Assignment_2_linreg

January 29, 2020

[5]:
```
from sklearn.datasets import fetch_california_housing
raw = fetch_california_housing()

X = raw.data
y = raw.target
# Show feature names
```

[5]:
```
['MedInc',
 'HouseAge',
 'AveRooms',
 'AveBedrms',
 'Population',
 'AveOccup',
 'Latitude',
 'Longitude']
```

[11]:
```
# Show dataset description
```

[11]:
```
'.. _california_housing_dataset:\n\nCalifornia Housing
dataset\n--------------------------\n\n**Data Set Characteristics:**\n\n
:Number of Instances: 20640\n\n    :Number of Attributes: 8 numeric, predictive
attributes and the target\n\n    :Attribute Information:\n        - MedInc
median income in block\n        - HouseAge      median house age in block\n
- AveRooms      average number of rooms\n        - AveBedrms     average number
of bedrooms\n        - Population    block population\n        - AveOccup
average house occupancy\n        - Latitude      house block latitude\n        -
Longitude     house block longitude\n\n    :Missing Attribute Values:
None\n\nThis dataset was obtained from the StatLib
repository.\nhttp://lib.stat.cmu.edu/datasets/\n\nThe target variable is the
median house value for California districts.\n\nThis dataset was derived from
the 1990 U.S. census, using one row per census\nblock group. A block group is
the smallest geographical unit for which the U.S.\nCensus Bureau publishes
sample data (a block group typically has a population\nof 600 to 3,000
people).\n\nIt can be downloaded/loaded using
the\n:func:`sklearn.datasets.fetch_california_housing` function.\n\n.. topic::
References\n\n    - Pace, R. Kelley and Ronald Barry, Sparse Spatial
Autoregressions,\n      Statistics and Probability Letters, 33 (1997) 291-297\n'
```

[12]:
```
# Show dimension of X
```

[12]:
```
(20640, 8)
```

[13]:
```
# Show dimension of y
```

[13]:
```
(20640,)
```

[36]:
```
# Split X, y into X_train, X_test, y_train, y_test with 7:3 ratio
```

[37]:
```
# Build a linear regression model with X_train, y_train
```

[37]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

[38]:
```
# y_pred from X_test
```

[38]:
```
array([ 4.34745277e-01,  9.66898932e-03, -1.05473236e-01,  6.37126713e-01,
       -5.78985090e-06, -3.19937113e-03, -4.27913326e-01, -4.40268982e-01])
```

[39]:
```
# find the argmax of coefficients
```

[39]:
```
3
```

[42]:
```
# Draw scatter plots of
#             argmax of X_train - y_train as 'x' marker
#             argmax of X_test - y_test as 'o' marker
```

[42]:
```
(2, 10)
```