Introduction to Machine Learning

Fall 2019

University of Science and Technology of China

Lecturer: Jie Wang

Name: Bowen Zhang

ID: PB17000215

Homework 1

Posted: Seq. 20, 2019

Due: Seq. 30, 2019

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Linear regression 20pts

Given a data set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R}$.

1. If we want to fit the data by a linear model

$$y = w_0 + w_1 x, \tag{1}$$

please find $\hat{w}_0$ and $\hat{w}_1$ by the least squares approach (you need to find expressions of $\hat{w}_0$ and $\hat{w}_1$ by $\{(x_i, y_i)\}_{i=1}^n$, respectively).

2. **Programming Exercise** We provide you a data set $\{(x_i, y_i)\}_{i=1}^{30}$. Consider the model in (1) and the one as follows:

$$y = w_0 + w_1 x + w_2 x^2. \tag{2}$$

Which model do you think fits better the data? Please detail your approach first and then implement it by your favorite programming language. The required output includes

   (a) your detailed approach step by step;

   (b) your code with detailed comments according to your planned approach;

   (c) a plot showing the data and the fitting models;

   (d) the model you finally choose [$\hat{w}_0$ and $\hat{w}_1$ if you choose the model in (1), or $\hat{w}_0$, $\hat{w}_1$, and $\hat{w}_2$ if you choose the model in (1)].

Solution:

1. 令误差函数

$E = \sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)^2$

$(w_0^*, w_1^*) = \arg\min_{(w_0,w_1)}(E)$

$\frac{\partial E}{\partial w_0} = 2\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)$

$\frac{\partial E}{\partial w_1} = 2\sum_{i=1}^{n} x_i(w_0 + w_1 x_i - y_i)$

要使得 E 最小，所以令两个偏导数为 0，那么可得

$nw_0 = \sum_{i=1}^{n}(w_1 x_i - y_i)$

$w_0 \sum_{i=1}^{n} x_i = n\sum_{i=1}^{n} x_i(w_1 x_i - y_i)$

联立方程解得

$w_0 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$

$w_1 = \frac{\sum_{i=1}^{n}(y_i - w_0 x_i)}{n}$

2.

(a) 主要思路是由 1. 中已经求得的 $w0, w1$ 表达式可以通过样本直接计算出对应的 $w0, w1$，再重新代入误差函数 E 中，求出最小误差。
同样的道理，二次拟合函数也可以通过计算其误差函数，对三个参数分别求偏导数并令其为 0，联立三个方程解得相应的三个参数，并代回到误差函数中求得最小误差。
以上两个最小误差进行比较即可看出哪个模型对数据的拟合较好。
具体到编程细节，首先读入数据并转换为 numpy array，分别进行线性拟合和二次拟合（其中二次拟合使用 numpy 求解线性方程组），求得最小误差后进行比较，并绘制图像。
参考文献：计算方法课程

(b) 代码见附件prob1.py。

(c) 图片见附件Figure.png。

(d) 最终比较之后选择二次拟合模型，得到 $w0 = 1.0295683746564661, w1 = 0.3861433340323225, w2 = -0.14215111308616024$。

■