Feature Engineering in Machine Learning

Ayush Singh[1]

Antern Department of Artificial Intelligence
ayush@antern.co

Abstract. This document contains contents on data preparation in machine learning and we also cover several components of data preparation like feature engineering, feature selection, dimensionality reduction, etc. We provide readers with several traditional and modern techniques to handle complicated data tasks.

Key words: data preparation, feature engineering, feature selection, data cleansing, data transformation, dimensionality reduction

1 Engineering Numerical Features

Scaling is important because many machine learning algorithms are sensitive to the size of the input features. If some features have much larger values than others, the algorithms may focus too much on those features, leading to less accurate results.

Consider a dataset with two features, 'Age' and 'Income':

Age	Income
20	2000
25	2500
30	3000
35	3500
40	4000

In this dataset, the 'Income' feature has a much larger magnitude than the 'Age' feature. A machine learning algorithm might give more importance to 'Income', even though 'Age' might also be a crucial factor.

To fix this issue, we can think of a technique that transforms the features to a common scale. One such technique is Min-Max scaling.

Min-Max scaling is a technique that scales the numerical features to a specific range, usually [0, 1]. The formula for Min-Max scaling is:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Dataset:

Age	Height
20	150
25	155
30	160
35	165
40	170

Min-Max Scaling:

First, we calculate the minimum and maximum values for both features:

- Age: min=20, max=40
- Height: min=150, max=170

Next, we apply Min-Max scaling to the dataset:

Scaled dataset:

Scaled Age	Scaled Height
0	0
0.25	0.25
0.5	0.5
0.75	0.75
1	1

Euclidean distance:

Original distance:

Distance $= \sqrt{(20-25)^2 + (150-155)^2} = \sqrt{25+25} = \sqrt{50}$

Scaled distance:

Distance $= \sqrt{(0-0.25)^2 + (0-0.25)^2} = \sqrt{0.0625+0.0625} = \sqrt{0.125}$

By applying Min-Max scaling, we can see that the distance calculation is more balanced and gives equal importance to both 'Age' and 'Height'. This will help the machine learning algorithm to make better predictions by considering both features fairly.

1.1 Several Transformations

Here's the colab which contains the detailed explanation for the same:

https://colab.research.google.com/drive/1D5N7EDT5KtuwKsr4aptNw866Boh0C4nC?usp=sharing

2 Interaction Effects

Interactions in prediction:

- Occur when the combined effect of two or more features on the outcome is different from their individual effects.
- Can improve predictions by considering the combined effects of features.
- Can occur between numerical, categorical, or mixed features.

Example 1 - Water and fertilizer on crop yield:

- No water + some fertilizer = No yield (water is essential).
- Sufficient water + no fertilizer = Some yield (not optimal).
- Sufficient water + sufficient fertilizer =
- Optimal yield (combined effect greater than individual effects).

Example 2 - Ames housing data (age of house, air conditioning, and sale price):

- Houses with air conditioning: positive relationship between age and sale price.
- Houses without air conditioning: no relationship between age and sale price.
- Interaction between age of house and presence of air conditioning, as their combined effect on sale price is different from their individual effects.

Importance of interactions:

- Help improve model performance and accuracy.
- Identify and incorporate interactions to better understand the relationships between features and outcomes.

Interaction representation in a simple linear model:

- Equation: y = 0 + 1x1 + 2x2 + 3x1x2 + error
- 0: overall average response
- 1 and 2: average rate of change due to x1 and x2, respectively
- 3: incremental rate of change due to the combined effect of x1 and x2

Estimating parameters:

- Use methods like linear regression (for continuous response) or logistic regression (for categorical response) to estimate the parameters from data

Evaluating interaction usefulness:

- Determine the usefulness of the interaction term (3x1x2) for explaining variation in the response after estimating the parameters
- Helps in understanding the significance of the interaction between predictors

Interaction Type	**Example**	**Description**
Additive	Exercise and healthy diet on weight loss	The combined effect of exercise and a healthy diet on weight loss is the sum of their individual effects.
Antagonistic	Sleep and caffeine intake on alertness	The combined effect of sleep and caffeine intake on alertness is less than the sum of their individual effects, as caffeine reduces the effectiveness of sleep on alertness.
Synergistic	Sunscreen and wearing a hat on preventing sunburn	The combined effect of sunscreen and wearing a hat on preventing sunburn is greater than the sum of their individual effects, providing better protection.
Atypical	Medication effect on pain relief in acute vs. chronic pain patients	The effect of medication on pain relief depends on the pain type (acute or chronic), but the main effect of one or both predictors on the response is not significant.

How to find interaction terms?

Concept	**Example**
Expert knowledge	A nutritionist's knowledge on the impact of different nutrients on health.
Experimental design	Designing a study to assess the effects of different types of exercise on weight loss.
Interaction hierarchy	In a pizza satisfaction study, pairwise interactions (crust-sauce, crust-cheese) should be considered before higher-order interactions (crust-sauce-cheese).
Effect sparsity	In the pizza satisfaction study, only a few factors (e.g., crust, cheese) and interactions (e.g., crust-sauce) might significantly impact customer satisfaction.

2.1 Heredity Principle

This principle is inspired by genetic heredity and states that an interaction term should only be considered if the preceding terms are effective in explaining the response variation.

Strong Heredity Example

Suppose you are studying the effect of three factors on plant growth: sunlight (x1), water (x2), and fertilizer (x3). You find that both sunlight (x1) and water (x2) have significant main effects on plant growth. According to the strong heredity principle, you can consider the interaction between sunlight and water (x1 × x2) in your model. However, if only sunlight (x1) had a significant main effect, you would not consider any interaction terms in the model, as strong heredity requires all lower-level preceding terms to be significant.

Weak Heredity Example

Using the same plant growth example with factors sunlight (x1), water (x2), and fertilizer (x3), let's say you find that only sunlight (x1) has a significant main effect on plant growth. According to the weak heredity principle, you can consider the interactions between sunlight and water (x1 × x2) and sunlight and fertilizer (x1 × x3) in your model, even though water (x2) and fertilizer (x3) don't have significant main effects. However, the interaction between water and fertilizer (x2 × x3) would not be considered, as neither of the main effects is significant.

3 Identifying Potential Interaction Terms

Imagine you are studying the effect of five factors on the sales of a product: price (x1), advertising (x2), packaging (x3), product quality (x4), and customer support (x5). You want to find the most important pairwise interactions that affect sales.

3.1 Brute-Force Approach

With the **brute-force approach**, you evaluate all possible pairwise interactions for an association with the response (in this case, sales). For five factors, there are 10 possible pairwise interactions: $(x1 \times x2)$, $(x1 \times x3)$, $(x1 \times x4)$, $(x1 \times x5)$, $(x2 \times x3)$, $(x2 \times x4)$, $(x2 \times x5)$, $(x3 \times x4)$, $(x3 \times x5)$, and $(x4 \times x5)$.

3.2 Drawbacks

As the number of evaluated interaction terms increases, the probability of identifying an interaction associated with the response due to random chance also increases. These terms, which are statistically significant only due to random chance and not because of a true relationship, are called **false positive findings**.

False positive findings can lead to overfitting and decrease a model's predictive performance. To protect against selecting these types of findings, an entire sub-field of statistics is devoted to developing methodology for controlling the chance of false positive findings.

3.3 Simple Screening

In the context of **Simple Screening**, let's consider an example where you want to predict house prices based on two factors: **square footage** (x1) and the **age** of the house (x2).

Main Effects Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + error \tag{2}$$

Interaction Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + error \tag{3}$$

These two models are called "*nested*" since the first model is a subset of the second. When models are nested, a statistical comparison can be made regarding the amount of additional information that is captured by the interaction term. For linear regression, the residual error is compared between these two models and the hypothesis test evaluates whether the improvement in error, adjusted for degrees of freedom, is sufficient to be considered real. The statistical test results in a p-value which reflects the probability that the additional information captured by the term is due to random chance. Small p-values, say less than 0.05, would indicate that there is less than a 5% chance that the additional information captured is due to randomness. It should be noted that the 5% is the rate of false positive findings, and is a historical rule-of-thumb. However, if one is willing to take on more risk of false positive findings for a specific problem, then the cut-off can be set to a higher value.

For linear regression, the objective function used to compare models is the statistical likelihood (the residual error, in this case). For other models, such as logistic regression, the objective function to compare nested models would be the binomial likelihood.