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 Feature Selection and How does it helps?

The process of selecting a subset of the most pertinent and instructive features from the initial collection of features in a dataset is known as feature selection. This is done to simplify the model, prevent overfitting, boost training effectiveness, and make the model easier to understand.

With feature selection, you can:

- Reduce overfitting: By limiting the model's ability to fit data noise by only using the most pertinent features, we improve generalisation to new data.
- Enhancing training effectiveness: The training procedure is quicker and uses fewer CPU resources when there are fewer features.
- Improving interpretability: In areas where explainability is critical, a model with fewer elements is simpler to comprehend and interpret.

Working illustration

Consider a dataset that contains details on homes, such as their age, location, square footage, number of rooms, and proximity to the city centre. Predicting housing prices is the objective.

House	Rooms	Sq. Footage	Age	Location	Distance from City Center	Price
1	3	1,200	10	Urban	2.0	250k
2	4	1,800	5	Suburban	5.5	300k
...						

You discover after examining the dataset that "Rooms" and "Sq. Footage" have a strong correlation (houses with more rooms generally have more square footage). You discover that the variable "Location" has little effect on the prices of houses in your dataset. You choose to perform feature selection and remove the "Sq.

Footage" and "Location" aspects in order to streamline your model and boost its functionality.

House	Rooms	Age	Distance from City Center	Price
1	3	10	2.0	250k
2	4	5	5.5	300k
...				

With fewer features to take into account, your model may be less overfitted, train more quickly, and produce predictions that are simpler to understand.

Many feature selection techniques exist, including filter techniques (such as correlation and mutual information), wrapper techniques (such as forward selection and backward removal), and embedding techniques (e.g., LASSO, Ridge Regression). The choice of method is based on the particular problem and dataset, each of which has strengths and disadvantages.