Data Curation Techniques

Siddharth R

Data Normalization

- Also called as Data Standardization
- Raw data into Standard format

Normalization is particularly important when:

- Features have different units or magnitudes.
- Distance-based algorithms (e.g., KNN, K-means, SVM) are used.
- Data needs to be prepared for better convergence in deep learning models.

Why is Data Normalization Important

1. Features with larger values may disproportionately influence model training.
2. Scaled values help reduce the complexity of optimization algorithms.
3. Many machine learning algorithms (e.g., gradient descent-based models) work better with normalized data.
4. Standardized scales make comparisons easier.

Common Types:

1. Min-Max Normalization
2. Z-score Normalization
3. Decimal Scaling
4. Robust Scaling

Min-Max Normalization

Normalized Value

Original Value

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Maximum Value of x

Minimum Value of x

When to use

- When data has no outliers and a known/stable range
- When preserving relative distances between data points is critical

When to avoid:

$$x'' = 2\frac{x - \min x}{\max x - \min x} - 1$$

- Sensitive to Outliers
- Not Robust to Changes in Data: If new data points introduce a new min or max, you must recompute normalization.

Special Case: When the data is centered around zero. What would you do?

Z-Score Normalization

Observed value

Mean value

$$z = \frac{x - \mu}{\sigma}$$

z-Score

Standard deviation

- Here instead of min and max, we are using the mean and standard deviation.

When to use:

- Z-score normalization is particularly useful when the data is approximately normally distributed.
- When mean and SD are meaningful

When not to use:

- When you are data is having skewed distributions / outlier
- When interpretability of original units is required
- Doesn't Bound Data

Decimal Scaling

- Dividing each feature by a power of 10.

$$X' = X / 10^j$$

The value of j is determined by the maximum absolute value in the dataset:

$$j = \text{ceil}(\log10(\max(|X|)))$$

When to use:

- For simplicity in preserving relative magnitudes without complex calculations
- When working with integer-heavy data

When to avoid:

- Not robust to outliers

Robust Scaling

$$\underbrace{x'}_{\text{Robust Standardised Value}} = \frac{\overbrace{x}^{\text{Original Value}} - \overbrace{\text{median}(x)}^{\text{Sample Median}}}{\underbrace{(Q3-Q1)}_{\text{Interquartile Range} = Q3 - Q1}}$$

- Median and IQR is used instead of mean and standard deviation

When to Use:

- For outlier-rich datasets / skewed distributions

When to avoid:

- When median/IQR misrepresents data (e.g., bimodal distributions)

Summary

Technique	**Outlier Handling**	**Preserves Distribution**	**Ideal data type**
Min-Max	Poor	Yes	Bounded, no outlier
Z-Score	Moderate	No (Gaussianizes)	Gaussian distributed
Decimal Scaling	Poor	Yes	Integer- heavy
Robust Scaling	Excellent	Partial (uses IQR)	Skewed / outlier

Handling Redundancy

- Redundancy vs Duplicates
- An attribute may be redundant if it can be “derived” from another attribute or set of attributes.
- Any examples?
- Identifiers: Attributes that uniquely identify an individual
- Quasi-Identifiers: Attributes that do not uniquely identify a person but, when combined, can lead to re-identification.
- Data type: continuous vs continuous , categorical vs categorical

Chi-Square Test

To determine whether there is a significant **association** between two **categorical / nominal** variables.

Steps:

1. Define Hypotheses :
    a. Null Hypothesis (H_0): The two categorical variables are independent (no relationship).
    b. Alternative Hypothesis (H_1): The two variables are dependent (have a relationship).
2. Create a Contingency Table :
    Suppose A has *c* distinct values,namely a_1,a_2, a_c.B has *r* distinct values,namely b_1,b_2, b_rThe data tuples described by A and B can be shown as a contingency table
3. Calculate Expected Frequencies :
$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$
4. Calculate Chi-square :
$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

O - Observed value
E - Expected value

Example:

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

Step 1: Hypothesis for this problem?

Step 2: Calculate expected frequency

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

Example (Contd.)

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Step 3: Calculate Chi-Square value

Step 4: Find degree of freedom => (R-1) * (C-1) , where R is the number of rows and C is the number of columns

Step 5: Based on the alpha value , refer the chi-square distribution table to compare the critical value with the chi-square value.

Step 6: If chi-square value is greater than critical value , then reject null hypothesis , else we fail to reject the null hypothesis

Example (Contd.)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

- Degree of freedom = (2-1) * (2-1) = 1
- For 1 degree of freedom, the value needed to reject the hypothesis at the 0.001 significance level is 10.828
- Since our computed value is above this,we can reject the hypothesis that gender and preferred reading are independent.
- Link to the table: https://www.di-mgt.com.au/chisquare-table.html

Exercise

	Program 1	Program 2	Current Program	Total
# Passed	112	94	130	336
# Failed	60	79	85	224
Total	172	173	215	560

Alpha value (p) = 0.05

Chi-Square Value = ?

	p					
v	0.100	0.050	0.025	0.010	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	14.6837	16.9190	19.0228	21.6660	23.5893	27.8772
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883

Thank You