

Open in app ↗



Search

Write



★ Member-only story

DBSCAN: Intution, Advantages, and Points to Remember



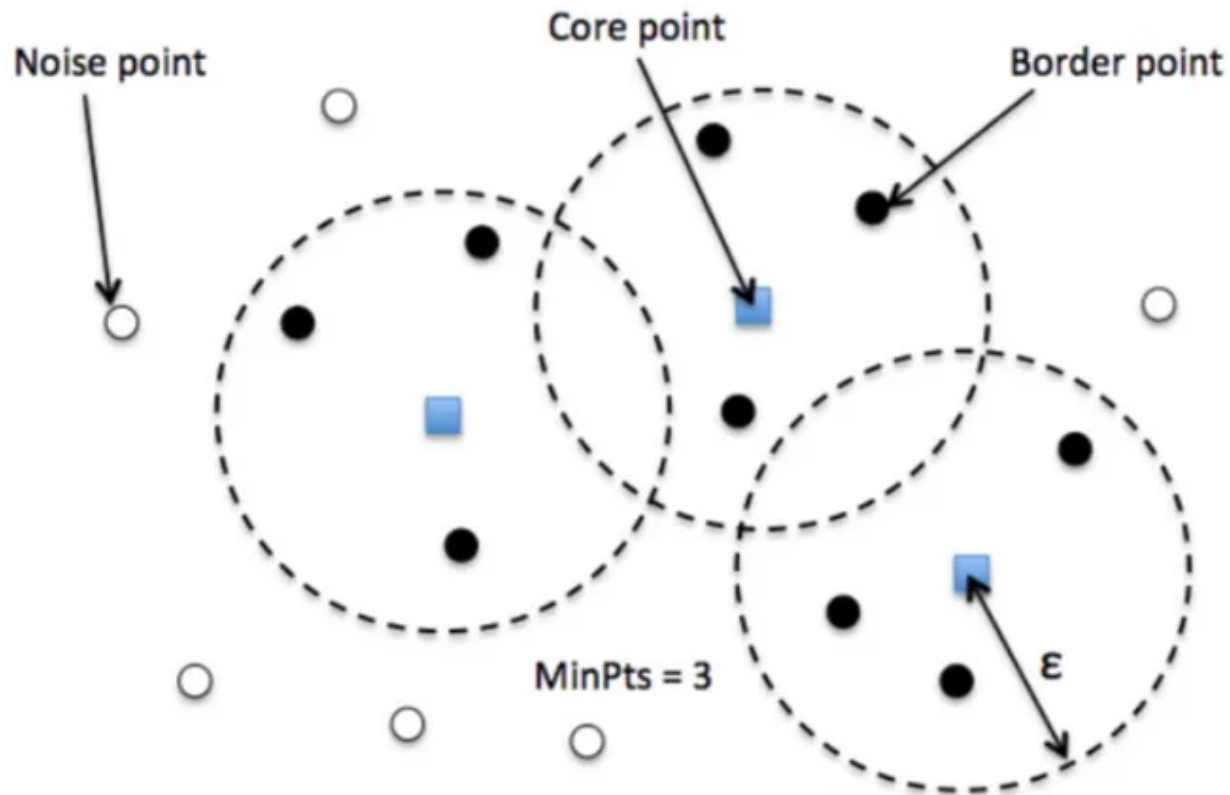
Rahul S · Following

3 min read · Sep 15



3





- **Core** — This is a point that has at least m points within distance n from itself.
- **Border** — This is a point that has at least one Core point at a distance n .
- **Noise** — This is a point that is neither a Core nor a Border. And it has less than m points within distance n from itself.

1. INTUITION

Imagine you have a field of stars in the night sky, and you want to group them based on how densely they are packed together rather than a predetermined number of clusters. This is where DBSCAN, which stands for *Density-Based Spatial Clustering of Applications with Noise*, shines like a cosmic beacon.

DBSCAN is a remarkable clustering algorithm that doesn't rely on predefining the number of clusters, making it particularly well-suited for finding clusters of varying shapes and sizes in your data.

2. WORKING

Here's how DBSCAN works:

1. **Density-Centered Clustering:** DBSCAN identifies clusters by looking at the density of data points. It defines a cluster as a dense region of data points that is separated by areas of lower point density.
2. **Core Points:** The algorithm starts by selecting a random data point and examines its neighborhood within a specified radius (epsilon, ϵ). If there are at least a minimum number of data points (minPts) within this neighborhood, it marks the central point as a "core point."
3. **Growing Clusters:** DBSCAN then expands the cluster around this core point by recursively adding nearby points that are also core points. This process continues until no more core points can be added.
4. **Border Points:** Any data points that are within the neighborhood of a core point but don't meet the density criteria to be core points themselves are considered "border points" and are assigned to the nearest cluster.
5. **Noise:** Data points that are not core points or border points and are not within the neighborhood of any core points are treated as noise and do not belong to any cluster.

3. ADVANTAGES

Now, let's explore some of the benefits of DBSCAN over other clustering algorithms:

1. **No Predefined Number of Clusters:** One of the most significant advantages of DBSCAN is that you don't need to specify the number of clusters beforehand. It adapts to the density and distribution of data, making it useful when you have no prior knowledge of the dataset's structure.

2. Robust to Outliers: DBSCAN is robust to noise and outliers since they are treated as noise points and don't affect the formation of clusters. This makes it effective for real-world data, which often contains noisy observations.

3. Handles Complex Shapes: Unlike K-Means, which assumes clusters are spherical, DBSCAN can identify clusters of varying shapes and sizes. It is excellent for datasets with irregularly shaped clusters.

4. No Need for Distance Metric Calibration: Unlike hierarchical clustering methods that require careful selection of distance metrics, DBSCAN uses a single parameter (ϵ) to define the neighborhood, which is often more intuitive and less sensitive to metric choices.

4. POINTS TO REMEMBER

However, there are important points to keep in mind when using DBSCAN:

1. Parameter Selection: While DBSCAN is robust, choosing appropriate values for ϵ and minPts can impact the results. These parameters may need to be adjusted based on the specific characteristics of your data.

2. Scalability: DBSCAN's computational complexity can be an issue for large datasets, especially if you use a naive implementation. Consider using optimized versions or subsampling if scalability is a concern.

3. Sensitivity to Density: DBSCAN's effectiveness relies on the assumption that clusters have different densities. If your data consists of clusters with similar densities, DBSCAN might not perform as well.

In conclusion, DBSCAN is a powerful clustering algorithm that excels at finding clusters in data with varying shapes and sizes, handling noise and

outliers gracefully, and not requiring you to specify the number of clusters in advance. However, parameter selection and scalability should be carefully considered for optimal results. Think of DBSCAN as your guiding star when exploring data with unknown structures, allowing you to uncover hidden patterns without the need for predefined expectations.

Machine Learning

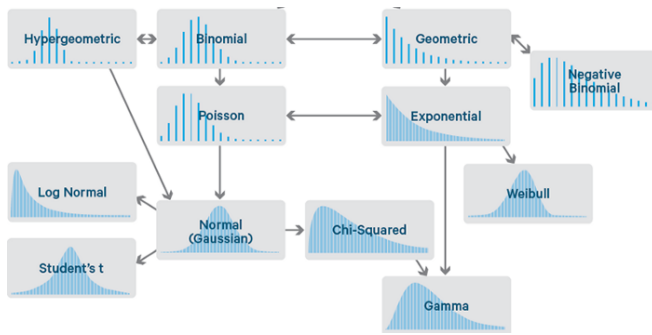
Clustering

**Written by Rahul S**

673 Followers

Following

[linkedin.com/in/rahultheogre](https://www.linkedin.com/in/rahultheogre) | NLP, Statistics, ML**More from Rahul S**



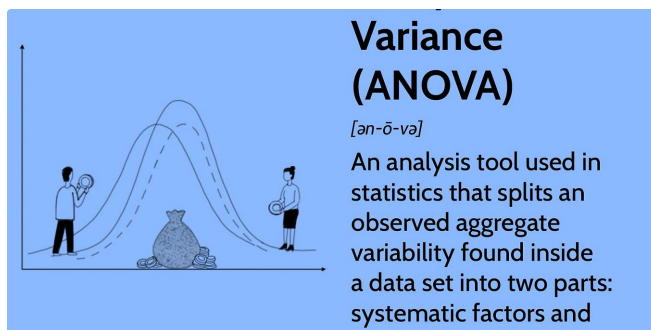
Rahul S

Statistics: Probability Distributions (An Overview)

Random Variable: All possible outcomes of a random experiment are random variables. A...

🌟 · 6 min read · Sep 5

30 1



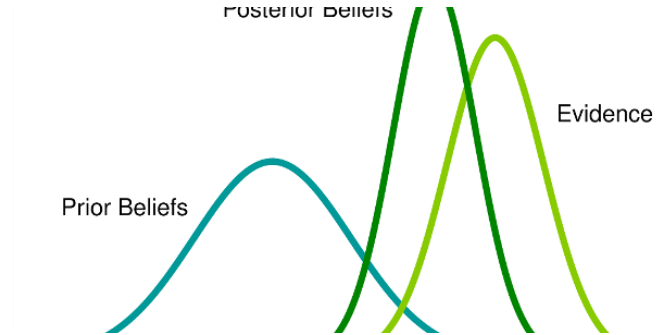
Rahul S

Statistics: ANOVA- An Intuition with Example

Without going into mathematical intricacies

🌟 · 7 min read · May 14

39



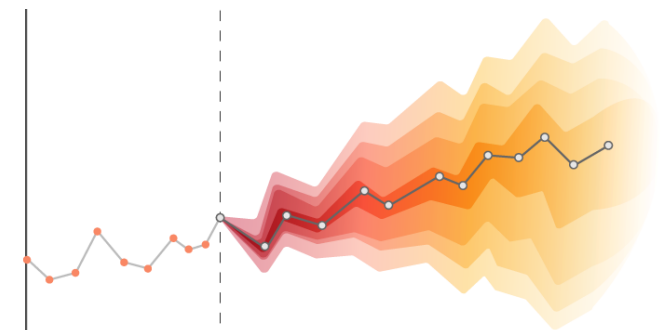
Rahul S

Bayesian Statistics for Kids—An Introduction

1. INTRODUCTION

🌟 · 3 min read · May 11

17 1



Rahul S

Time Series Forecasting: A Comparative Analysis of SARIMA...

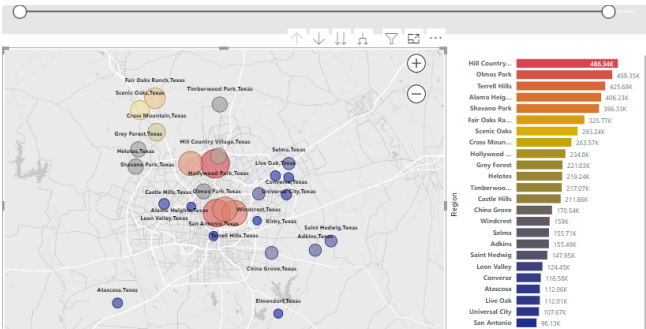
SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous...

🌟 · 3 min read · May 14

117

See all from Rahul S

Recommended from Medium



 Rayan Yassminh www.linkedin.com/in/rayan-yassminh/

Clustering Time Series : Analysis of Housing Prices in Bexar County, T...

Clustering time series is a powerful technique utilized to group similar time series data...

9 min read · May 6


 17










 Deniz Gunay

Clustering

Clustering

20 min read · Sep 18

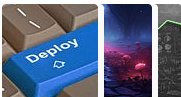
 28







Lists



Predictive Modeling w/ Python

20 stories · 482 saves



Practical Guides to Machine Learning

10 stories · 554 saves



Natural Language Processing

698 stories · 309 saves



The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 141 saves



Yannis Poulakis

Is Silhouette the Right Clustering Evaluation Metric for You?

While popular, simple experiments prove that Silhouette index may not be the index to...

4 min read · Jul 4



17



misun_song

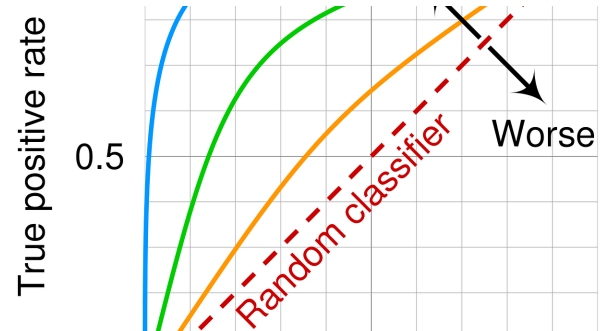
Understanding the ROC-AUC Curve

Evaluating Classification Model Performance Simply

6 min read · Sep 23



62



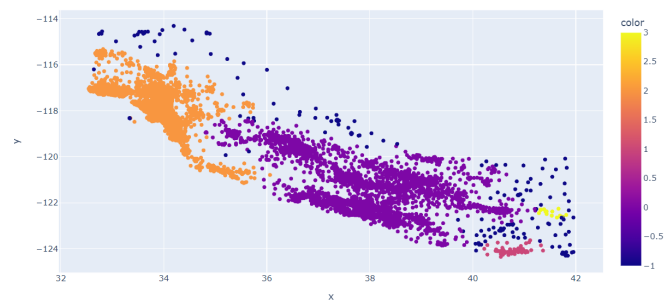
Description	Quantity	InvoiceD
3 HEART T-LIGHT HOLDER	6	2010-12-01 08:26
WHITE METAL LANTERN	6	2010-12-01 08:26
D HEARTS COAT HANGER	8	2010-12-01 08:26
FLAG HOT WATER BOTTLE	6	2010-12-01 08:26
LY HOTTIE WHITE HEART.	6	2010-12-01 08:26



nurazizah vidya

Customer Segmentation using clustering

Customer Segmentation is the subdivision of a market into discrete customer groups that...



Revag

DBSCAN—an Easy Clustering Algorithm and also how to optimi...

DBSCAN stands for “Density-Based Spatial Clustering of Applications with Noise.”


5 min read · Jun 8

9 min read · Aug 9

 60 

 24  1

See more recommendations