BINOMIAL_30_MAKAROFF_PAVARD_part2

Nicolas Makaroff

08/12/2019

Train-Test split

Before doing anything, we voluntarily set aside part of the data to create a test set without looking at it. This way we won't stumbled upon some seemingly pattern and do data snooping which could leads us to focus on a particular Learning Algorithm and alter our prediction.

The data provided comes from 12 stations in Beijing, China. The city is wide and resuming all sets into one would affect the features available as the weather can be complitely different in different part of the city. Hence we will lower our study in 3 sectors : North, Center and South Beijing following the representation in the study [**1**]

We will conduct two types models in this project.

The first one is the most obvious : a regression. For this one, as a result on our study of the data, it emerged that the data are time-series. Time series are a specific problem of machine learning as the data aren't independant. They are correlated throught time. As such we will keep the data provided for the year between 2013/03 and 2015/02 and set aside 2016/03-2017/02 as the test set.

The second one is a way of going over the fact that the data is a time-series. As a fact, what most people are interesting in is if the air-pollution is a danger to them or not. We will look frame the problem as a classification-problem where the class will be set a from good to bad air-quality. We will in a more classic way select 20% of data randomly to create our test set.

For the following parts, we use the same strategy on both train and test set.

Preparing the data for learning algorithm

Most Machine Learning algorithm cannot work with missing features. We must take care of the missing data. Our plan for the first modelisation doesn't allow us to just drop rows containing missing features. Therefore, we will use an interpolation between the points where there is some missing data to approximate their values.

As we saw there is categorical attributes in the data, for example *WD* the wind direction. We transform the wind direction to map the trigonometrical cercle to keep a strong link between the wind direction and the new representation as numerical. After some research on weather link to air-pollution, we came to the conclusion that while the temperature is important, it isn't because it's cold outside that the pollution will be higher. The same goes for *DEWP* and most meteorogical paper doesn't use it but use instead humidity. Following [**2**], we transformed *DEWP* and *TEMP* to get humidity from them.Time is also an important feature but the learning algorithm won't understand that 12 comes before 1 and the same goes for december and january. To solvethis we do as for the wind direction and map them to the trigonometrical cercle.

Every features where then scaled using a standardization.

Considering our time series and only using a linear regression or variant of it forces us to find ways not to ignore previous data that are more than precious. To solve this we create new features from copies of the others by shifting them on the time line, technics use in [**3**].

For the classification we transformed using [**4**], giving standards for particule pollution and the AQI (Air Quality Index). We tried for this modelisatoin not two take into account previous values on the time line.

Select and train a model

To get an insight on our model we look at the AIC and try to minimize it.

To select our model we will act as the following :

- Create a base-model to have a starting quality
- Add relevant variables interaction (done in most parts previously)
- Remove insignificant variables using p-values and AIC improvement
- Remove any outlying datapoints
- Finally evaluate on the test set

This step will be completed with the use of Lasso and Ridge regression. Using a Ridge regression will force the learning algorithm to fit the data but also to keep the model weights as small as possible. Constraining the weights helps reguralizing the model. Opposite to the Ridge regression, using a lasso regression will tend to eliminate the weights of the least important features. We will prefer Lasso over Ridge regression if we suspect that only a few features are actually useful. Finally, we will prefer Elastic Net which is the middle ground between Lasso and Ridge because Lasso may behave erractically since several features are strongly correlated (the new features introducted).

Furthermore, to see if we can get better result testing an other model, we will try a K-NearestNeighbor regression. To select the best k hyperparameter we'll use a grid search.

The difference, attain model selection between our two modelisation, will be the class of the problem, switching from regression to classification.

Evaluating the selected model

To evaluate our model, we will train it using cross validation and we will use mean squared error and root mean squared error a since both aren't acting in a similar manner with respect to outliers, to evaluate the model on the test set and get its generalization power. Out of the training process, we will use a R2 score to see its performances.

Bibliography

[**1**] Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. and Chen, S.X. (2017) Cautionary Tales on Air-Quality Improvement in Beijing. Proceedings of the Royal Society A, Volume 473, No. 2205, Pages 20170457.

[**2**] Alduchov, O.A. and R.E. Eskridge, 1996: Improved Magnus Form Approximation of Saturation Vapor Pressure. J. Appl. Meteor., 35, 601–609, https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2

[**3**] Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, Somayeh Talebiesfandarani, PM2.5 Prediction Based on Random Forest, XGBoost,and Deep Learning Using Multisource Remote Sensing Data, 4 July 2019, Atmosphere 2019, 10, 373;

[**4**] The National Ambient Air Quality Standards for Particle Pollution, REVISED AIR QUALITY STANDARDS FOR PARTICLE POLLUTION AND UPDATES TO THE AIR QUALITY INDEX (AQI), 2016, https://www.epa.gov/sites/production/files/2016-04/documents/2012_aqi_factsheet.pdf