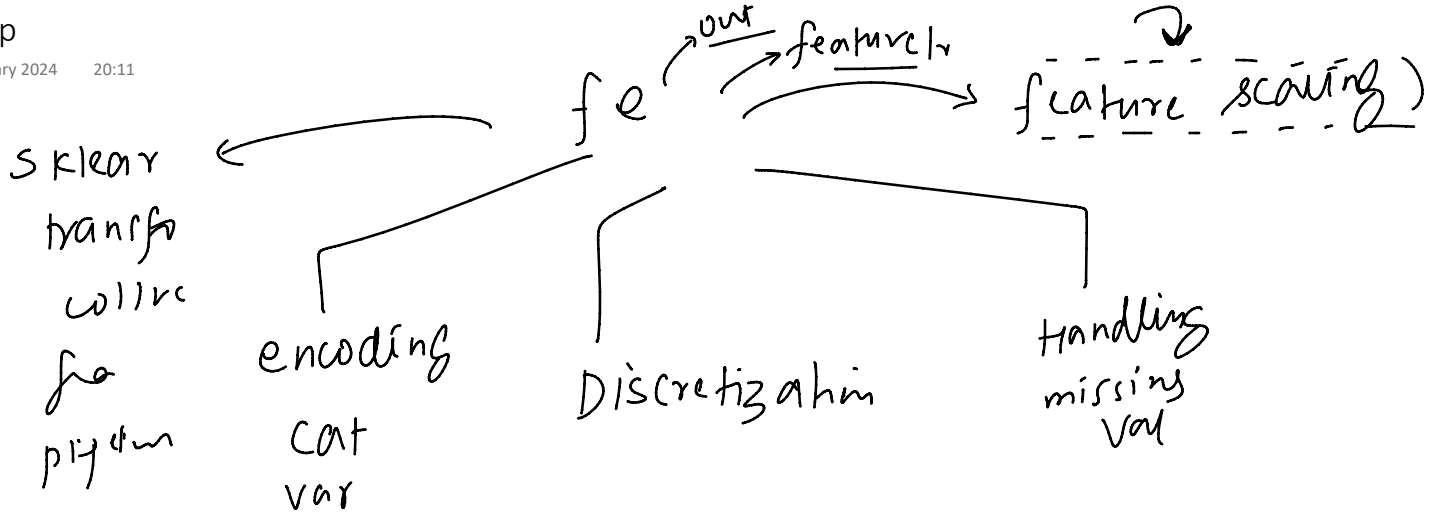
Recap

29 February 2024 20:11



What is Feature Scaling

29 February 2024 14:38

Feature scaling is a method used in data pre-processing for machine learning and data science that involves scaling the range of variables or features of data. The goal of feature scaling is to normalize the range of independent variables or features of data within a specific range, such as between -1 and 1 or 0 and 1, or to standardize the features to have a mean of 0 and a standard deviation of 1. This process is important because features on different scales can distort the distance between data points in some algorithms and can also affect the performance of algorithms negatively.

Why do we need feature scaling?

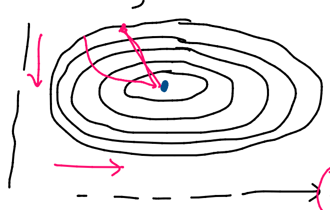
1. Improves the performance of Distance based algorithms
2. Improves in optimization techniques
3. [redacted] → deep learning

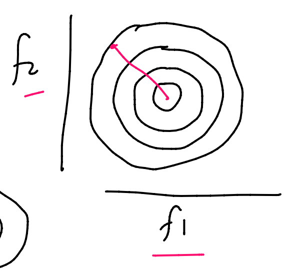
Which algorithms are affected if the features are not scaled?

1. Gradient Descent based algorithms
   a. Linear Regression
   b. Logistic Regression
   c. Neural Network
2. SVM
3. Distance based algorithms
   a. KNN
   b. K-Means
4. Dimensionality Reduction Techniques
   a. PCA
   b. LDA
5. Regularization Techniques
   a. Ridge Regression
   b. Lasso Regression

Which algorithms are not affected?

1. Tree Based Algorithms
   a. CART
   b. Random Forest
   c. Adaboost
   d. Gradient Boosting
   e. XGBoost
2. Naïve Bayes





Types of Feature Scaling Techniques

$$w_2 = w_2 - \eta \frac{\partial L}{\partial w_2}$$

Vishnu

1. Standardization → scaling

29 February 2024 15:06

Description: Scales features to have zero mean and unit variance, $z = \frac{(x-\mu)}{\sigma}$ where μ is the mean and σ is the standard deviation.

$x_i' = \frac{x_i - \mu}{\sigma}$ — $(x-\mu)$ → mean centering — $\frac{1}{\sigma}$ → scaling

100 users: age, $\mu = 30$, $\sigma = 10$; 35, 45, 55, 39, ...

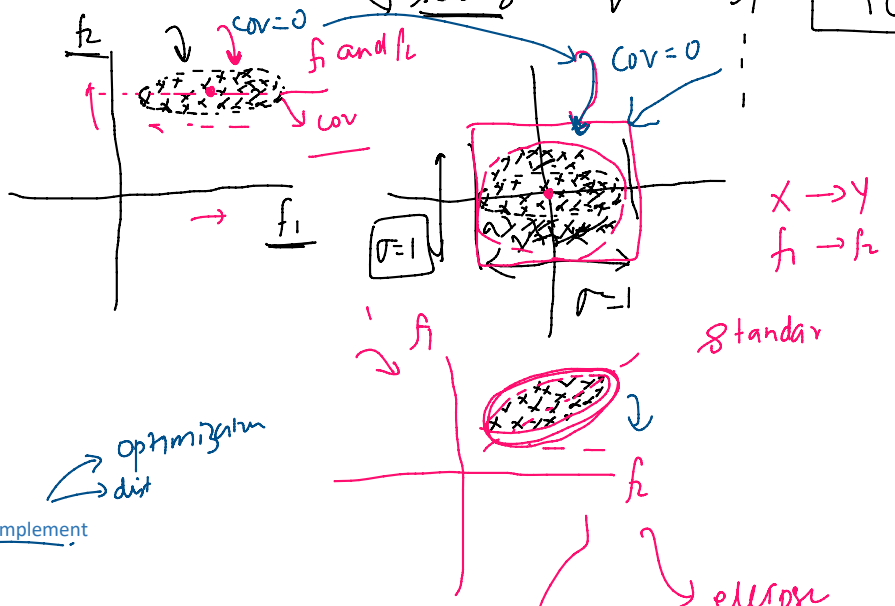
Standardization: $\frac{35-30}{10} = \frac{5}{10} = 0.5$; $\frac{45-30}{10} = 1.5$

age
0.5
1.5
2.5
...

mean = 0, $\sigma = 1$

Geometric Intuition:

f_1	f_2
—	—
—	—
—	—
—	—



Advantages:

1. Simple and efficient to implement

Disadvantages:

1. Does not work with algos that assume non negative values.

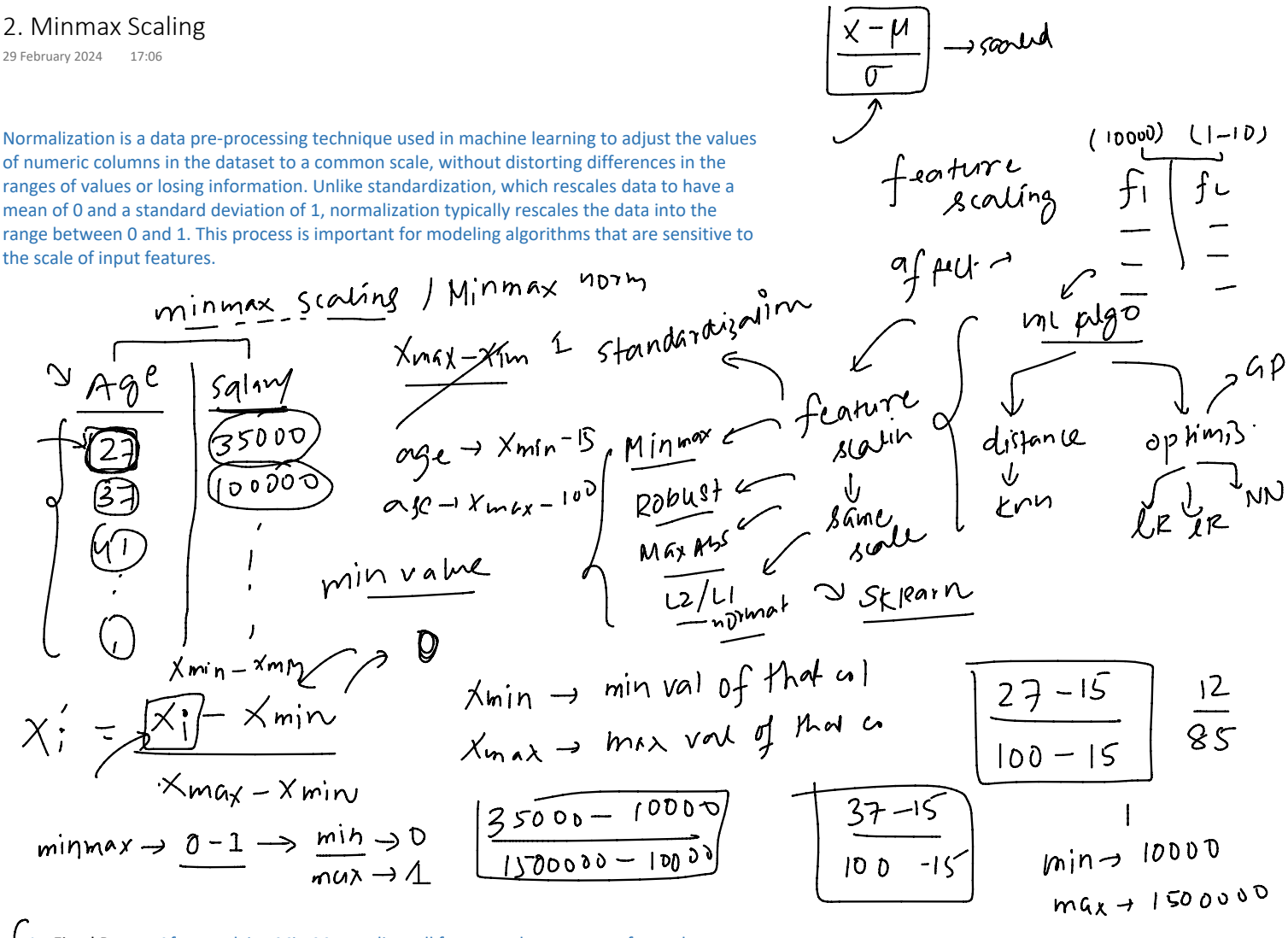
algo → negative → transformation → Normalization

multinomial naive bayes — dup → x dim

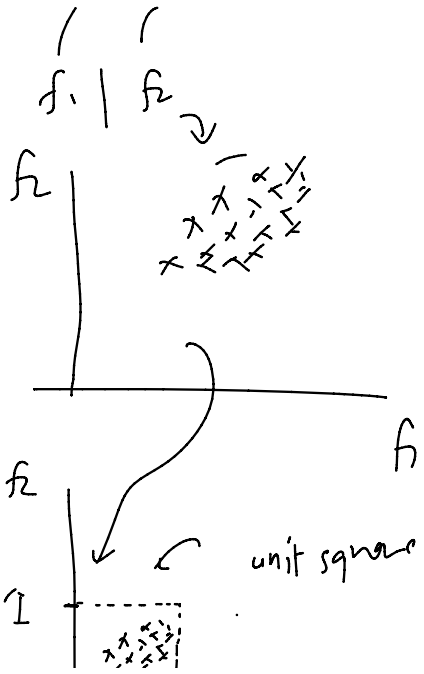
2. Minmax Scaling

29 February 2024 17:06

Normalization is a data pre-processing technique used in machine learning to adjust the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values or losing information. Unlike standardization, which rescales data to have a mean of 0 and a standard deviation of 1, normalization typically rescales the data into the range between 0 and 1. This process is important for modeling algorithms that are sensitive to the scale of input features.

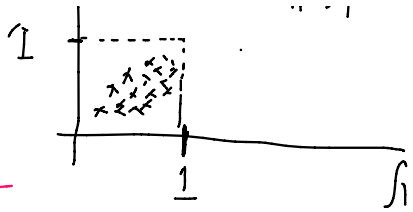


1. Fixed Range: After applying Min-Max scaling, all feature values are transformed to lie within a predefined range, typically [0, 1]. If X is a feature, the transformed value 'X'will be in the range where the minimum value of the feature becomes 0, and the maximum value becomes 1.
2. Relative Distances Preserved: The relative distances between values in each feature are preserved. While the absolute values change to fit within the new range, the proportions relative to the minimum and maximum of the original data are maintained.
3. No Impact on Shape of the Data Distribution: Min-Max scaling does not alter the shape of the feature's distribution. If the original data is normally distributed, skewed, or has any other distribution, the scaled data will maintain that distribution, albeit within the new range.
4. Dependency on Min and Max Values: The scaled values are highly sensitive to the minimum and maximum values in the data. If there are outliers, they will compress the rest of the data into a very small portion of the range, potentially reducing the effectiveness of the scaling.
5. Zero Values Transformation: If the original data contains zero values, they may no longer be zero after scaling unless the minimum value in the original data is 0. The transformation shifts all values proportionally between the new minimum and maximum.
6. Impact on Algorithms: Min-Max scaling can significantly impact the performance of machine learning algorithms that are sensitive to the magnitude of the input values, such as gradient descent-based algorithms, nearest neighbors, and neural networks. The scaling ensures that features contribute equally to the model's learning process, preventing features with larger scales from dominating the learning.



values, such as gradient descent-based algorithms, nearest neighbors, and neural networks. The scaling ensures that features contribute equally to the model's learning process, preventing features with larger scales from dominating the learning.

7. Reversibility: The Min-Max scaling process is reversible. Given the scaled value, along with the original minimum and maximum, you can calculate the original value. This is important for interpreting model outputs in the original scale.



Bin about → 0 → 0.3

high sensitive to outliers

max

outliers

more

less

0 1

$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}}$?

outliers ×

0 — 1 ✓

0-1

outliers

NN → activation → sigmoid / relu → 0-1

Z- [Standard

Norm]

$X_i = \frac{X_i - \mu}{\sigma}$

$X_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$

bounded

image

0-255

0-1

scaling $f_1 f_2 \cdots$

f_1 → did normal

normal-like

Standardization vs Minmax Normalization

29 February 2024 17:25

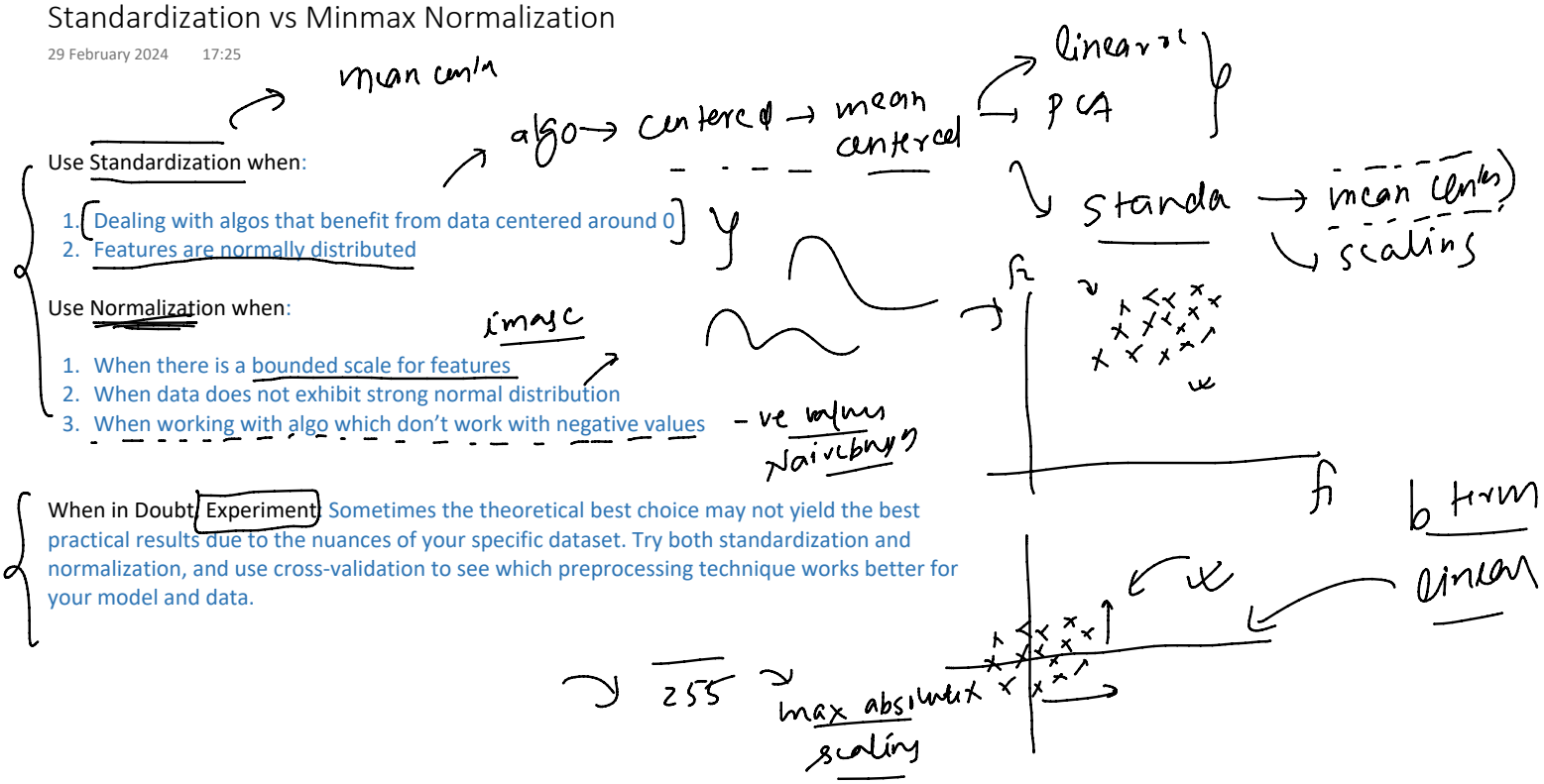
Use Standardization when:

1. Dealing with algos that benefit from data centered around 0
2. Features are normally distributed

Use Normalization when:

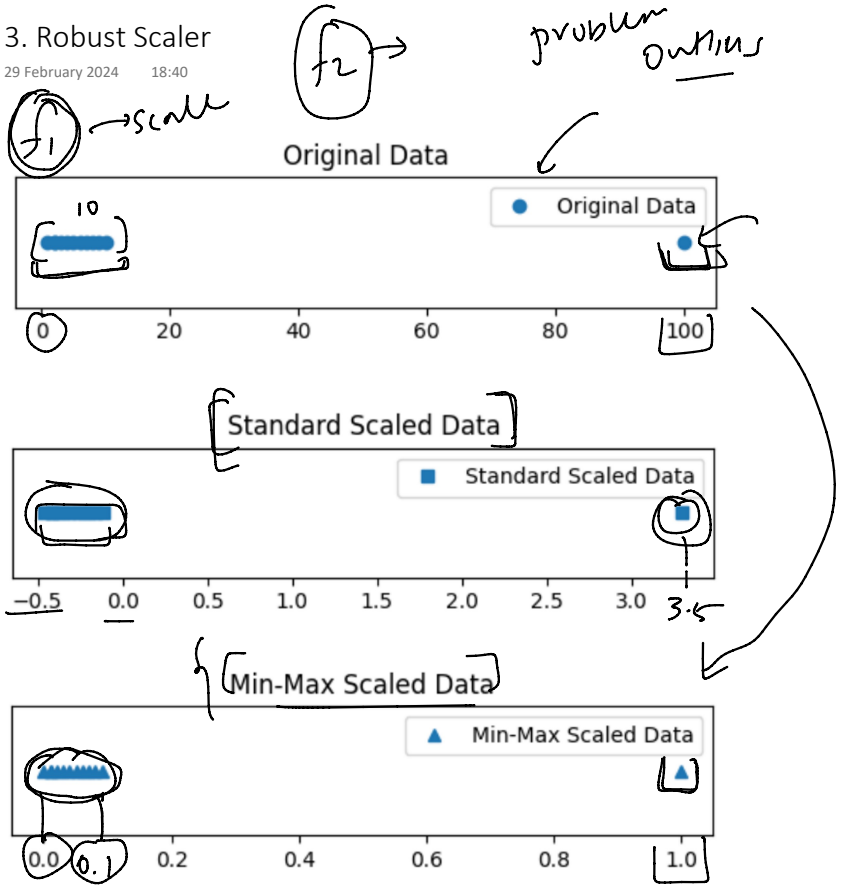
1. When there is a bounded scale for features
2. When data does not exhibit strong normal distribution
3. When working with algo which don't work with negative values

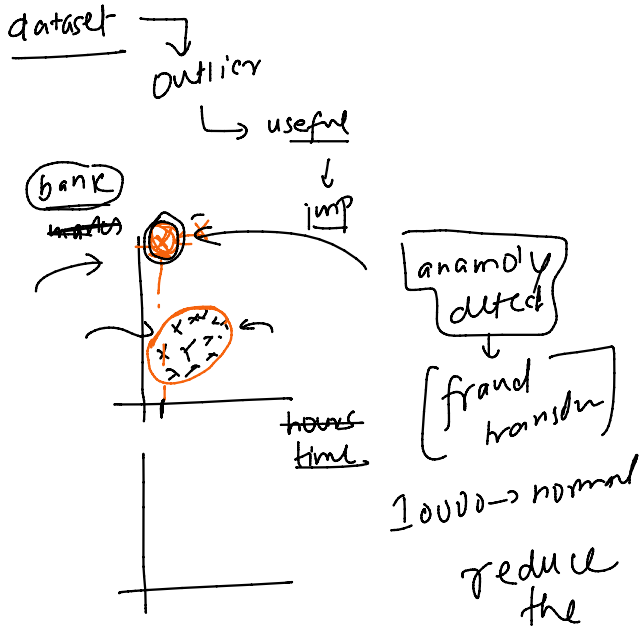
When in Doubt, Experiment: Sometimes the theoretical best choice may not yield the best practical results due to the nuances of your specific dataset. Try both standardization and normalization, and use cross-validation to see which preprocessing technique works better for your model and data.



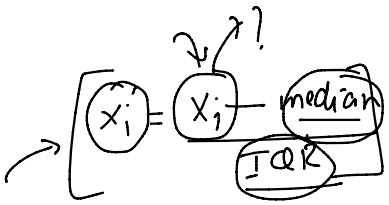
3. Robust Scaler

29 February 2024 18:40





Robust scaler
↳ Robust to outlier



25 pr

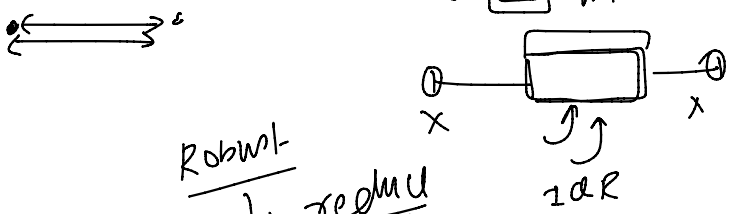
75pr − 25pr

Robust → Stand scale

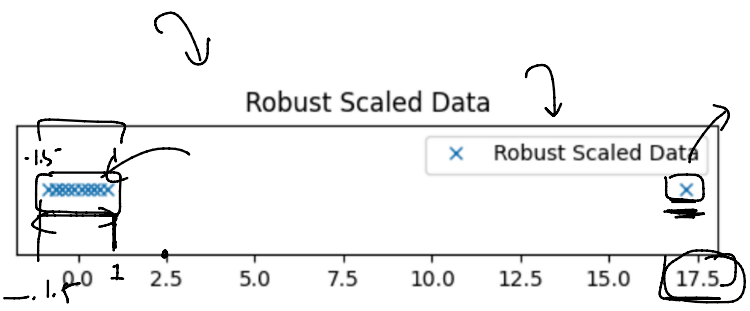
mean outlier

$x_i' = x_i - \mu$

σ middle



Robust
↳ reduce outlier



1. Median Centering: The median of each feature will be shifted to zero. Since the robust scaler subtracts the median from each data point, the center of the scaled data for any feature will be its median, aligning it at 0 in the scaled dataset.

2. IQR Scaling: The IQR (the range between the 25th and 75th percentiles) becomes the scaling reference, set to 1. This means that the difference between the first quartile (Q1) and the third quartile (Q3) in the scaled data will be 1. Consequently, data points that were at Q1 and Q3 in the original data will be at -1 and 1, respectively, in the scaled data.

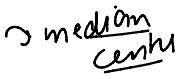
3. Outlier Influence Minimization: The scaling is less sensitive to outliers compared to methods like standardization or min-max scaling. Since the median and IQR are more robust against outliers, extreme values have less influence on the scaling process, ensuring that the core data is not skewed by anomalous points.

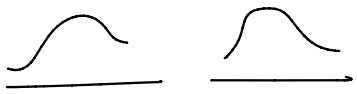
4. Data Distribution Shape: The overall shape of the data distribution is maintained. While the scale and location of the data change, the relative positioning of data points within the distribution remains constant.

5. No Fixed Range for Scaled Data: Unlike min-max scaling, the robust scaler does not confine the data to a fixed range. Scaled values can be greater than 1 or less than -1, especially for data points that lie outside the interquartile range.

6. Impact on Learning Algorithms: Scaling data with a robust scaler can improve the performance of machine learning algorithms sensitive to outliers. By ensuring that the feature's core data is effectively normalized, models can focus on the more representative part of the data.

7. Reversibility: The transformation is reversible, provided that you retain the original median and IQR values. Knowing these parameters allows you to reconstruct the original data values from the scaled data.





4 Max Absolute Scaler

29 February 2024 18:40

The formula for max absolute scaling a value X_i in a feature is:

$$X' = \frac{X_i}{\max(|X|)}$$

- **Range**: The scaled data will be within the range [-1, 1]. If the original data contains both positive and negative values, the scaled data will span this entire range. If the original data is all non-negative or non-positive, the scaled data will lie within [0, 1] or [-1, 0], respectively.
- **Preservation of Zero**: If a value is 0 in the original data, it remains 0 in the scaled data, as scaling does not introduce a shift.
- **Outlier Sensitivity**: Like min-max scaling, max absolute scaling is sensitive to outliers since extreme values will dictate the scaling factor for the entire feature.

1. Sign Preservation: The scaling maintains the sign of each value. Positive values remain positive, and negative values stay negative, ensuring that the relative relationships in terms of direction (positive or negative) are preserved.
2. Zero Values: Zero values in the data remain unchanged. This characteristic is crucial for datasets where the zero point carries specific meaning or indicates the absence of a feature.
3. Relative Distances: The relative distances between values in each feature are preserved. While the scale changes, the proportional relationships between data points remain consistent.
4. Outlier Sensitivity: The scaling process is sensitive to outliers. Extreme values can dominate the scaling factor, affecting how the rest of the data is scaled.
5. No Centering: The data is not centered around the mean or median; only the scale changes. The center of the distribution remains the same as in the original data.
6. Dimensional Consistency: Each feature is scaled independently, ensuring that the scaling does not distort the relationships between different features in multivariate data.
7. Sparse Data Suitability: The method is particularly suitable for sparse data, as it

scaling does not distort the relationships between different features in multivariate data.

7. Sparse Data Suitability: The method is particularly suitable for sparse data, as it does not alter zero entries, maintaining the sparsity pattern of the dataset.
8. Scale Interpretation: Post-scaling, the interpretation of a unit change in the data is consistent across features, facilitating comparisons and analyses that rely on the magnitude of values.
9. Reversibility: The scaling transformation is reversible. Knowing the original maximum absolute value allows you to rescale the data back to its original range, preserving the integrity of the data for inverse transformations if needed.

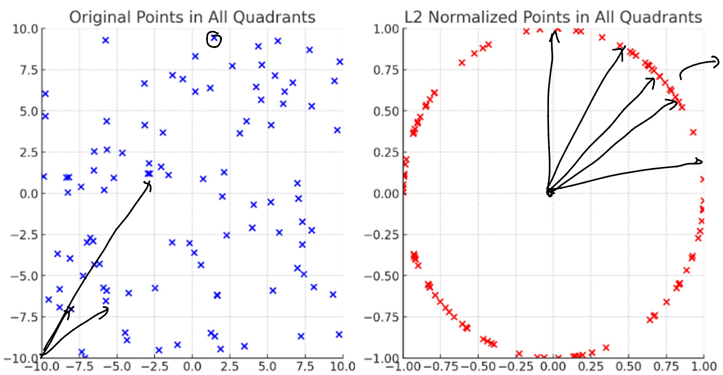
X → Scaling

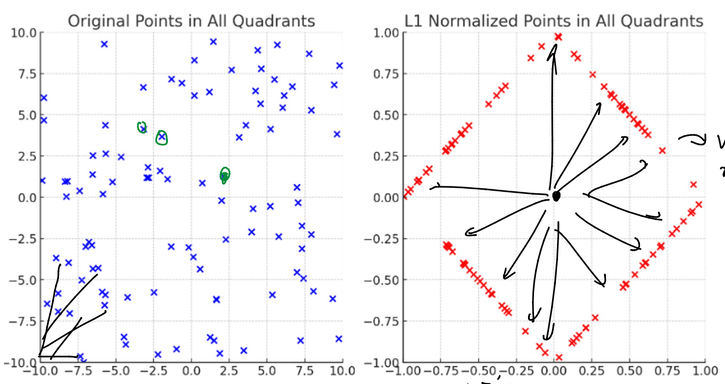
5. L2/L1 Normalization

05 March 2024 10:50

The What

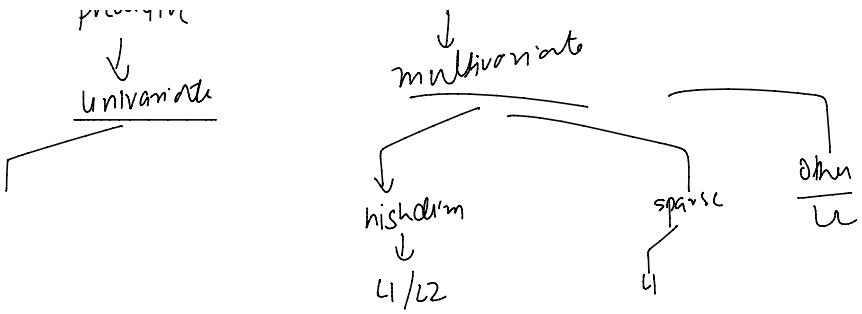
Geometric Intuition





Use-case

1. Text Preprocessing -> sparse features -> L1
2. Similarity based algos [cosine similarity]
3. Image processing based applications
4. Neural network embeddings


univariate
multivariate
hisdim
L1/L2
sparse
L1
other
L2

Comparison

02 March 2024 17:35

Univariate

Parameter/Criteria	Standardization	Min-Max Scaling	Robust Scaling	Max Absolute Scaling
Formula	$\frac{X-\mu}{\sigma}$	$\frac{X-\text{Min}}{\text{Max}-\text{Min}}$	$\frac{X-\text{Median}}{\text{IQR}}$	(\frac{X} {\text{Max}}(
Centering	Centers data around mean (0)	No centering; shifts data to start at 0	Centers data around the median	No centering; retains original data center
Scaling	Scales data by standard deviation	Scales data to a specified range, typically [0, 1]	Scales data by the interquartile range	Scales data relative to the maximum absolute value
Sensitivity to Outliers	Yes, outliers can significantly affect mean and standard deviation	Yes, outliers affect the minimum and maximum values	No, uses median and IQR which are less influenced by outliers	Yes, the maximum absolute value determines the scaling factor

Range of Scaled Data	No fixed range; typically within [-3, 3] for data close to normal distribution	Fixed and known range, typically [0, 1] or [-1, 1]	No fixed range; values can vary widely based on the IQR	Typically [-1, 1], but not strictly bounded if there are outliers
Impact on Distribution Shape	Maintains the shape of the distribution	Maintains the shape of the distribution	Maintains the shape of the distribution	Maintains the shape of the distribution
Suitability for Sparse Data	Not particularly suitable as it shifts zero values	Not suitable as it changes zero values unless the range starts at 0	Not particularly suitable as it can shift zero values	Suitable, especially when zero values are meaningful
Preservation of Zero	Zero values are shifted unless the mean is 0	Zero values are shifted unless the minimum value is 0	Zero values are shifted unless the median is 0	Zero values are preserved
Common Use Cases	Data with a Gaussian distribution, linear models, clustering, PCA	Neural networks, image processing, when data needs to be in a [0, 1] range	Data with outliers, robust machine learning models, financial data	Data where maximum value is crucial, text processing, sparse data
Interpretability	Interpretation in terms of how many standard deviations from the mean	Direct interpretability due to fixed range	Interpretation in terms of median and IQR	Interpretation in terms of a proportion of the maximum value
Impact on Feature Importance	Equalizes features' importance based on variance	Equalizes features' importance based on their range	Equalizes features' importance based on their spread (IQR)	Normalizes features' influence based on their maximum absolute value

Algorithmic Suitability	Algorithms assuming features with Gaussian distribution, linear models	Algorithms sensitive to feature scale, neural networks	Algorithms robust to outliers, median-based clustering	Algorithms where magnitude and sign are crucial, preserving data structure
Invariance to Transformations	Not invariant to multiplicative transformations of input	Invariant to linear transformations of input	Not invariant to multiplicative transformations of input	Invariant to scaling transformations of input
Mathematical Properties	Linear transformation	Linear transformation	Non-linear transformation (if outliers are present)	Linear transformation
Parameter Dependency	Depends on mean and standard deviation	Depends on minimum and maximum values	Depends on median and IQR	Depends on the maximum absolute value

Effect on Data Ordering	Maintains the order of data	Maintains the order of data	Maintains the order of data	Maintains the order of data
Typical Value Range Post-Scaling	Mostly within [-3, 3] but can have outliers	[0, 1] or a user-specified range like [-1, 1]	Varies widely; central data points fall between [-1, 1]	[-1, 1], but extremes can exceed this range if outliers are present
Ease of Reversibility	Easy to reverse with knowledge of original mean and std	Easy to reverse with original min and max values	Reversible with original median and IQR	Easy to reverse with the original max absolute value

multivariate

Parameter/Criteria	L2 Normalization	L1 Normalization
Formula	$\frac{X}{\|X\|_2}$ where $\|X\|_2 = \sqrt{\sum x_i^2}$	$\frac{X}{\|X\|_1}$ where (|X|_1 = \sum
Centering	No centering; does not alter the mean of the data	No centering; does not alter the mean of the data
Scaling	Scales the feature vector so its Euclidean norm (L2 norm) is 1	Scales the feature vector so its Manhattan norm (L1 norm) is 1
Sensitivity to Outliers	Moderately sensitive; outliers affect the norm, thus the scaling	Less sensitive; since it sums absolute values, the impact of outliers is diluted
Range of Scaled Data	No fixed range; the vector's length is 1, but individual elements can vary	No fixed range; elements are scaled relative to the sum of absolute values
Impact on Distribution Shape	Changes the magnitude but not the direction of data points in feature space	Changes the magnitude but not the direction of data points in feature space
Suitability for Sparse Data	Suitable; does not alter zero entries	Suitable; does not alter zero entries

Preservation of Zero	Zero values remain zero	Zero values remain zero
Common Use Cases	Text classification, clustering, image processing where direction matters	Text processing, creating sparse representations, feature selection
Interpretability	Normalizes data points to unit length; useful in cosine similarity	Scales data points by their aggregate absolute magnitude; useful in sparsity
Impact on Feature Importance	Normalizes overall influence of a feature vector, not individual features	Balances the overall influence by the total magnitude, promoting sparsity
Algorithmic Suitability	Algorithms where the magnitude of vectors isn't as important as their direction	Algorithms benefiting from sparsity and robustness to outliers or scale
Invariance to Transformations	Not invariant to additive transformations, invariant to scalar multiplications	Not invariant to additive or scalar multiplicative transformations
Mathematical Properties	Non-linear transformation (based on Euclidean norm)	Non-linear transformation (based on Manhattan norm)
Parameter Dependency	Dependent on the entire feature vector's norm	Dependent on the entire feature vector's norm
Effect on Data Ordering	Does not alter the order of features but normalizes their magnitude	Does not alter the order of features but normalizes their magnitude
Typical Value Range Post-Scaling	Each data point is scaled such that the sum of squares is 1	Each data point is scaled such that the sum of absolute values is 1
Ease of Reversibility	Not directly reversible without the original vector's norm	Not directly reversible without the original vector's norm