Unsupervised Learning

Part 5: Feature selection

By: Noureddin Sadawi, PhD

University of London

What is feature selection?

- Automatic selection of the most informative (i.e. most relevant) feature subset from the original features.
- In other words, removal of unimportant features and keeping the features that contribute the most to the prediction of the outcome.

Why feature selection is useful

- Uninformative features can have a negative impact on the performance of many models.
- In particular, linear models such as logistic and linear regression are usually affected by irrelevant features.

Benefits of feature selection

- Accuracy improvement: having only informative and relevant data leads to improvement in model performance.
- Reduction of overfitting: when the model is trained on informative and relevant features it does not rely on unimportant information to make decisions
- Faster training time: as data size is reduced by the removal of unimportant features, the model should train faster!

Remove features with low variance

- ‘A simple baseline approach to feature selection.
- It removes all features whose variance doesn’t meet some threshold.
- By default, it removes all zero-variance features, i.e. features that have the same value in all samples.’
- Variance is the average of the squared differences from the mean.

$$Variance = \frac{\sum_{i=0}^{n}(x_i - \mu)^2}{n-1}$$

https://scikit-learn.org/stable/modules/feature_selection.html

Correlation-based feature selection

- Correlation measures the strength of a linear relationship between two numeric variables (i.e. quantitative variables).
- A regression problem (where the outcome is quantitative) is suitable if it has quantitative features(s).
- The idea here is to drop features that have a low correlation with the outcome variable.
- A predefined threshold can be used.

Recursive feature elimination

- ‘RFE selects features by recursively considering smaller and smaller sets of features.
- First, an estimator is trained on the initial set of features and the importance of each feature is obtained.
- Then, the least important features are pruned from current set of features.
- That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.’

https://scikit-learn.org/stable/modules/feature_selection.html