Final_Project_Proposal

September 29, 2019

Course: ***DATA 1030***

Author: ***Cheng Zeng***

Project description

The final project lives on the github (see final project repo). The directory structure looks like this

```
.
  data
     processed
     raw
  figures
  models
  proposal
     refs
  reports
  src
      features
      models
      visualization
```

It aims to solve a regression problem. Given several features of a building, we hope to find a regression model which can predict the energy efficiency of it. The energy efficiency are described by two target variables names *heating load* and *cooling load*. It is important because it could help to design more environemntally friendly buildings from the theoretical perspective. Real-life energy efficiency experiments of buidings could be both time and effort consuming.

Questions to be answered

- Regression questions
  - Predict the heating load and cooling load of the building
  - Estimate the importance of the features
- Classification quesitons
  - Convert the output variables into a high-level concept such as low-, medium- and high-enregy efficiency and predict what kind of buildings can be classified into certain groups.

- Unsupervised quesitons
  - Using the first nine features and group them into the similar buildings and test if buildings in the same cluster also performs similarly on the 10 attribute

Dataset

The dataset if from UCI energy efficiency. In this dataset, **768** simulation points are provided. **10 numerical attributes** are used to represent the main features of each building.

Two related public projects

1. The first project related to the dataset is described in this work https://www.sciencedirect.com/science/article/pii/S037877881200151X. They collected the data by computational simulation of buildings which are assumed to be in Athens, Greece. Classical linear regression model is compared to a non-parametric non-linear model, random forest, to estimate and predict the heating and cooling loadings of the buildings.
2. The second project lives in https://doi.org/10.1016/j.enbuild.2014.07.036, in which artificial neural network, support vector regression and ensemble inference models are employed and it concludes that the ensemble model displays superior performance than other existing models for this dataset.

Attribute description Of the 10 attributes, **8** attributes serve as the dependent variables and the other **2** are response variables to be predicted. In the following table the 10 features of the buildings, the number of possible values and the units/description are detailed.

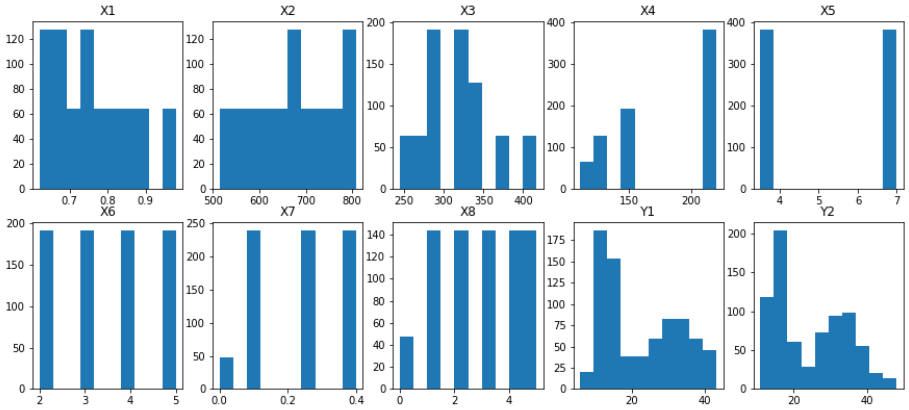
Mathematical symbols	Attribute	# possible values	units/description
X1	Relative compactness	12	absolute unit
X2	Surface area	12	m^2
X3	Wall area	7	m^2
X4	Roof area	4	m^2
X5	Overall height	2	m
X6	Orientation	4	2:North, 3:East, 4:South, 5:West
X7	Glazing area	4	0%, 10%, 25%, 40% (of floor area)

Mathematical symbols	Attribute	# possible values	units/description
X8	Glazing area distribution	5	0:Unknow, 1:Uniform, 2:North, 3:East, 4:South, 5:West
y1	Heating load	586	kWh/m^2
y2	Cooling load	636	kWh/m^2

To make it more clear, the meaning of target values are elaborated. * heating load: the amount of heat energy that would need to be added to a space to maintain the temperature in an acceptable range. * cooling load: the amount of heat energy that would need to be removed from a space (cooling) to maintain the temperature in an acceptable range.

For the mathematical convention, X is used to denote the input variables and y is for output variables. One thing to be noted it that it is not a real-life dataset. It comes from building simulations.

The histograms of the ten attributes are shown in the plot



Dataset preprocessing

The values in the original dataset are all numerical. It can be as it is for learning a regresssion model. However, if one needs to do the classification, features such as **orientation (X6)**, **glazing area (X7)** and **glazing area distribution (X8)** need to be converted to categorical features.

As of now, we will do the preprocessing for the regression problem and classification problem, respectively. The current preprocessing is evaluated on all the dataset.

In the next steps the dataset will be splited to training, cross-validation and testing set.

The jupyter notebook for proprocessing and the resulting preprocessed dataset is in the folder "./data/processed".

- The dataset for regression problem is named "df_preprocessed_regression.csv"
- The dataset for classification problem corresponds to "df_preprocessed_classification.csv"

In terms of the regression problem, from the histograms one can see that all the numerical values are bounded. Therefore the `MinMaxScaler` scaler is used for all the features. The preprocessed dataset still has ten features.

In regard to the classification problem, since the attributes **orientation**, **glazing area** and **glazing area distribution** cannot be ordered, `OneHotEncoder` is applied. In terms of the other 5 exactly numerical attributes, `MinMaxScaler` is employed. The two output variables are firstly transformed to two group of classes ('low efficiency' and 'high efficiency') using the mean values of the dataset as the criterion, then they are proprocessed with `LabelEncoder`. The preprocessed data in total has **18** features.