Feature Engineering

CH1 Machine learning pipeline

- **Data**(Real-world phenomena) => **Task**(**WHY** do we collect data?)
- **Models**: A mathematical model of data describes the relationships between different aspects of the data.
- **Features**: A feature is a numeric representation of raw data.
- **Feature engineering** is the process of formulating the most appropriate features given the data, the model, and the task.

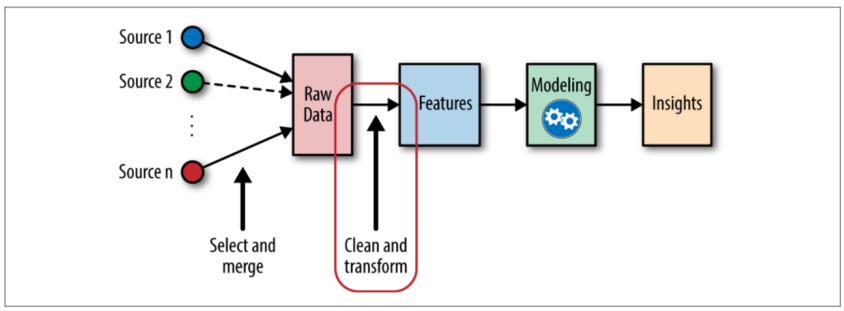
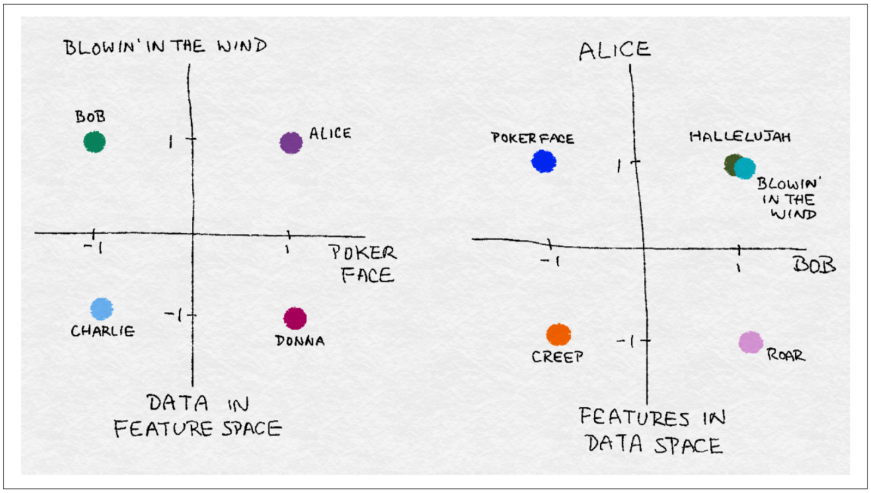


Figure 1-2. The place of feature engineering in the machine learning workflow

CH2 Fancy Tricks with Simple Numbers

- numeric feature = scalar
- a ordered list of scalars = vector
- vectors sits within a vector space

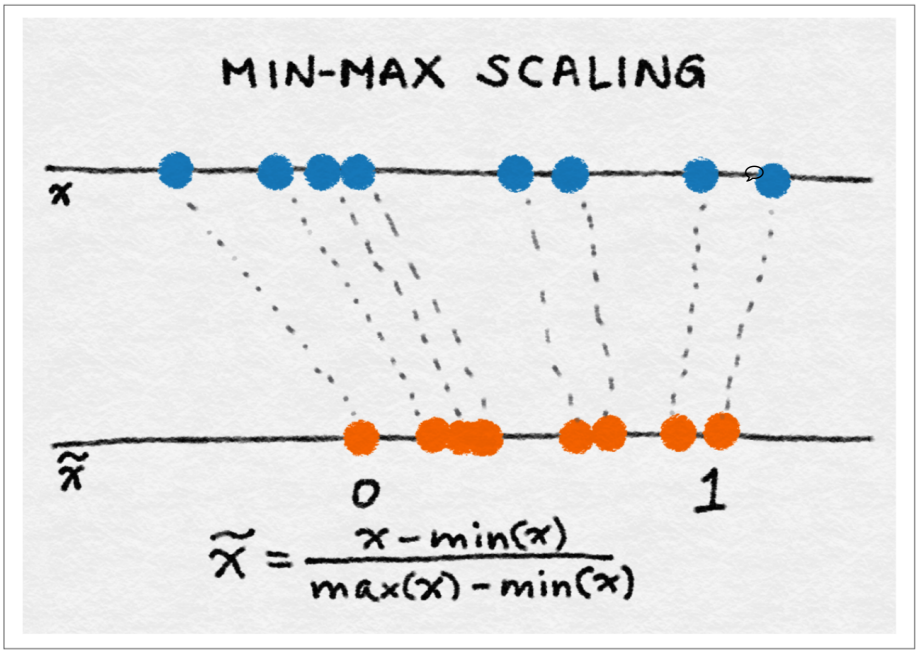


Binarization / Quantization

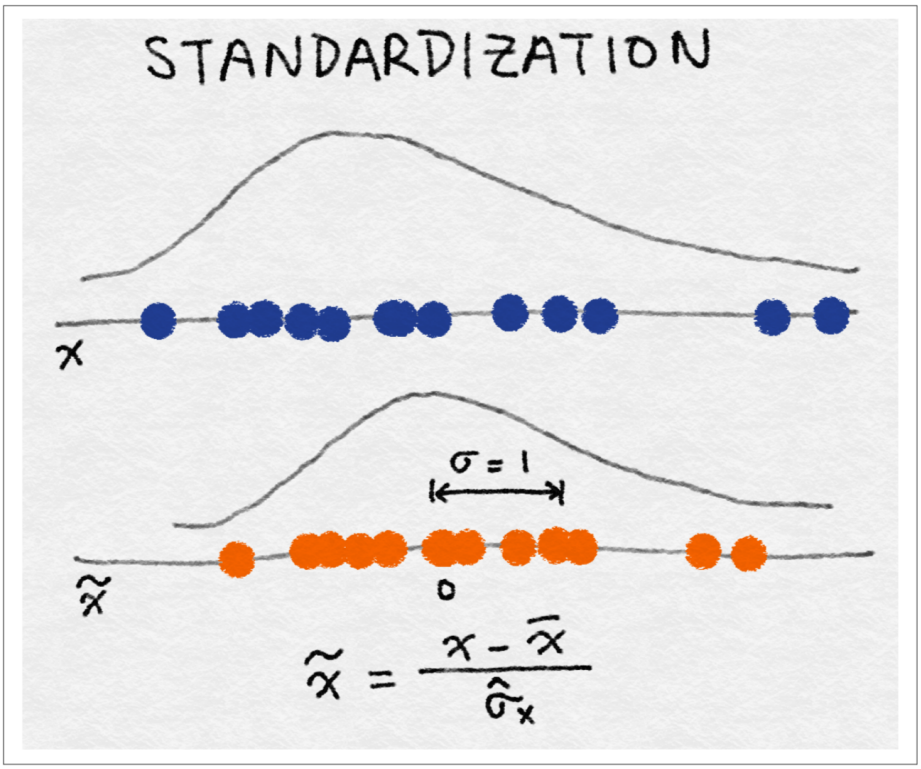
Log Transformation

Normalization scaling

Min-Max

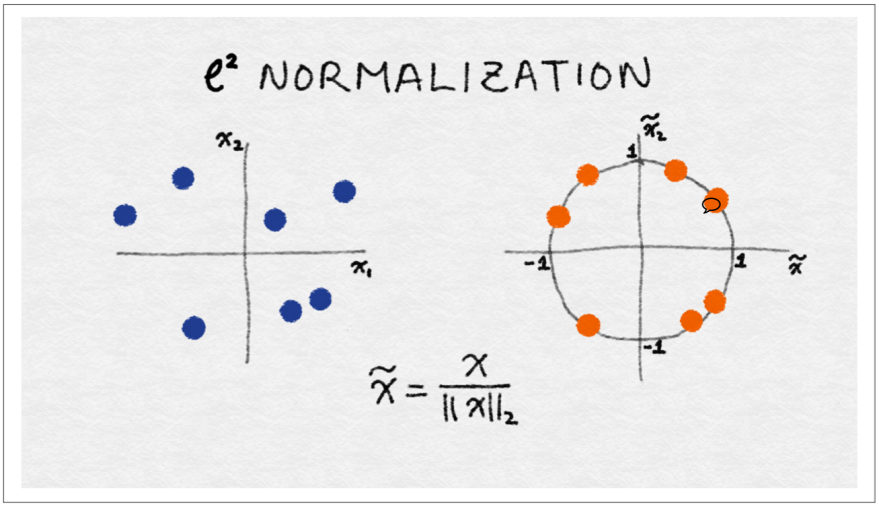


Standardization (Variance Scaling)



l2 Normalization

- l2 norm, also known as the Euclidean norm



feature selection

feature selection techniques fall into three classes:

- Filtering:

1. 篩選掉對模型不重要的特徵
2. 計算和Y或其他特徵的相關係數
3. cheap

- Wrapper methods:

1. 為每個特徵子集訓練一個新模型，使用預測模型給特徵子集打分

- Embedded methods:

1. 構建線性模型的LASSO方法