Lab: Simple linear regression

In this lab, you will load data, plot data, perform simple mathematical manipulations, and fit a simple linear regression model. Before doing this lab, you can go through the demo (./demo2_auto_mpg.ipynb) to see an example of these operations on an automobile dataset. The lab use the Boston housing data set, a widely-used machine learning data set for illustrating basic concepts.

Loading the data

The Boston housing data set was collected in the 1970s to study the relationship between house price and various factors such as the house size, crime rate, socio-economic status, etc. Since the variables are easy to understand, the data set is ideal for learning basic concepts in machine learning. The raw data and a complete description of the dataset can be found on the UCI website:

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names)

In the lab, you will complete all the code marked `TODO`.

First, complete the following code that uses the `pd.read_csv` command to read the data from the file located at

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data)

I have supplied a list `names` of the column headers. You will have to set the options in the `read_csv` command to correctly delimit the data in the file and name the columns correctly.

In [1]:

```
import pandas as pd
import numpy as np
names =[
    'CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM',
    'AGE',  'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'PRICE'
]
```

Display the first six rows of the data frame

In [2]:

```python
df = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/hou
sing/housing.data", delim_whitespace=True, names=names, index_col=False, engine
="python")
df.head(6)
```

Out[2]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	39
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	39
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	39
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	39
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	39
5	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222.0	18.7	39

Basic Manipulations on the Data

What is the shape of the data? How many attributes are there? How many samples? Print a statement of the form:

```
num samples=xxx, num attributes=yy
```

In [3]:

```python
# len rows == num samples
samples = df.shape[0]

# num cols == num attr
attrs = df.shape[1]

print("num samples=" + str(samples) + ", num attributes=" + str(attrs))
```

```
num samples=506, num attributes=14
```

Create a response vector `y` with the values in the column `PRICE`. The vector `y` should be a 1D `numpy.array` structure.

In [4]:

```python
y = np.array(df['PRICE'])
# print(y)
```

Use the response vector `y` to find the mean house price in thousands and the fraction of homes that are above $40k. (You may realize this is very cheap. Prices have gone up a lot since the 1970s!). Create print statements of the form:

```
The mean house price is xx.yy thousands of dollars.
Only x.y percent are above $40k.
```

In [5]:

```
price_avg = np.mean(y)
print("The mean house price is", round(price_avg, 2), "thousands of dollars.")

percent = np.mean(y > 40) * 100
print("Only", round(percent, 1), "percent are above $40k.")
```

```
The mean house price is 22.53 thousands of dollars.
Only 6.1 percent are above $40k.
```

Visualizing the Data

Python's `matplotlib` has very good routines for plotting and visualizing data that closely follows the format of MATLAB programs. You can load the `matplotlib` package with the following commands.

In [6]:

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

Similar to the `y` vector, create a predictor vector `x` containing the values in the `RM` column, which represents the average number of rooms in each region.

In [7]:

```
x = np.array(df["RM"])
# print(x)
```

Create a scatter plot of the price vs. the `RM` attribute. Make sure your plot has grid lines and label the axes with reasonable labels so that someone else can understand the plot.

In [8]:

```
plt.scatter(x, y)
plt.title("Price v. Total Rooms in 1970s Homes")
plt.xlabel("Rooms")
plt.ylabel("Price (in thousands)")
plt.grid()
plt.show()
```

Fitting a Simple Linear Model

We will write a simple function to perform a linear fit. Use the formulae given in the class, to compute the parameters β_0, β_1 in the linear model

$$y = \beta_0 + \beta_1 x + \epsilon$$

as well as the coefficient of determination R^2.

In [9]:

```
def fit_linear(x,y):
    """
    Given vectors of data points (x,y), performs a fit for the linear model:
       yhat = beta0 + beta1*x,
    The function returns beta0, beta1 and rsq, where rsq is the coefficient of
 determination.
    """

    # sample means
    xm = np.mean(x)
    ym = np.mean(y)

    sxx = np.mean((x - xm)**2)
    syy = np.mean((y - ym)**2)
    sxy = np.mean((y - ym)*(x-xm))

    beta1 = sxy / sxx
    beta0 = ym - (beta1 * xm)

    yhat = (beta0 + beta1 * x)
    rss = np.mean((y - yhat)**2)
    rsq = (syy - rss) / syy

    return beta0, beta1, rsq
```

Using the function `fit_linear` above, print the values `beta0`, `beta1` and `rsq` for the linear model of price vs. number of rooms.

In [10]:

```
beta0, beta1, rsq = fit_linear(x, y)
print(beta0, beta1, rsq)
```

```
-34.67062077643857 9.10210898118031 0.4835254559913342
```

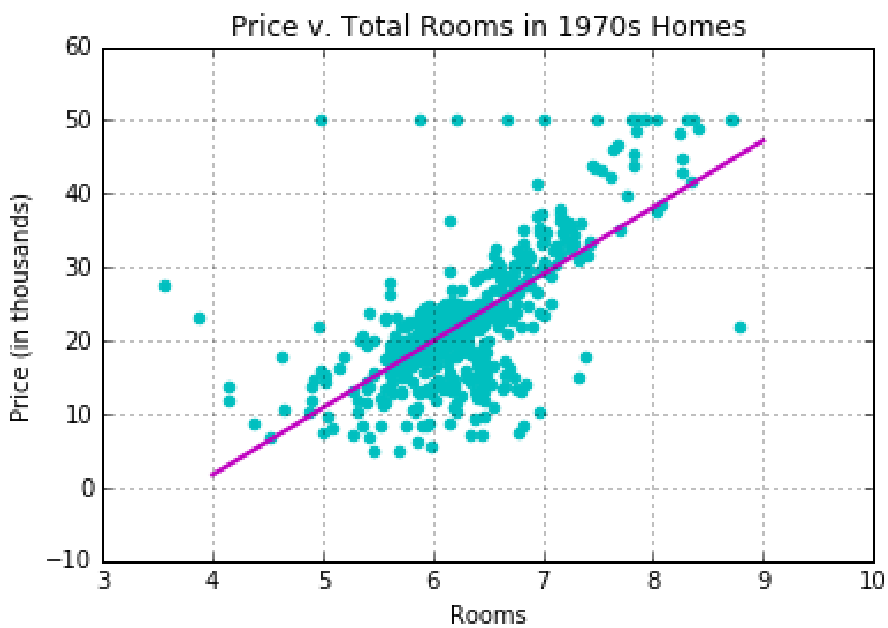
Replot the scatter plot above, but now with the regression line. You can create the regression line by creating points `xp` from say 4 to 9, computing the linear predicted values `yp` on those points and plotting `yp` vs. `xp` on top of the above plot.

In [11]:

```
# data points
plt.scatter(x, y, color = 'c')

# regression
xplt = np.array([4, 9])
yplt = beta1 * xplt + beta0
plt.plot(xplt, yplt, linewidth = 2, color = 'm')

# formatting
plt.title("Price v. Total Rooms in 1970s Homes")
plt.xlabel("Rooms")
plt.ylabel("Price (in thousands)")
plt.grid()
plt.show()
```



Compute coefficients of determination

We next compute the R^2 values for all the predictors and output the values in a table. Your table should look like the following, where each the first column is the attribute name and the second column is the R^2 value.

```
CRIM        0.151
ZN          0.130
INDUS       0.234
...         ...
```

To index over the set of colunms in the dataframe `df`, you can either loop over the items in the `names` lists (skipping over the final name `PRICE`) or loop over integer indices and use the method, `df.iloc`.

In [14]:

```
rsqs = {}
for name in names:
    if name is not "PRICE":
        x = df[name]
        rsq = fit_linear(x, y)[2]
        rsqs[name] = rsq

rsq_table = pd.DataFrame.from_dict(rsqs, orient='index', columns=['VALUE'])
rsq_table.head(14)
```

Out[14]:

	VALUE
B	0.111196
RM	0.483525
TAX	0.219526
CRIM	0.150780
CHAS	0.030716
INDUS	0.233990
RAD	0.145639
AGE	0.142095
DIS	0.062464
NOX	0.182603
ZN	0.129921
LSTAT	0.544146
PTRATIO	0.257847