Why we have to remove highly correlated features in Machine Learning?

Sujatha Mudadla · Follow

2 min read · Nov 28, 2023

In machine learning, highly correlated features refer to variables that have a strong linear relationship with each other. The presence of highly correlated features can lead to several issues, and removing them is often beneficial. Let's delve into the reasons behind removing highly correlated features:

1. Redundancy:

- Explanation: Highly correlated features essentially provide the same information about the target variable. When two features are strongly correlated, one of them becomes redundant as it doesn't add new or unique information to the model.
- Impact: Redundant features do not contribute significantly to the model's predictive power but can increase the complexity of the model without

adding value.

2. Multicollinearity:

- Explanation: In linear regression and other linear models, multicollinearity arises when two or more features are highly correlated. This can lead to issues in estimating the individual coefficients of the correlated features because their effects on the target variable become difficult to distinguish.
- Impact: Multicollinearity can result in unstable and inaccurate coefficient estimates. The model becomes sensitive to small changes in the data, making it less reliable.

3. Computational Efficiency:

- Explanation: Models with fewer features are computationally less intensive and faster to train. Removing highly correlated features reduces the dimensionality of the dataset, making it more efficient to process.
- Impact: Faster training times are especially crucial in scenarios where large datasets or real-time processing are involved.

4. Model Interpretability:

- Explanation: Highly correlated features can make it challenging to interpret the importance of individual features. Feature importance becomes ambiguous when two or more features convey similar information.
- Impact: Interpretable models are essential for understanding the factors influencing predictions. Removing correlated features aids in clearer interpretation and better decision-making.

5. Enhancing Model Generalization:

- Explanation: Models trained on datasets with highly correlated features may perform well on the training data but struggle to generalize to new, unseen data.
- Impact: Removing correlated features helps improve a model's ability to generalize to new data, reducing the risk of overfitting and making the model more robust.

6. Improved Model Performance:

- Explanation: Correlated features can lead to a phenomenon known as the "curse of dimensionality," where the model's performance degrades as the number of features increases.
- Impact: Removing redundant features can lead to simpler models that are less prone to overfitting and perform better on unseen data.

7. Stability of Feature Importance:

- Explanation: The importance of features in a model can be unstable when features are highly correlated. Small changes in the dataset may lead to significant variations in feature importance.
- Impact: Removing correlated features enhances the stability of feature importance rankings, making the model more reliable.

In summary, removing highly correlated features is a crucial preprocessing step in machine learning to improve model performance, interpretability, generalization, and computational efficiency. It helps create more robust and efficient models that are better suited for real-world applications.

Written by Sujatha Mudadla

2K Followers · 1 Following

Follow

M.Tech(Computer Science),B.Tech (Computer Science) I scored GATE in Computer Science with 96 percentile.Mobile Developer and Data Scientist.

No responses yet

John Bennett

What are your thoughts?

More from Sujatha Mudadla

Sujatha Mudadla

Sujatha Mudadla

Interview questions on Android Jetpack Compose.

What is Android Jetpack Compose, and why was it introduced?

Jun 8, 2023 450 4

Android Kotlin Coroutines interview questions

1. Q: What are Kotlin Coroutines, and how are they different from traditional threading?

Dec 19, 2023 751 3

Sujatha Mudadla

AWS Data Engineering Interview Questions.

Question: What is Amazon S3, and how is it commonly used in data engineering?

Jan 11, 2024 308 3

Sujatha Mudadla

What is the difference between Map and FlatMap in Spark

In Apache Spark, both map and flatMap are transformation operations that can be applie...

Sep 12, 2023 39 2

See all from Sujatha Mudadla

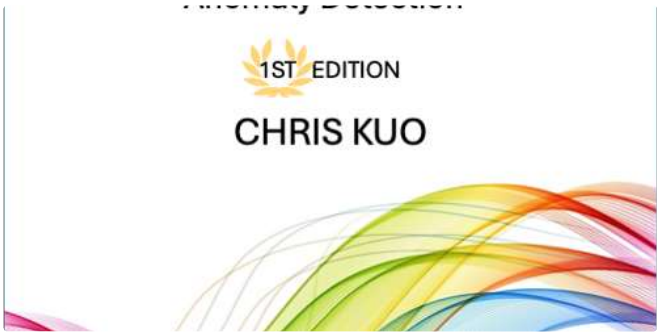
Recommended from Medium

JABERI Mohamed Habib

Feature Selection Techniques in Machine Learning

Feature selection is a critical step in the data preprocessing phase of machine learning. It...

Sep 27, 2024



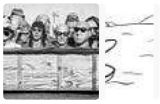
In Dataman in AI by Chris Kuo/Dr. Dataman

Tree-based XGB, LightGBM, and CatBoost Models for Multi-period...

Sample eBook chapters (free): https://github.com/dataman-git/modern-...

Apr 6, 2024 162

Lists



Staff picks

819 stories · 1637 saves

Stories to Help You Level-Up at Work

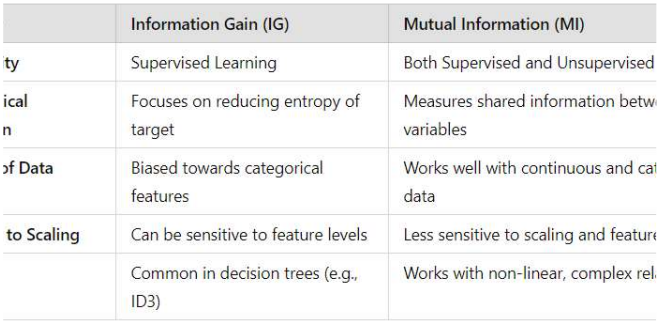
19 stories · 944 saves

Self-Improvement 101

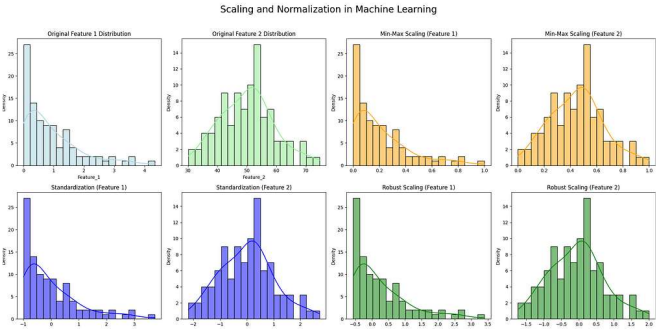
20 stories · 3328 saves

Productivity 101

20 stories · 2797 saves



	Information Gain (IG)	Mutual Information (MI)
ty	Supervised Learning	Both Supervised and Unsupervised
ical n	Focuses on reducing entropy of target	Measures shared information betw variables
of Data	Biased towards categorical features	Works well with continuous and cat data
to Scaling	Can be sensitive to feature levels	Less sensitive to scaling and feature
	Common in decision trees (e.g., ID3)	Works with non-linear, complex rel



Ebrahim Mousavi

In Biased-Algorithms by Amit Yadav

Information Gain and Mutual Information for Machine Learning

"In machine learning, the art of making decisions is as much about what you remove...

Sep 28, 2024 1

ML Series: Day 47—Scaling and Normalization

Balancing Data Magnitudes for Model Stability

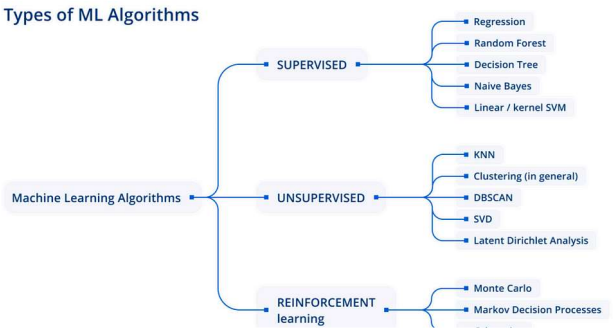
Feb 9 70

Emmanuel Ikogho

Data Science is dying; here's why

Why 85% of data science projects fail

Sep 3, 2024 2.2K 102



John Vastola

10 Must-Know Machine Learning Algorithms for Data Scientists

Machine learning is the science of getting computers to act without being explicitly...

Dec 6, 2022 1.4K 26

See more recommendations