Feature Engineering: Filtering and Scaling

Motivation: Selecting relevant features to use for model training
- Remove channels from an image if color is not important
- Remove frequencies from audio if power is less than a threshold

Scaling: Motivation - risk
Many algorithms are sensitive to features being on different scales, e.g. gradient descent, KNN
Align all features on the same scale

Some algorithms (like decision trees and random forests) aren't sensitive to features on different scales

Important: fit the scaler to training data only, then transform both train and validation data.

Common choices in sklearn
- mean/variance standardization
- min/max scaling
- matabs scaling
- robust scaling
- normalizer

Mean/Variance Transform:

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$$

Scaled values are centered around mean $\mu_j = 0$ with std deviation $\sigma_j = 1$ for each data column
sklearn preprocessing StandardScaler
For feature x, find mean & std dev, remove mean & divide by std dev

Many algs scale better with smaller values, keeps outliers but minimizes their influence

Filtering and Scaling -2

Min Max

Transform $x^*_{i,j} = \frac{x_i - \min x_j}{\max x_j - \min x_j}$

Scales vals so min = 0, max = 1

Very robust to small std deviations

MaxAbs Scaling

$x^*_{i,j} = \frac{x_{i,j}}{\max(|x_j|)}$

sklearn preprocessing MaxAbsScaler

Doesn't destroy sparsity

Robust Scaling

$x^*_{ij} = \frac{x_i - Q_{25}(x)}{Q_{75}(x) - Q_{25}(x)}$

RobustScaler

Robust to outliers

Normalizer

$x^*_{i,j} = \frac{x_{j,j}}{\sigma_j}$

applied to a row, scaling applied to a column

scaled values are scaled with standard deviation $\sigma_j = 1$

sklearn.preprocessing.Normalizer

apply when multiple numeric features

$x_{i,j} = \frac{x_j - \mu_x}{\sigma_x}$

Rescales x_j to unit norm based on

- L1 norm
- L2 norm
- Max norm

Widely used in text analysis