Machine Learning, Spring 2018
Homework 2

Xiaohe He(48655395)

1 The relationship between the maximum likelihood and distance metric

Suppose $y = f(\theta; \mathbf{x})$ is our model (such as classifier or regression model), $\mathbf{x} \in \Re^n$ is feature, $y \in \Re$ is response, and $\theta \in \Re^K$ is the parameters of model f. As you know, the standard processing in machine learning is: We first collocate train samples $\{(\mathbf{x}_i, y_i)\}, i = 1, 2, \cdots, n$, and some loss function will be defined, then we can find the optimal (or suboptimal) θ^* by minimizing the loss function. Mean square error (MSE) is a common loss function.

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2.$$

Suppose $y = f(\theta; \mathbf{x}) + \varepsilon$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

(1) Write down the log likelihood function of θ;(5 points)

(2) Use the maximum likelihood principle to explain why MSE is a good loss function.(10 points)

(3) Explain why MSE is not robust to outliers in Homework 1.(10 points)

Solution:

(1) Since $y = f(\theta; \mathbf{x}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $E = \frac{1}{n}\sum\limits_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$. We colud easy to get

$$Pr(Y_i|\boldsymbol{x_i}) \sim \mathcal{N}(\boldsymbol{x_i^T\theta}, \sigma^2)$$

$$L_n(\boldsymbol{x_i^T\theta}; y|x) = \prod_{i=1}^{n} f_{Y|x}(y_i|\boldsymbol{x_i}; \boldsymbol{x_i^T\theta})$$

$\Rightarrow$

$$l_n(\boldsymbol{x_i^T\theta}; y|x) = -\frac{m}{2}ln\sigma^2 - \frac{m}{2}ln(2\pi) - \frac{1}{\sigma^2}\frac{1}{2}||\boldsymbol{X\theta} - \boldsymbol{y}||^2$$

(2) log is a strictly increasing function.
The maximum of $l_n(\boldsymbol{x_i^T\theta}; y|x)$ and the minimum of $||\boldsymbol{X\theta} - \boldsymbol{y}||^2$ are in the same point.

(3) As in MSE, the points which are more far away from the hyperplane have more right to influence the results than the close ones. Outliesrs have more right to make resluts worse.

2 Linear Regression via Gradient Descent Method

The case study "**GPA.txt**" contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Question:

(1) Formulate this problem with the linear regression and give its expression. Give the expression of the cost function $J(\boldsymbol{\theta})$. (5 points)

(2) Use the stochastic gradient descent method to solve this linear regression problem, and show your termination criterion in the report. Your termination criterion is required to ensure the convergence of the algorithm. (15 points)

(3) Plot both your fitted curve and the convergence result. (10 points)

Please finish your simulation with MATLAB/Python and compress your codes into one file and sent it to TAs. **(Do not use any existing solvers. Add annotations to your code, if your code is poorly readable, you won't get the points!)**

Notice: In pseudocode, the stochastic gradient descent method can be presented as follows, where the data is shuffled for each pass to prevent cycles.

Algorithm 1 Stochastic Gradient descent

Given the desired accuracy ϵ.
Initialize the parameter $\boldsymbol{\theta}$ and the learning rate α.
repeat
  Randomly shuffle examples in the training set.
  for $i = 1, \cdots, n$ **do**
    Update $\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha \nabla J_i(\boldsymbol{\theta})$.
  end for
until *Your termination criterion.*
return $\boldsymbol{\theta}$.

Solution:

(1) Denote $x = (x_1, \cdots, x_n)^T$ and $y = (y_1, \cdots, y_n)^T$ as the high_GPA corresponding univ_GPA, respectively.

$$h_i(\theta) = \theta_1 x_i + \theta_0, \quad where\ \theta = (\theta_0, \theta_1)^T \quad and\ i = 1, \cdots, n;$$

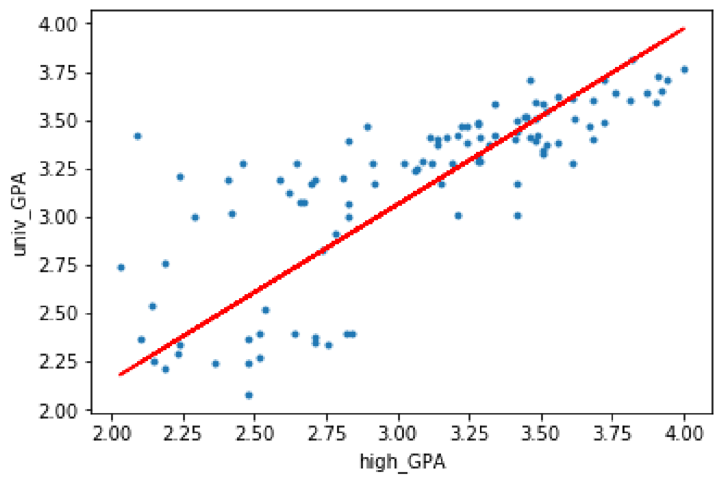
$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (h_i(\theta) - y_i)^2$$

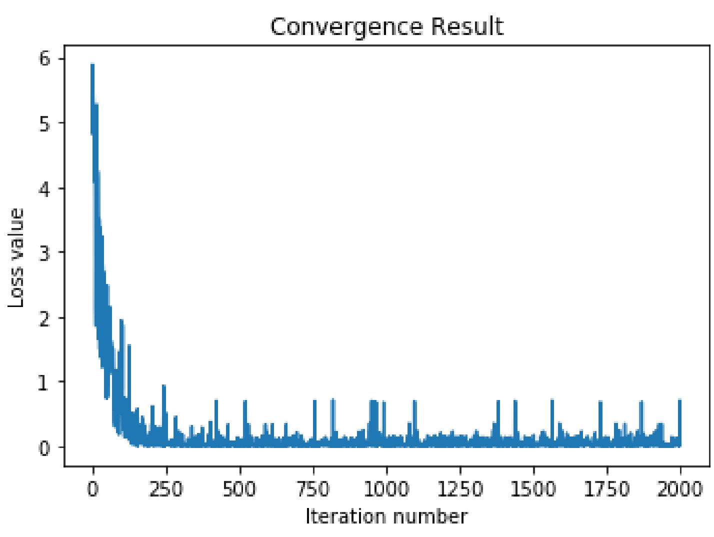
(2) My termination criterion is 2000 iterates, each iterate use 1 samples. And Learning Rate is 0.001.

(3) Convergence result:

$$univ\_GPA = 0.921791 high\_GPA + 0.334685$$

Fitted curve and convergence result are as follows:





3 Multivariable Linear Regression

Please refer to the data **car-data.xls**, there is a representative sample of cars' attributes and price. The columns in the table are described as:

Price: suggested retail price of the car

Mileage: number of miles the car has been driven

Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet

Model: specific models for each car manufacturer such as Ion, Vibe, Cavalier

Trim (of car): specific type of car model such as SE Sedan 4D, Quad Coupe 2D

Type: body type such as sedan, coupe, etc.

Cylinder: number of cylinders in the engine

Liter: a more specific measure of engine size

Doors: number of doors

Cruise: indicator variable representing whether the car has cruise control (1 = cruise)

Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)

Leather: indicator variable representing whether the car has leather seats (1 = leather)

Notice: Please finish your simulation with MATLAB/Python and compress your codes into one file and sent it to TAs. You should also write your answer to the questions below. **(Do not use any existing solvers. Add annotations to your code, if your code is poorly readable, you won't get the points!)**

(1) Find the linear regression equation for Mileage vs Price.

  (i) Write the linear regression equation, plot the data and your equation. (5 points)

  (ii) How to determine whether your equation is a good fit for this data? Find it and evaluate it. (Key: You may like to use R^2 score, 0.8 and greater is considered a strong correlation.) (5 points)

  (iii) Does the Mileage affect the price? How big is the impact? Why? (5 points)

(2) Use Mileage, Cylinders, Liters, Doors, Cruise, Sound, and Leather to find the linear regression equation about price.

  (i) Write the linear regression equation. (10 points)

  (ii) How to determine whether your equation is a good fit for this data? Find it and evaluate it. (10 points)

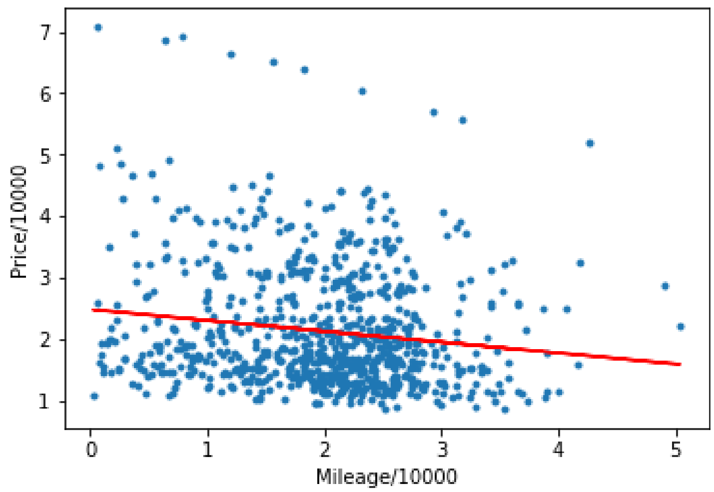
  (iii) Find the combination of the factors that is the best predictor for Price. (10 points)

Solution:

(1) (i) Linear regression equation is:

$$Price = -0.155324 * Mileage + 24792.574497$$

Plot is as follows:



(ii)

$$R^2 \qquad = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2} \qquad = 0.020450253615 61381$$

Hence the equation is not a good fit for this data.

(iii) Mileage afect the price.
Points on the above of the pic. are all Cadillac XLR-V8. We colud easy to find out that they have a significant linear correlation.

(2) (i)

(ii)

(iii)