GUIDE TO FEATURE SELECTION AND TRANSFORMATION (UNIT 7)

COURSE: MACHINE LEARNING I

MASTER IN BIG DATA ANALYTICS

Ricardo Aler Mur

- First, preprocessing is introduced as a step in the Machine Learning pipeline
- One of the most important preprocessing processes is attribute selection.
- Attribute selection is important to remove redundant and irrelevant attributes, and ease the curse of dimensionality.
- The curse of dimensionality can happen even in linear classifiers.
- An important idea in attribute selection is that the unit of selection might not be an attribute but a set of attributes, because sometimes, two (or more) attributes do not work well when used in isolation, but work well when used together.
- The filter / wrapper and single / subset classification of selection algorithms are introduced.
- Some single attribute selection algorithms are introduced (ranking): based on entropy (which was already used to select the best attribute in decision trees), based on mutual-information, and based on chi-square.
- The correlation feature selection (CFS) and Wrapper algorithms are introduced as subset selection algorithm.
- Finally, it is explained the difference between attribute selection and attribute transformation: selection selects the important attributes, while generation

transforms the original attributes into new ones, which might have more desirable properties.

- Two main attribute transformation algorithms are explained: PCA and random projections.
- PCA is a non-supervised algorithm and while it can be used to generate new attributes and reduce dimensionality, it has to be used with care when doing classification, given that PCA does not consider the class of the data. Random Projections is introduced as a faster method for doing PCA-like attribute generation and selection. Random Projections and PCA can obtain similar results when the dimensionality of the target space is still large.

Associated material

In addition to the lecture slides and some exercises (check the course guide), a new Python lab for attribute selection is provided.