্বাধীন বা নির্ভরশীল হতে পারে। নির্ভরশীল ঘটনার জন্য, একটি পূর্ববর্তী ঘটনা পরবর্তী ঘটনাকে প্রভাবিত করে। ধরুন আমাদের কাছে M & Ms এর একটি ব্যাগ আছে এবং একটি লাল M & M এলোমেলোভাবে নির্বাচন করা হবে এমন সম্ভাবনা নির্ধারণ করতে চেয়েছিল। প্রতিবার যখন আমরা ব্যাগ থেকে নির্বাচিত M & M সরিয়ে ফেলি, পূর্ববর্তী ইভেন্টের প্রভাবের কারণে লাল পরিবর্তন বাছার সম্ভাবনা।

BIAS - পক্ষপাত

As explained in the statistics, we often use data samples to estimate the entire data set. Similarly, we will use some training data for predictive modelling and create a model that can make predictions about new data.
Bias is the tendency of a statistical or predictive model to underestimates a parameter. This is often due to the method of obtaining a sample or the way errors are measured. There are different types of distortions in statistics. Here is a brief description of two of them.
Selection Distortion - This occurs when the sample is not randomly selected. An example of data science can be to stop an AB test prematurely when the test runs or choose data to train a machine learning model from a specific period of time, masking seasonal effects.
Confirmation Distortion - This occurs when the person performing an analysis has a predetermined assumption about the data. In this situation, there may be a tendency to spend more time studying variables that are likely to support this assumption.
As explained earlier, the mean in a data sample is the central value. The variance measures how far each value in the record is from the mean. Essentially, it is a measurement of the variation of numbers in a data set.
The standard deviation is a common measure of the variation of da-ta with the normal distribution. It is a calculation that specifies a value that indicates how far the values are distributed. A low standard deviation indicates that the values tend to be reasonably close to the mean, while a high standard deviation indicates that the values are more distributed.
If the data does not follow a normal distribution, other variance measures are used. The interquartile range is usually used. This measurement is derived by first dividing the values by rank and then dividing the data points into four equal parts, called quartiles. Each quartile describes where 25% of the data points are according to the median. The interquartile range is calculated by subtracting the median for the two central quarters, also known as Q1 and Q3.

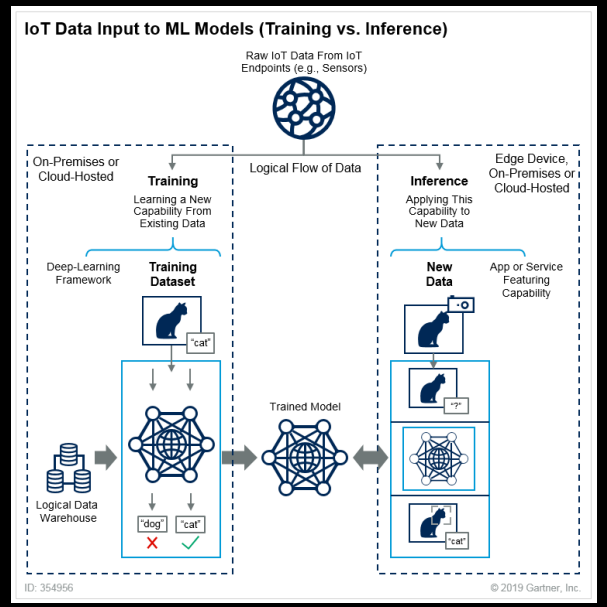
পক্ষপাত - হচ্ছে কোন তথ্যকে এমনভাবে অনুসন্ধান করা, ব্যাখ্যা করা, মনে করা বা তার পক্ষ নেয়া যাতে ব্যক্তির মধ্যে থাকা পূর্বের বিশ্বাস বা তার পূর্বের কোন অনুকল্পকে নিশ্চিত করা হয়।
বায়াস হল একটি পরিসংখ্যান বা ভবিষ্যদ্বাণীমূলক মডেলের প্রবণতা যা একটি প্যারামিটারকে অবমূল্যায়ন করে। এটি প্রায়শই একটি নমুনা পাওয়ার পদ্ধতি বা ত্রুটিগুলি পরিমাপের কারণে হয়। পরিসংখ্যানগুলিতে বিভিন্ন ধরণের বিকৃতি রয়েছে। তাদের দুটির সংক্ষিপ্ত বিবরণ এখানে।
নির্বাচন বিকৃতি- এটি ঘটে যখন নমুনা এলোমেলোভাবে নির্বাচিত হয় না। ডেটা সায়েন্সের একটি উদাহরণ হতে পারে একটি AB পরীক্ষা অকালে বন্ধ করা যখন পরীক্ষা চলবে বা নির্দিষ্ট সময় থেকে মেশিন লার্নিং মডেলকে প্রশিক্ষণের জন্য ডেটা বেছে নেবে।
নিশ্চিতকরণ বিকৃতি- এটি ঘটে যখন বিশ্লেষণ করা ব্যক্তির ডেটা সম্পর্কে পূর্বনির্ধারিত অনুমান থাকে। এই পরিস্থিতিতে, ভেরিয়েবলগুলি অধ্যয়ন করার জন্য আরও বেশি সময় ব্যয় করার প্রবণতা থাকতে পারে যা এই ধারণাটিকে সমর্থন করার সম্ভাবনা রয়েছে।

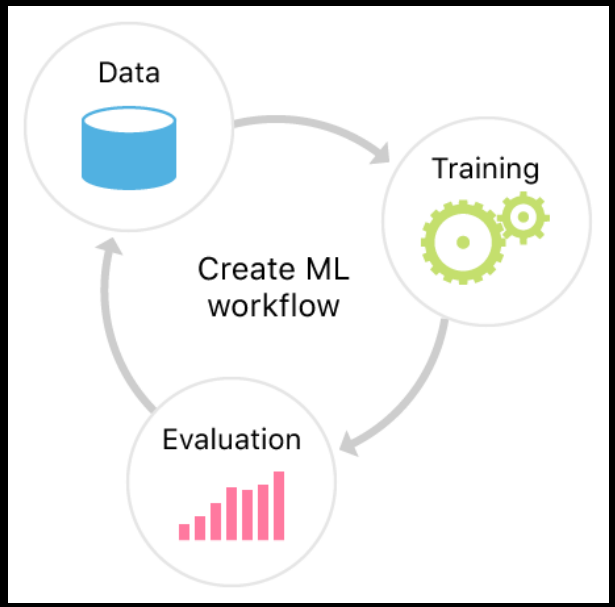


COMPROMISE BETWEEN PRELOAD AND VARIANCE - প্রিলোড এবং বৈচিত্র্যের মধ্যে আপস

The concepts of bias and variance are essential for machine learning. When we create a machine learning model, we use a sample of data called a training record. The model learns patterns in this data and generates a mathematical function that can be used to associate the correct target label or target value (y) with a series of inputs (X).
When generating this mapping function, the model uses several assumptions to approximate the target. For example, the linear regression algorithm assumes a linear relationship (straight line) between the input and the target. These assumptions distort the model.
The variance is the difference between the mean prediction generated by the model and the actual value in the calculation.
If we were to train a model using different training data samples, we would vary the returned predictions. The variance in machine learning is a measure of how big this difference is.
In machine learning, bias and variance are the overall expected flaw for our predictions. In an ideal world, we would have both low distortion and low friction. In practice, however, minimising the preload usually leads to an increase in variance and vice versa. The bias/variance compromise describes the process of compensating these two errors to reduce the overall error for a model.

মেশিন লার্নিং এর জন্য পক্ষপাত এবং বৈকল্পিকতার ধারণা অপরিহার্য। যখন আমরা একটি মেশিন লার্নিং মডেল তৈরি করি, তখন আমরা একটি প্রশিক্ষণ রেকর্ড নামক তথ্যের নমুনা ব্যবহার করি। মডেল এই ডেটার মধ্যে প্যাটার্ন শিখে এবং একটি গাণিতিক ফাংশন তৈরি করে যা সঠিক টার্গেট লেবেল বা টার্গেট ভ্যালু (y) কে ইনপুট সিরিজ (x) এর সাথে যুক্ত করতে ব্যবহার করা যেতে পারে।
এই ম্যাপিং ফাংশনটি তৈরি করার সময়, লক্ষ্যটি আরও ভালভাবে অনুমান করতে মডেলটি বেশ কয়েকটি অনুমান ব্যবহার করে। উদাহরণস্বরূপ, লিনিয়ার রিগ্রেশন অ্যালগরিদম ইনপুট এবং টার্গেটের মধ্যে একটি রৈখিক সম্পর্ক (সরলরেখা) ধরে নেয়। এই অনুমানগুলি মডেলকে বিকৃত করে।

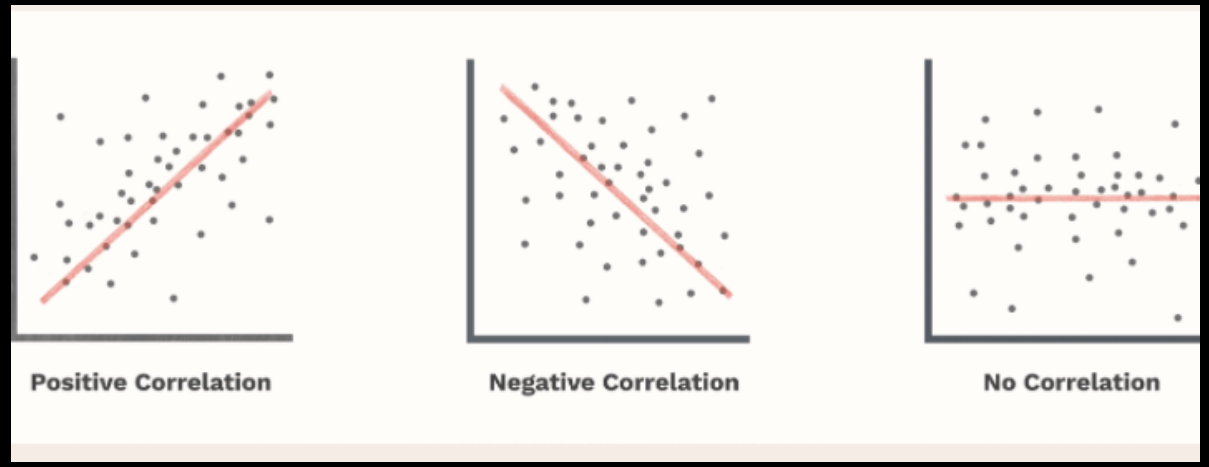




CORRELATION - পারস্পরিক সম্পর্ক

Correlation is a statistical technique used to measure relationships between two variables. The correlation is assumed to be linear (it forms a line when displayed in a chart) and is expressed as a number between +1 and -1. This is called a correlation coefficient.
A correlation coefficient of +1 denotes an entirely positive correlation (if the value for one variable also increases the value of the second variable), a coefficient of 0 does not mean correlation, and a coefficient of -1 denotes an entirely negative correlation.
Statistics is a wide and complex field. This article is intended as a brief introduction to some of the most commonly used statistical techniques in data science. Data science courses often require prior knowledge of these basic concepts or start with descriptions that are too complex and difficult to understand. I hope this article will serve as a re-fresher for a selection of basic statistical techniques used in data science before going into more advanced topics.

পারস্পরিক সম্পর্ক একটি পরিসংখ্যান কৌশল যা দুটি ভেরিয়েবলের মধ্যে সম্পর্ক পরিমাপ করতে ব্যবহৃত হয়। পারস্পরিক সম্পর্কটি রৈখিক বলে ধরে নেওয়া হয় (এটি একটি চার্টে প্রদর্শিত হলে এটি একটি লাইন গঠন করে) এবং +1 এবং -1 এর মধ্যে একটি সংখ্যা হিসাবে প্রকাশ করা হয়। একে বলা হয় পারস্পরিক সহগ।



THANK YOU AND NEXT WE LEARN

- METHODS AND METRICS