1 Linear Regression

1.1 What is Linear Regression?

If there is a "linear relationship" between two or more variables, then we can figure out the "routines" between variables through historical data and establish an effective linear model to predict the future variable results.

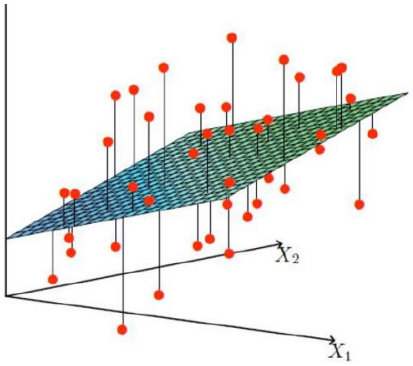
1.2 Example: Loan from bank

Assume that someone want to loan some from bank, their salary and age will influence the outcome of the bank loan, so the question is to predict how much can be loan for another person based on the following data.

Group	Salary ($)	Age	Loan ($)
1	4,000	25	20,000
2	8,000	30	70,000
3	5,000	28	35,000
4	7,500	33	50,000
5	12,000	40	85,000

- Data：salary, age (both two features influence the outcome of the bank loan)
- Aim：predict how much someone can load from bank? (Label)
- Consider：how much do salary and age influence the outcome of the bank loan? (parameters)

1.3 Build Model



We assume the model would be a plane as follow:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \tag{1}$$

Let's say $\theta = \begin{pmatrix}\theta_0\\ \theta_1\\ \theta_2\end{pmatrix}$, $x = \begin{pmatrix}1\\ x_1\\ x_2\end{pmatrix}$, we can rewrite the above equation as:

$$h_\theta(x) = (\theta_0, \theta_1, \theta_2)\begin{pmatrix}1\\ x_1\\ x_2\end{pmatrix} = \sum_{i=1}^{2} \theta_i x_i = \theta^T x \tag{2}$$

θ: parameters; $\boldsymbol{\theta_0}$**:** bias term; **x:** features;

i: no. of features or no. of parameters; $\boldsymbol{h_\theta(x)}$**:** predicted value

How about we get m samples and n features?

$$h_\theta(x) = (\theta_0, \theta_1, \theta_2, \dots, \theta_n)\begin{pmatrix}1\\ x_1\\ x_2\\ \vdots\\ x_n\end{pmatrix} = \sum_{i=1}^{m} \theta_i x_i = \theta^T x \tag{3}$$

1.4 Prediction Error

There must be a difference between the real value and the predicted value. The error is:

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)} \tag{4}$$

y: real value; $\boldsymbol{\varepsilon}$**:** error between real value and predicted value

Assume the error obeys the Gaussian Distribution: $\varepsilon^{(i)} \sim N(0, \sigma^2)$

$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}} \tag{5}$$

Plug Eq. (3) into Eq. (4):

$$p\left(y^{(i)}|x^{(i)};\,\theta\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y^{(i)}-\theta^{T}x^{(i)})^{2}}{2\sigma^{2}}} \tag{6}$$

1.5 Maximum Likelihood Estimation (MLE)

In order to solve what parameters are closest to the real value after combining with our data, we construct the maximum likelihood function:

$$\mathrm{L}(\theta)\;=\prod_{i=1}^{m}p\left(y^{(i)}|x^{(i)};\,\theta\right)=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y^{(i)}-\theta^{T}x^{(i)})^{2}}{2\sigma^{2}}} \tag{7}$$

m: no. of samples
Get logarithm for both sides of Eq. (4):

$$log\mathrm{L}(\theta)=log\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\left(y^{(i)}-\theta^{T}x^{(i)}\right)^{2}}{2\sigma^{2}}}=mlog\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{\sigma^{2}}\cdot\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\;\theta^{T}x^{(i)}\right)^{2} \tag{8}$$

In order to **maximize the likelihood function** (the same after the logarithmic transformation, we should ensure that **minimize the following formula (Loss Function) is minimal**:

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\;\theta^{T}x^{(i)}\right)^{2} \tag{9}$$

1.6 Solve Using Least Square Method

The object function is $J(\theta)=\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\;\theta^{T}x^{(i)}\right)^{2}$, we divide it by m and rewrite as matrix format:

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)}-\;\theta^{T}x^{(i)}\right)^{2}=\frac{1}{2m}(X\theta-y)^{T}(X\theta-y) \tag{10}$$

X: coefficient matrix(m × n)

$$\begin{gathered}h_{\theta}\left(x^{(1)}\right)=\theta_{0}+\;\theta_{1}x_{1}^{(1)}+\;\theta_{2}x_{2}^{(1)}+\cdots+\theta_{n}x_{n}^{(1)}\\ h_{\theta}\left(x^{(2)}\right)=\theta_{0}+\;\theta_{1}x_{1}^{(2)}+\;\theta_{2}x_{2}^{(2)}+\cdots+\theta_{n}x_{n}^{(2)}\\ \cdots\cdots\\ h_{\theta}\left(x^{(m)}\right)=\theta_{0}+\;\theta_{1}x_{1}^{(m)}+\;\theta_{2}x_{2}^{(m)}+\cdots+\theta_{n}x_{n}^{(m)}\end{gathered}$$

$$X=\begin{pmatrix}1 & x_{1}^{(1)} & x_{2}^{(1)} & \dots & x_{n}^{(1)}\\ 1 & x_{1}^{(2)} & x_{2}^{(2)} & \dots & x_{n}^{(2)}\\ \dots & \dots & \dots & \dots & \\ 1 & x_{1}^{(m)} & x_{2}^{(m)} & \dots & x_{n}^{(m)}\end{pmatrix}$$

$$\theta=\begin{pmatrix}\theta_{0}\\ \theta_{1}\\ \theta_{2}\\ \vdots\\ \theta_{n}\end{pmatrix}$$

Therefore, we can get:

$$h_{\theta}\left(x^{(i)}\right)=X\theta \tag{11}$$

Get the partial derivative:

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \left(\frac{1}{2}(X\theta - y)^T(X\theta - y)\right)m \\ &= \nabla_\theta \left(\frac{1}{2}(\theta^T X^T - y^T)(X\theta - y)\right)m \\ &= \nabla_\theta \left(\frac{1}{2}(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y)\right)m \\ &= \frac{1}{2}(X^T X\theta + (\theta^T X^T X)^T - X^T y - (y^T X)^T)m \\ &= (X^T X\theta - X^T y)m\end{aligned} \tag{12}$$

Set the partial derivative is equal to 0 and get θ:

$$(X^T X\theta - X^T y)m = 0$$

$$\theta = (X^T X)^{-1} X^T y \tag{13}$$

1.7 Solve Using Gradient Descent

Why do we use gradient descent?

① $X^T X$ is not always invertible.

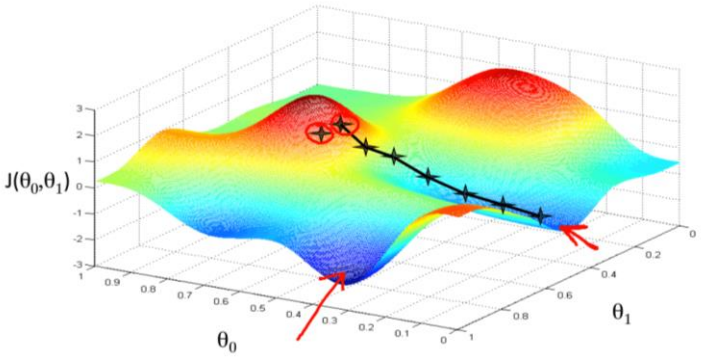
② When the data is large, the least square method takes a long time to solve. In general, if m is less than 10,000, you can use the analytic solution, and if m>10,000, you can use gradient descent. Of course this is not necessary, we need to understand why we do not choose the analytic solution, mainly because our computer is too slow (too complex, resulting in too long), this experience value can be adjusted as the computer gets faster and faster.

How do we solve using gradient descent?

For gradient descent with multiple variable, we need to update θ_j simultaneously.

j ∈[0, n]. n: no. of features.

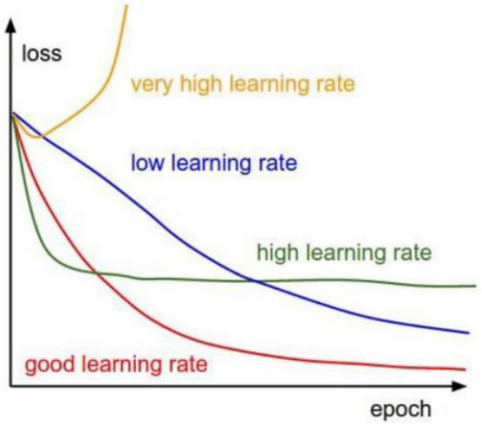
$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\theta_j} = \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)} \tag{14}$$



1.8 Learning Rate (α)

The effect of learning rate on results:

- Too high: the loss function increases (divergent)
- Big: loss function decreases, but reaches the local minimum, not global minimum
- Small: loss function decreases, but it takes a long time to reach global minimum (low speed)
- Fit: perfect



1.9 Evaluation for Model

We evaluate a model with R-squared (R^2), which is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. We think the model is better if R^2 is close to 1.

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$R^2\text{: } 1 - \frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

1.10 Pros and Cons

Pros:

- modeling is fast, does not require very complicated calculations
- each variable can be understood and explained according to the coefficient

Cons:

Nonlinear data cannot be fitted well. Thus we need to determine whether the variables are linear or not.

2 Notes

2.1 Feature Scaling

Feature Scaling is used to reduce the range to [-1, 1] by dividing by s_i=max.

$$x_i = \frac{x_i}{s_i}$$

Example: [-2, 2, 4, 8] → divide by max=8 → [-0.25, 0.25, 0.5, 1]

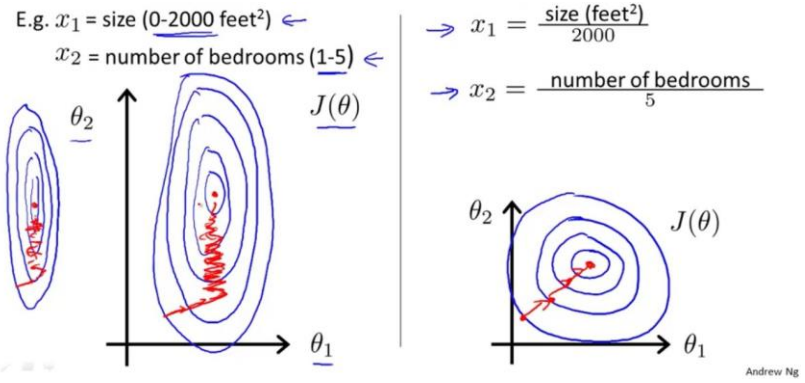
2.2 Mean Normalization

$$x_i = \frac{x_i - \mu_i}{s_i}$$

$\mu_i = mean\ of\ feature$
$s_i = max - min$

2.3 Pros for Feature Scaling

Divergent rapidly.



3 Review

Question 1:
Given how well a student did in her first year. Specifically, let x be equal to the number of "A" grades (including A-. A and A+ grades) that a student receives in freshmen year. We would like to predict the value of y, which we define as the number of "A" grades they get in sophomore year. It would be referred by the following questions.

Question 2
Consider the following training set of m=4 training examples:
x y
1 0.5
2 1
4 2
0 0
Consider the linear regression model $h_\theta(x)=\theta_0+\theta_1 x$. What are the values of θ_0 and θ_1 that you would expect to obtain upon running gradient descent on this model? (Linear regression will be able to fit this data perfectly.)
A. $\theta_0=0$, $\theta_1=0.5$
B. $\theta_0=1$, $\theta_1=1$
C. $\theta_0=0.5$, $\theta_1=0.5$
D. $\theta_0=1$, $\theta_1=0.5$
E. $\theta_0=0.5$, $\theta_1=0$
F. $\theta_0=0$, $\theta_1=0.5$
Answer: A

Question 3
Suppose we set $\theta_0=-1$, $\theta_1=2$ in the linear regression hypothesis from Q1. What is $h_\theta(6)$?
Answer: -1 + 2*6 = 11

Question 4

Let f be some function so that $f(\theta_0,\theta_1)$ outputs a number. For this problem, f is some arbitrary/unknown smooth function. Suppose we use gradient descent to try to minimize $f(\theta_0, \theta_1)$ as a function of θ_0 and θ_1. Which of the following statements are true? (Check all that apply.)

A. No matter how θ_0 and θ_1 are initialized, so long as α is sufficiently small, we can safely expect gradient descent to converge to the same solution.
B. If the first few iterations of gradient descent cause $f(\theta_0, \theta_1)$ to increase rather than decrease, then the most likely cause is that we have set the learning rate α to too large a value.
C. If θ_0 and θ_1 are initialized at the global minimum, then one iteration will not change their values.
D. Setting the learning rate α to be very small is not harmful, and it can only speed up the convergence of gradient descent.

Answer: B、C

5. Question 5

Suppose that for some linear regression problem (say, predicting housing prices as in the lecture), we have some training set, and for our training set we managed to find some θ_0, θ_1 such that $J(\theta_0, \theta_1)=0$. Which of the statements below must then be true? (Check all that apply.)

A. For this to be true, we must have $\theta_0=0$ and $\theta_1=0$ so that $h_\theta(x)=0$
B. We can perfectly predict the value of y even for new examples that we have not yet seen. (e.g., we can perfectly predict prices of even new houses that we have not yet seen.)
C. For these values of θ_0 and θ_1 that satisfy $J(\theta_0, \theta_1)=0$, we have that $h_\theta(x(i))=y(i)$ for every training example (x(i), y(i))
D. This is not possible: By the definition of $J(\theta_0, \theta_1)$, it is not possible for there to exist θ_0 and θ_1 so that $J(\theta_0, \theta_1)=0$

Answer: C