DATA PREPROCESSING

PC LEE

DEC. 2018

WHY DATA PREPROCESSING – GARBAGE IN AND GARBAGE OUT

- We do not know which factors are the most important ones before our data collection process. So, sometimes the collected features are too many to practical computational analysis and sometimes these features are dependent.
- In a real-world study, we get data with missing entry
- We collected only categorical factor e.g., S, M, L, XL and this have to be transformed before further computation
- Sometimes the range of a factor is just too huge comparing to others e.g. driving miles and age. This would influence the analysis.

WHY DATA PREPROCESSING – GARBAGE IN AND GARBAGE OUT

- We do not know which factors are the most important ones before our data collection process. So, sometimes the collected features are too many to practical computational analysis and sometimes these features are dependent.
- In a real-world study, we get data with missing entry
- We collected only categorical factor e.g., S, M, L, XL and this have to be transformed before further computation
- Sometimes the range of a factor is just too huge comparing to others e.g. driving miles and age. This would influence the analysis.

TOOLS AND DATA

- For .NET development: .NET framework >= 4.6, Deedle (BSD) with Accord.Net 3.8.0 (LGPL)
- Python >= 3.6 with pandas, scikit-learn, matplotlib
- Data: Wine quality

MISSING DATA

- Usually there are NaNs or N/As appearing in some entries
- Drop
  - This is a brute force strategy
- Filling something
  - Median/mean of a row (column)
  - Most frequent value

CATEGORICAL DATA

- We can not directly compute with literal nouns e.g., ***red***, ***green***, ***blue***. Before computation, they shall be transformed into numerical presentation
- **Direct value mapping** e.g., ***red*** → 0, ***green*** → 1, and ***blue*** → 2
  Disadvantage: Imply that blue is larger than red; none sense!
- **One-hot encoding**: index to binary vector mapping e.g.,
  red → (1, 0 ,0), ***green*** → (0, 1, 0), and ***blue*** → (0, 0, 1)
  In most cast, this is a more acceptable method and will bring out more meaningful analysis.

FEATURE SCALING

- ***This is a very critical step*** for many applications**!** Considering a 3-vector feature case, (x_1, x_2, x_3) and the distance two samples are $\mathrm{sum}(\mathrm{pow}(f_1\text{-}f_2, 2))$. If the range of x_2 is 2-order larger than x_1 and x_3, almost surely the analysis result is dominated by the factor, x_2.

Normalization	**Standarization**
• Mapping to zero-one interval	• Mapping to a normal distribution

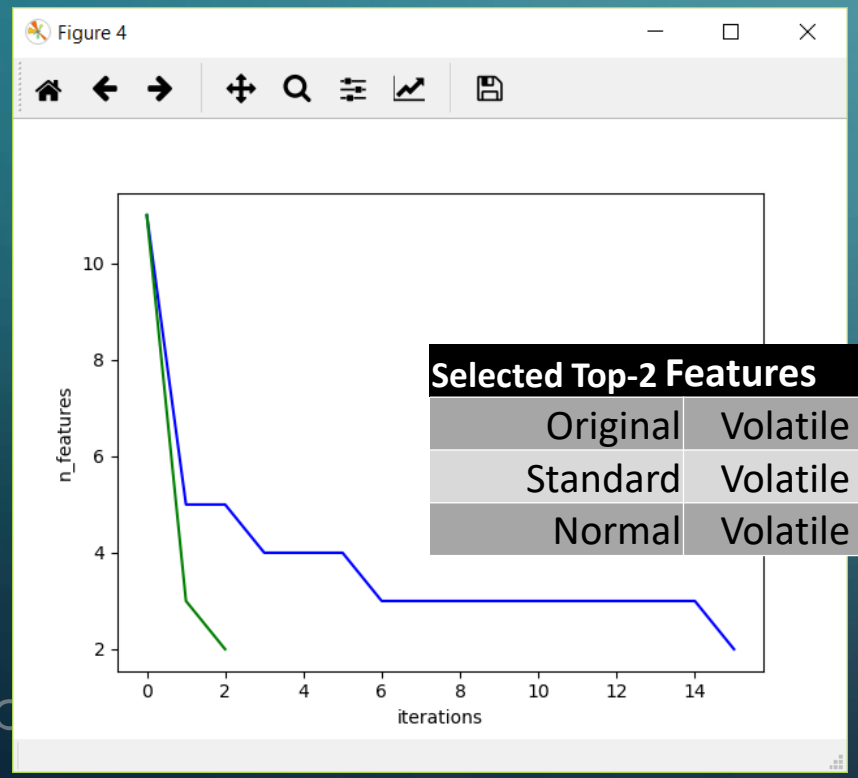
TOY CASE: WINE QUALITY

"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6

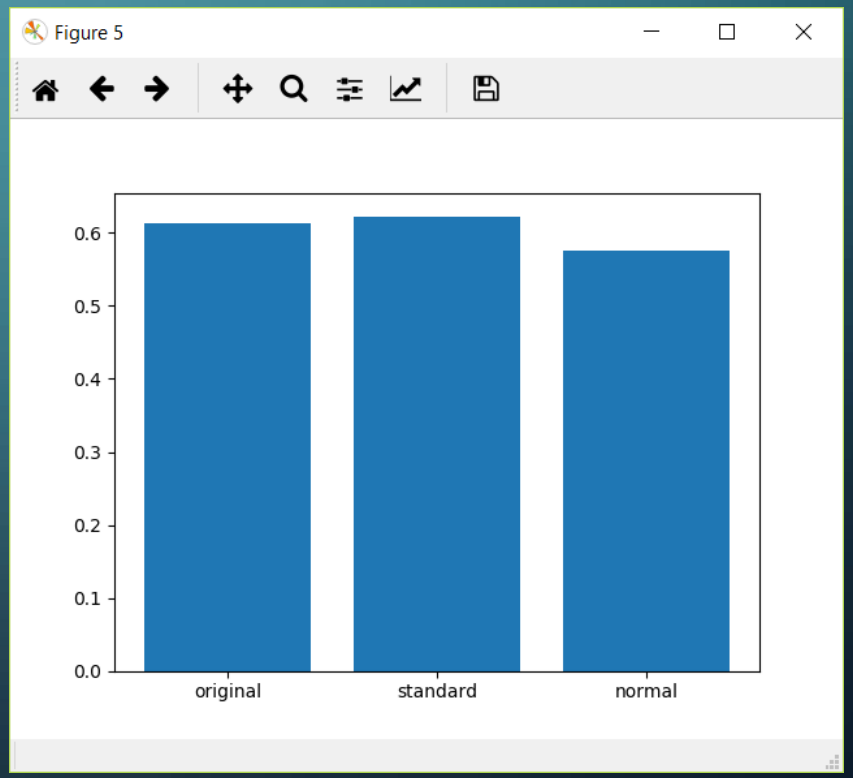
Data source: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

FEATURE SCALING

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8



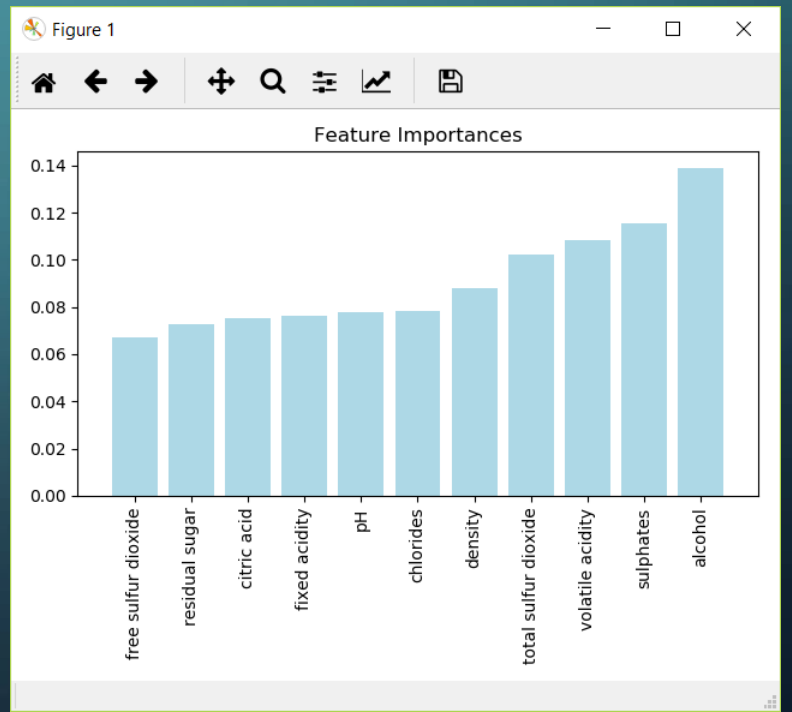
Selected Top-2 Features		
Original	Volatile acidity	pH
Standard	Volatile acidity	alcohol
Normal	Volatile acidity	alcohol



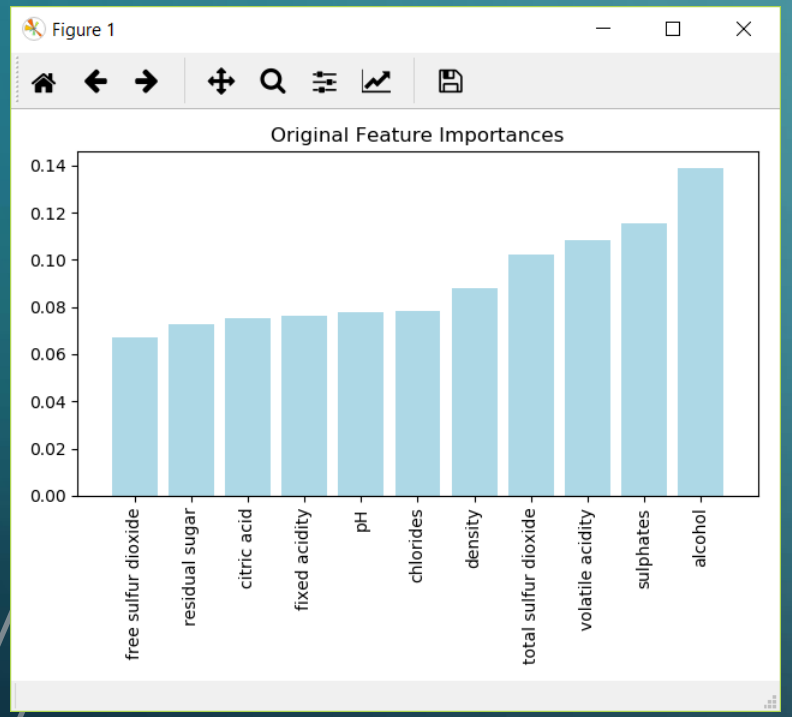
FEATURE SELECTION

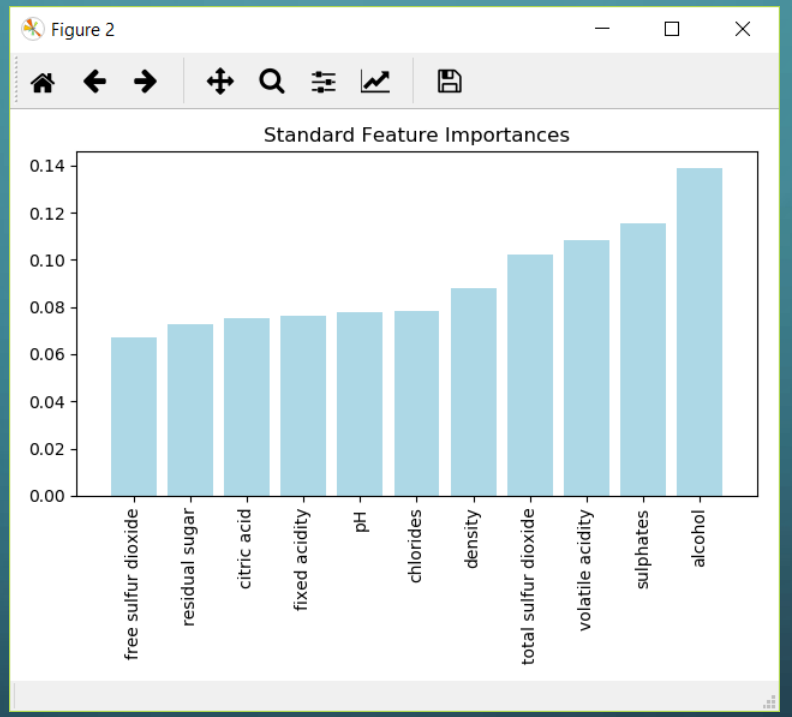
- Some selection method is scale-invariant, one of the most famous is ***Random forest*** selection method. The distance between samples are measured in geodesic rather than Euclidean distance

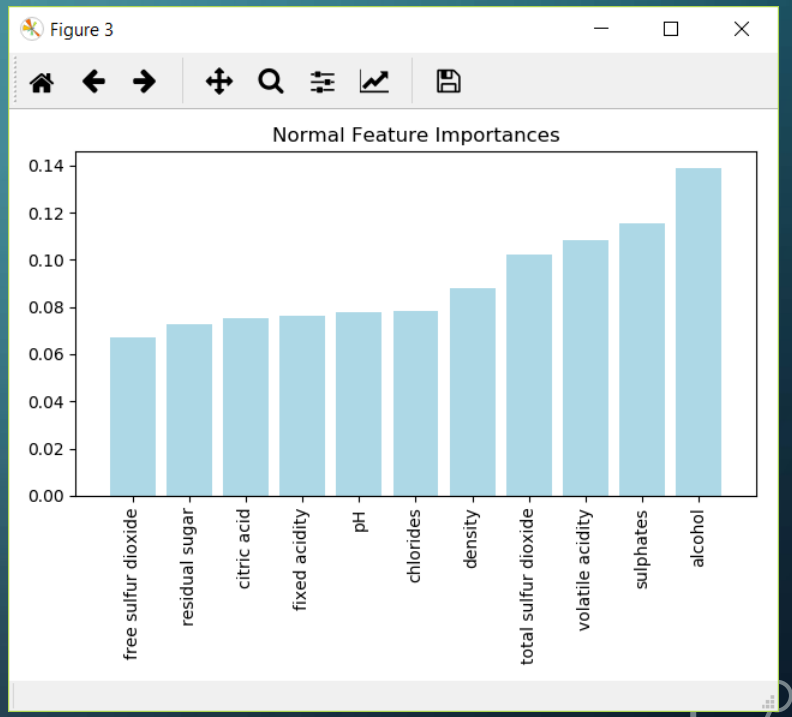
Selected Top-2 Features (Linear regression)		
Original	Volatile acidity	pH
Standard	Volatile acidity	alcohol
Normal	Volatile acidity	alcohol
Selected Top-2 Features (Random forest)		
	alcohol	sulphates



FEATURE SELECTION







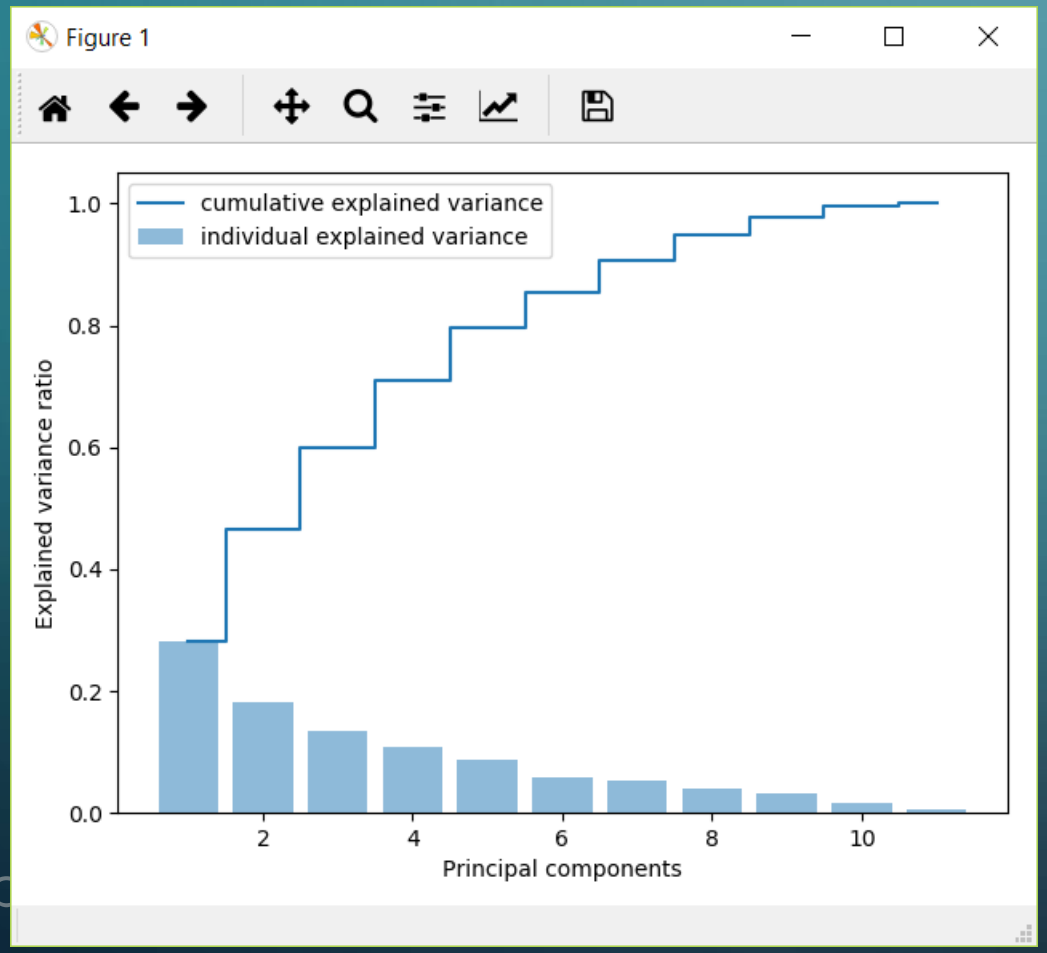
FEATURE EXTRACTION

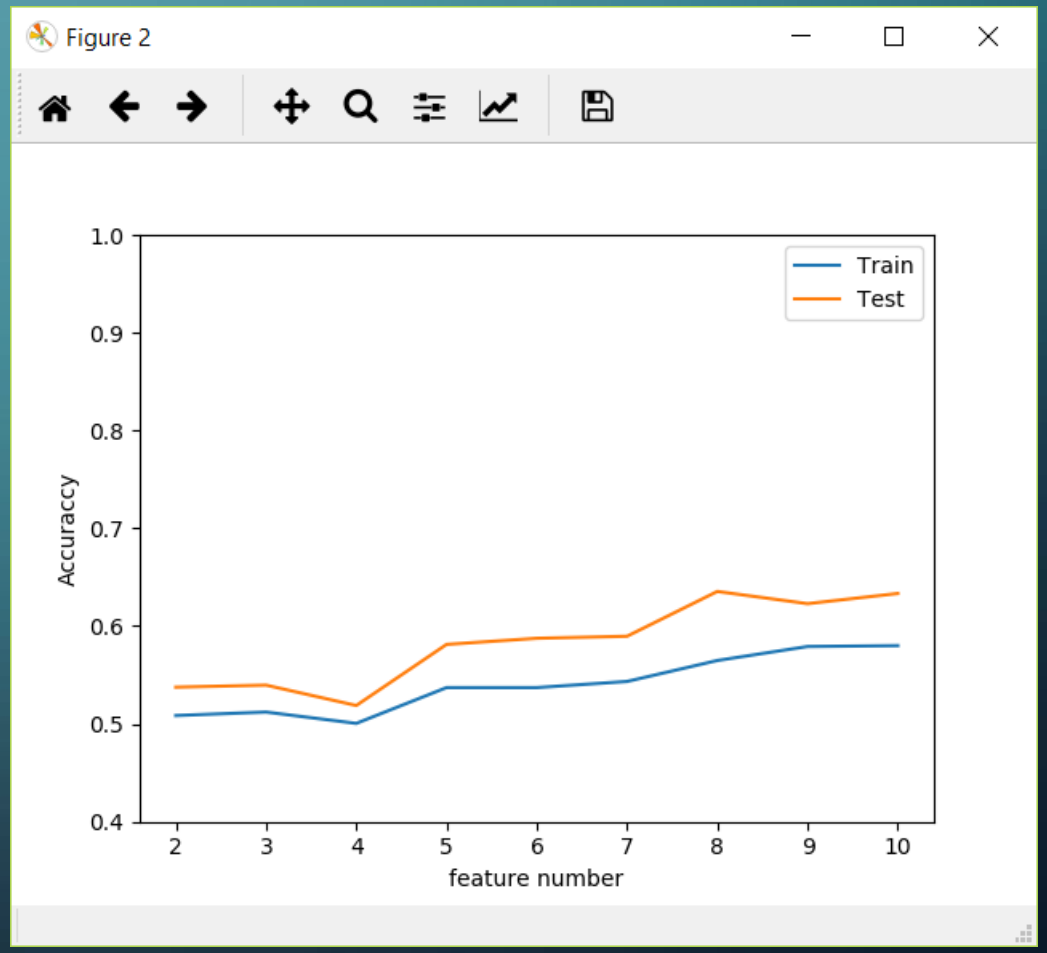
- Differences between extraction and selection
  - **Extraction**: Input data are transformed into another space, i.e. the description of data sample are changed, beforehand. Dimension reduction is done in the transformed space. The key concept is: Keep the most information of data set while try to reduce the factors to describe the data set.
  - **Selection**: Directly remove some factors from the description of data sample.

EXAMPLE: PRINCIPLE COMPONENT ANALYSIS

- Differences between extraction and selection
  - **Extraction**: Input data are transformed into another space, i.e. the description of data sample are changed, beforehand. Dimension reduction is done in the transformed space. The key concept is: Keep the most information of data set while try to reduce the factors to describe the data set.
  - **Selection**: Directly remove some factors from the description of data sample.

EXAMPLE: PRINCIPLE COMPONENT ANALYSIS





NEXT …

- When talking about machine learning, there are several independent facets for discussion
  - **Learning strategy:** Supervised, unsupervised, and reinforcement learning
  - **Task:** classification and prediction
  - **Model:** Generative or discriminative model. For the former, the distribution of either input or output is assumed.