Machine Learning With Python: Linear Regression Multiple Variables

Sample problem of predicting home price in monroe, new jersey (USA) ¶

Below is the table containing home prices in monroe twp, NJ. Here price depends on **area (square feet), bed rooms and age of the home (in years)**. Given these prices we have to predict prices of new homes based on area, bed rooms and age.
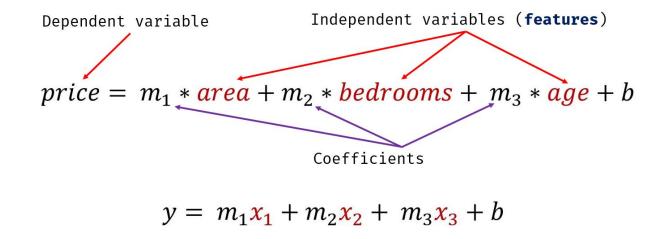
area	bedrooms	age	price
2600	3	20	550000
3000	4	15	565000
3200		18	610000
3600	3	30	595000
4000	5	8	760000
4100	6	8	810000

Given these home prices find out price of a home that has,

3000 sqr ft area, 3 bedrooms, 40 year old

2500 sqr ft area, 4 bedrooms, 5 year old

We will use regression with multiple variables here. Price can be calculated using following equation,

Dependent variable   Independent variables (**features**)

$$price = m_1 * area + m_2 * bedrooms + m_3 * age + b$$

Coefficients

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + b$$

Here area, bedrooms, age are called independant variables or **features** whereas price is a dependant variable

```
In [1]:  import pandas as pd
         import numpy as np
         from sklearn import linear_model
```

```
In [2]:  df = pd.read_csv('homeprices.csv')
         df
```

Out[2]:

	area	bedrooms	age	price
0	2600	3.0	20	550000
1	3000	4.0	15	565000
2	3200	NaN	18	610000
3	3600	3.0	30	595000
4	4000	5.0	8	760000
5	4100	6.0	8	810000

Data Preprocessing: Fill NA values with median value of a column

```
In [3]:  df.bedrooms.median()
```

Out[3]: 4.0

```
In [5]:  df.bedrooms = df.bedrooms.fillna(df.bedrooms.median())
         df
```

Out[5]:

	area	bedrooms	age	price
0	2600	3.0	20	550000
1	3000	4.0	15	565000
2	3200	4.0	18	610000
3	3600	3.0	30	595000
4	4000	5.0	8	760000
5	4100	6.0	8	810000

```
In [6]:  reg = linear_model.LinearRegression()
         reg.fit(df.drop('price',axis='columns'),df.price)
```

Out[6]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                  normalize=False)

In [7]:
```
reg.coef_
```

Out[7]: `array([  112.06244194, 23388.88007794, -3231.71790863])`

In [8]:
```
reg.intercept_
```

Out[8]: `221323.00186540408`

Find price of home with 3000 sqr ft area, 3 bedrooms, 40 year old

In [9]:
```
reg.predict([[3000, 3, 40]])
```

Out[9]: `array([498408.25158031])`

In [10]:
```
112.06244194*3000 + 23388.88007794*3 + -3231.71790863*40 + 221323.00186540384
```

Out[10]: `498408.25157402386`

Find price of home with 2500 sqr ft area, 4 bedrooms, 5 year old

In [11]:
```
reg.predict([[2500, 4, 5]])
```

Out[11]: `array([578876.03748933])`

Exercise

In exercise folder (same level as this notebook on github) there is **hiring.csv**. This file contains hiring statics for a firm such as experience of candidate, his written test score and personal interview score. Based on these 3 factors, HR will decide the salary. Given this data, you need to build a machine learning model for HR department that can help them decide salaries for future candidates. Using this predict salaries for following candidates,

2 yr experience, 9 test score, 6 interview score

12 yr experience, 10 test score, 10 interview score

Answer

53713.86 and 93747.79