glm

November 18, 2019

[1]:
```
%%html
<style>
.container{width: 100%}
</style>
```

```
<IPython.core.display.HTML object>
```

[2]:
```
%load_ext autoreload
%autoreload 2
```

[3]:
```
import warnings
warnings.filterwarnings("ignore")
```

[4]:
```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
%matplotlib inline
```

[5]:
```
import os
os.sys.path.insert(0, "../")
```

0.0.1 Load Data

[6]:
```
from tools import load_boston
data, desc = load_boston("../data_base")
data = data.rename(columns = {"target": "MEDV"})
features = data.drop("MEDV", axis = 1)
prices = data.MEDV
```

[7]:
```
print(desc)
```

```
.. _boston_dataset:

Boston house prices dataset
---------------------------
```

Data Set Characteristics:

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - CRIM   per capita crime rate by town
        - ZN   proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS   proportion of non-retail business acres per town
        - CHAS   Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
        - NOX   nitric oxides concentration (parts per 10 million)
        - RM   average number of rooms per dwelling
        - AGE   proportion of owner-occupied units built prior to 1940
        - DIS   weighted distances to five Boston employment centres
        - RAD   index of accessibility to radial highways
        - TAX   full-value property-tax rate per $10,000
        - PTRATIO pupil-teacher ratio by town
        - B   1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT   % lower status of the population
        - MEDV   Median value of owner-occupied homes in $1000's

    :Missing Attribute Values: None

    :Creator: Harrison, D. and Rubinfeld, D.L.

This is a copy of UCI ML housing dataset.
https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.   Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley, 1980.   N.B. Various transformations are used in the table on pages 244-261 of the latter.

The Boston house-price data has been used in many machine learning papers that address regression
problems.

.. topic:: References

```
   - Belsley, Kuh & Welsch, 'Regression diagnostics: Identifying Influential
Data and Sources of Collinearity', Wiley, 1980. 244-261.
   - Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In
Proceedings on the Tenth International Conference of Machine Learning, 236-243,
University of Massachusetts, Amherst. Morgan Kaufmann.
```

0.0.2 Making a model

Train Test Splitting

[8]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(features, prices, test_size␣
 ↪= .3, random_state = 42)
```

OLS Benchmark

[9]:
```
from tools import cal_benchmark_perf
benchmark_perf = cal_benchmark_perf(X_train, y_train, X_test, y_test)
```

GLM

Gamma

[10]:
```
import statsmodels.api as sm
```

[11]:
```
gamma_mod = sm.GLM(y_train, X_train, family = sm.families.Gamma())
regr = gamma_mod.fit()
```

```
C:\ProgramData\Anaconda3\lib\site-
packages\statsmodels\genmod\generalized_linear_model.py:273: DomainWarning: The
inverse_power link function does not respect the domain of the Gamma family.
  DomainWarning)
```

[12]:
```
from tools import cal_benchmark_perf
```

[13]:
```
benchmark_perf = cal_benchmark_perf(X_train, y_train, X_test, y_test)
```

[14]:
```
from sklearn.metrics import mean_squared_error

y_test_hat = regr.predict(X_test)
mean_squared_error(y_test_hat, y_test) / benchmark_perf
```

[14]:
```
0.6267896342164393
```

[ ]:
```

```