# (선형) 회귀 분석

# (Linear Regression Analysis)

# 선형 회귀 분석(Linear Regression Analysis)

- 선형성이라는 기본 가정이 충족된 상태에서 독립변수와 종속변수의 관계를 설명하거나 예측하는 통계방법

- 회귀분석에서 독립변수에 따라 종속변수의 값이 일정한 패턴으로 변해 가는데, 이러한 변수간의 관계를 나타내는 회귀선이 직선에 가깝게 나타나는 경우

- 종 류
  - 단순회귀분석 - 독립변수가 하나인 경우
  - 다중회귀분석 - 독립변수가 여러 개인 경우

# 선형 회귀 분석(Linear Regression Analysis)

$$y = Wx + b$$

- x : 독립변수
- y : 종속 변수
- W : 직선의 기울기(가중치 : weight)
- b : y 절편(bias)

# 선형 회귀 분석(Linear Regression Analysis)

## - 편 차(Deviation)

- 수학 및 통계학에서 편차는 자료값 또는 변량과 평균의 차이를 나타내는 수치
- 편차를 살펴보면 자료들이 평균을 중심으로 얼마나 퍼져 있는지를 알 수 있다.
- 자료값이 평균보다 크면 편차는 양의 값을, 평균보다 작으면 음의 값을 갖는다.
- 편차의 크기는 차이의 크기를 나타낸다.
- 편차의 절댓값은 절대편차, 편차의 제곱은 제곱편차라고 한다.

# 선형 회귀 분석(Linear Regression Analysis)

## - 용어 정의

- 잔차(Residual)
  - 회귀분석에서 종속변수와 적합값(예상값)의 차이.
  - 잔차는 (종속변수 - 적합값)으로 정의.
- 분산(Variance)
  - 편차의 제곱
- 표준 편차(Standard Deviation)
  - 분산의 제곱근

# 선형 회귀 분석(Linear Regression Analysis)

## 예상 시험 점수 : regression

| x(hours) | y(score) |
|---|---|
| 10 | 90 |
| 9 | 80 |
| 3 | 50 |
| 2 | 30 |

# 선형 회귀 분석(Linear Regression Analysis)

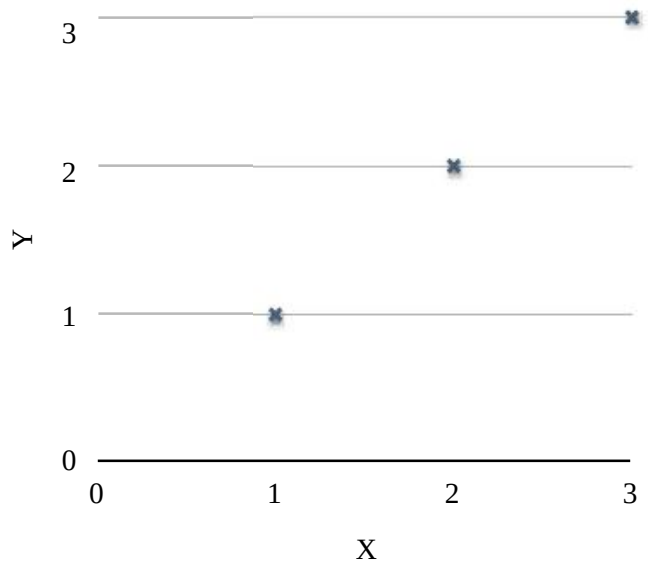
## Regression (data)

| x | y |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |

# 선형 회귀 분석(Linear Regression Analysis)

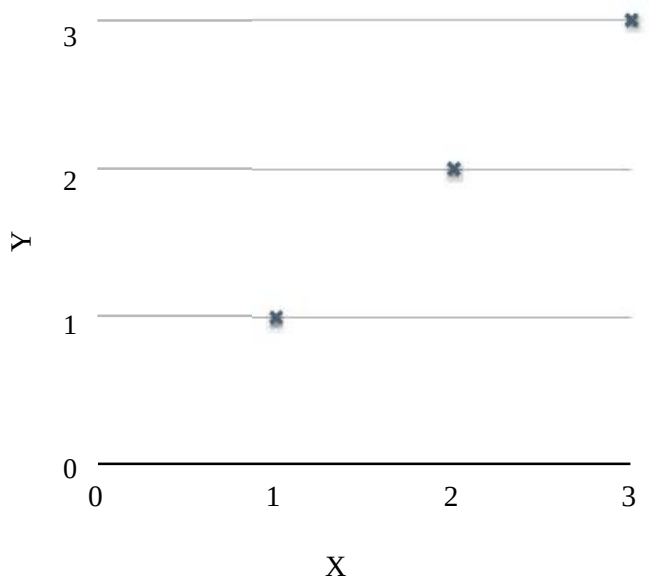
## Regression (presentation)

| x | Y |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |



# 선형 회귀 분석(Linear Regression Analysis)

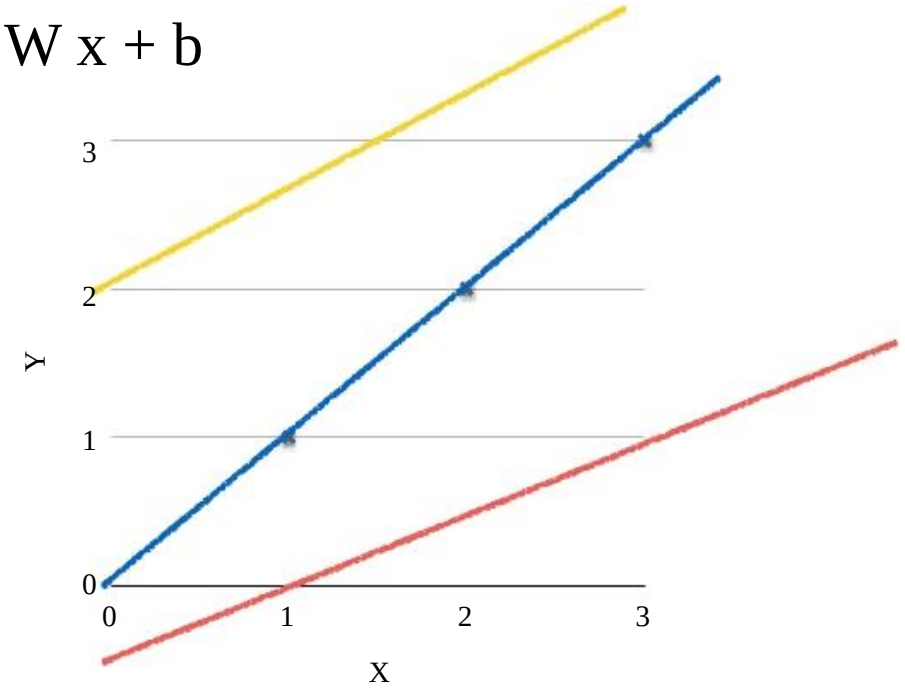
## (Linear) Hypothesis



# 선형 회귀 분석(Linear Regression Analysis)

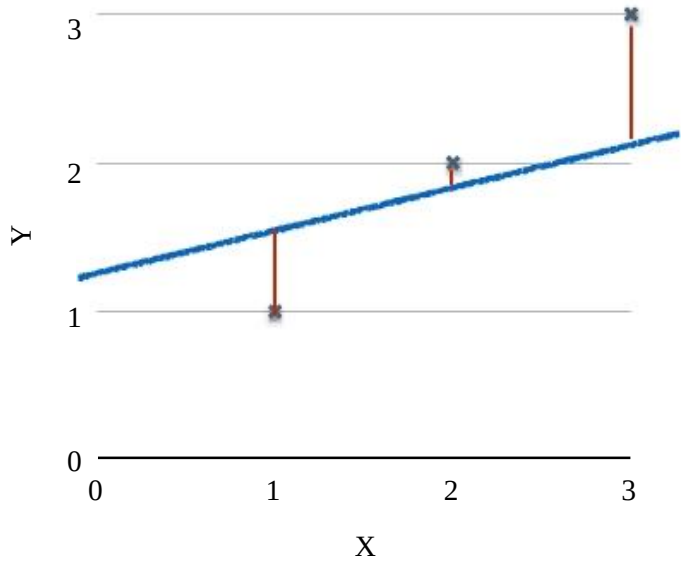
## (Linear) Hypothesis

$$H(x) = Wx + b$$



# 선형 회귀 분석(Linear Regression Analysis)

## Which hypothesis is better?

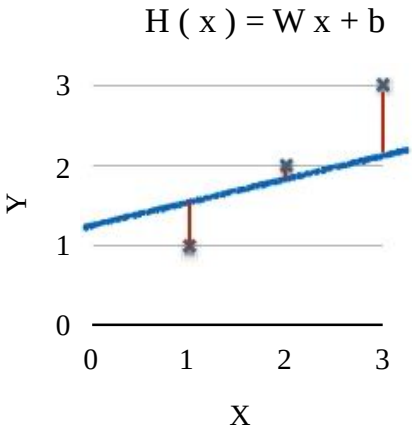


# Cost function

- How fit the line to our (training) data

$$H(x) - y$$

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$




$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

# Cost function

$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

$$H(x) = Wx + b$$

$$cost(W, b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

# 선형 회귀 분석(Linear Regression Analysis)

Goal: Minimize cost

$$\underset{W,b}{\text{minimize}} \ \text{cost}(W, b)$$

# Multivariable Linear Regression

# 다중 선형 회귀 분석

- Hypothesis

$$H(x) = Wx + b$$

- Cost Function

$$cost(W, b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

- Gradient descent algorithm

$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

# 다중 선형 회귀 분석

## 예상 시험 점수 :
regression using one input (x)

one-variable
one-feature

| x (hours) | y (score) |
|---|---|
| 10 | 90 |
| 9 | 80 |
| 3 | 50 |
| 2 | 60 |
| 11 | 40 |

# 다중 선형 회귀 분석

## 시험 성적 예측:
one input (x1, x2, x3)을 사용한 회귀 분석

multi-variable/feature

| $x_1$ (quiz 1) | $x_2$ (quiz 2) | $x_3$ (midterm 1) | Y (final) |
|---|---|---|---|
| 73 | 80 | 75 | 152 |
| 93 | 88 | 93 | 185 |
| 89 | 91 | 90 | 180 |
| 96 | 98 | 100 | 196 |
| 73 | 66 | 70 | 142 |

# 다중 선형 회귀 분석

## 가설(Hypothesis)

$$H(x) = Wx + b$$

$$H(x_1, x_2, x_3) = w_1x_1 + w_2x_2 + w_3x_3 + b$$

# Cost function

$$H(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

$$cost(W, b) = \frac{1}{m} \sum_{I=1}^{m} (H(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}) - y^{(i)})^2$$

# 다중 선형 회귀 분석

## Multi-variable

$$H(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

$$H(x_1, x_2, x_3, ..., x_n) = w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_n x_n + b$$

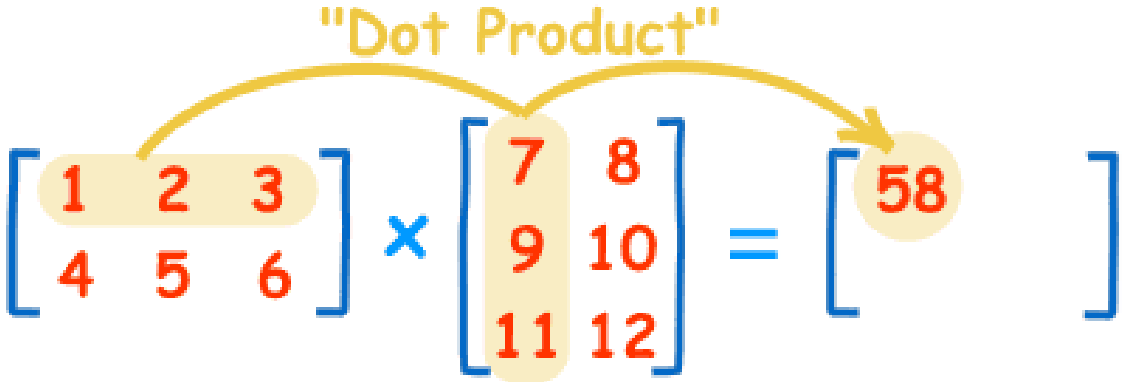
# 다중 선형 회귀 분석

Matrix

$$w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

# 다중 선형 회귀 분석

## Matrix multiplication



# 다중 선형 회귀 분석

matrix를 사용한 Hypothesis

$$w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_1w_1 + x_2w_2 + x_3w_3 \end{pmatrix}$$

$$H(X) = XW$$

# 다중 선형 회귀 분석

## matrix를 사용한 Hypothesis

$$H(x_1, x_2, x_3) = x_1w_1 + x_2w_2 + x_3w_3$$

| $x_1$ | $x_2$ | $x_3$ | Y |
|---|---|---|---|
| 73 | 80 | 75 | 152 |
| 93 | 88 | 93 | 185 |
| 89 | 91 | 90 | 180 |
| 96 | 98 | 100 | 196 |
| 73 | 66 | 70 | 142 |

Test Scores for General Psychology

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_1w_1 + x_2w_2 + x_3w_3 \end{pmatrix}$$

$$H(X) = XW$$

# 다중 선형 회귀 분석

## matrix를 사용한 Hypothesis

| $x_1$ | $x_2$ | $x_3$ | Y |
|---|---|---|---|
| 73 | 80 | 75 | 152 |
| 93 | 88 | 93 | 185 |
| 89 | 91 | 90 | 180 |
| 96 | 98 | 100 | 196 |
| 73 | 66 | 70 | 142 |

$$w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_{11}w_1 + x_{12}w_2 + x_{13}w_3 \\ x_{21}w_1 + x_{22}w_2 + x_{23}w_3 \\ x_{31}w_1 + x_{32}w_2 + x_{33}w_3 \\ x_{41}w_1 + x_{42}w_2 + x_{43}w_3 \\ x_{51}w_1 + x_{52}w_2 + x_{53}w_3 \end{pmatrix}$$

$$H(X) = XW$$

# 다중 선형 회귀 분석

## matrix를 사용한 Hypothesis

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_{11}w_1 + x_{12}w_2 + x_{13}w_3 \\ x_{21}w_1 + x_{22}w_2 + x_{23}w_3 \\ x_{31}w_1 + x_{32}w_2 + x_{33}w_3 \\ x_{41}w_1 + x_{42}w_2 + x_{43}w_3 \\ x_{51}w_1 + x_{52}w_2 + x_{53}w_3 \end{pmatrix}$$

[5, 3] [3, 1] [5, 1]

$$H(X) = XW$$

# 다중 선형 회귀 분석

## matrix를 사용한 Hypothesis

$$H(x_1, x_2, x_3) = x_1w_1 + x_2w_2 + x_3w_3$$

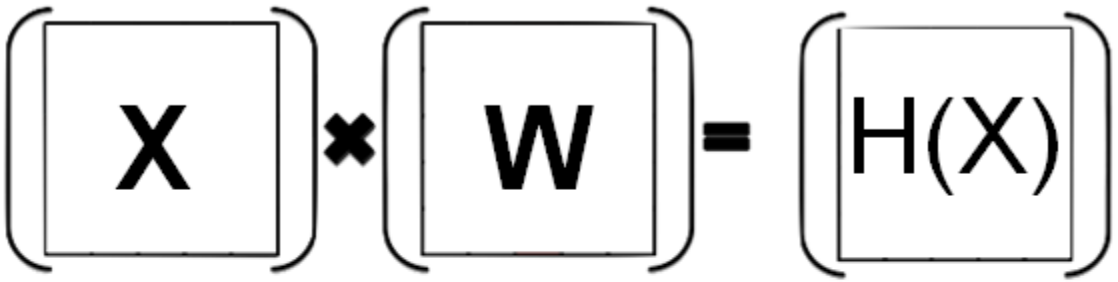
| $x_1$ | $x_2$ | $x_3$ | Y |
|---|---|---|---|
| 73 | 80 | 75 | 152 |
| 93 | 88 | 93 | 185 |
| 89 | 91 | 90 | 180 |
| 96 | 98 | 100 | 196 |
| 73 | 66 | 70 | 142 |

Test Scores for General Psychology

```
x1_data = [73., 93., 89., 96., 73.]
x2_data = [80., 88., 91., 98., 66.]
x3_data = [75., 93., 90., 100., 70.]
y_data = [152., 185., 180., 196., 142.]

# placeholders for a tensor that will be always fed.
x1 = tf.placeholder(tf.float32)
x2 = tf.placeholder(tf.float32)
x3 = tf.placeholder(tf.float32)

Y = tf.placeholder(tf.float32)

w1 = tf.Variable(tf.random_normal([1]), name='weight1')
w2 = tf.Variable(tf.random_normal([1]), name='weight2')
w3 = tf.Variable(tf.random_normal([1]), name='weight3')
b = tf.Variable(tf.random_normal([1]), name='bias')

hypothesis = x1 * w1 + x2 * w2 + x3 * w3 + b
```

# 다중 선형 회귀 분석

matrix를 사용한 Hypothesis




$$H(X) = XW$$

# 다중 선형 회귀 분석

## matrix를 사용한 Hypothesis

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_{11}w_1 + x_{12}w_2 + x_{13}w_3 \\ x_{21}w_1 + x_{22}w_2 + x_{23}w_3 \\ x_{31}w_1 + x_{32}w_2 + x_{33}w_3 \\ x_{41}w_1 + x_{42}w_2 + x_{43}w_3 \\ x_{51}w_1 + x_{52}w_2 + x_{53}w_3 \end{pmatrix}$$

`[n, 3]` `[3, 1]` `[n, 1]`

$$H(X) = XW$$

# 다중 선형 회귀 분석

matrix를 사용한 Hypothesis(n output)

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \; ? \; = \begin{pmatrix} x_{11}w_{11} + x_{12}w_{21} + x_{13}w_{31} & x_{11}w_{12} + x_{12}w_{22} + x_{13}w_{32} \\ x_{21}w_{11} + x_{22}w_{21} + x_{23}w_{31} & x_{21}w_{12} + x_{22}w_{22} + x_{23}w_{32} \\ x_{31}w_{11} + x_{32}w_{21} + x_{33}w_{31} & x_{31}w_{12} + x_{32}w_{22} + x_{33}w_{32} \\ x_{41}w_{11} + x_{42}w_{21} + x_{43}w_{31} & x_{41}w_{12} + x_{42}w_{22} + x_{43}w_{32} \\ x_{51}w_{11} + x_{52}w_{21} + x_{53}w_{31} & x_{51}w_{12} + x_{52}w_{22} + x_{53}w_{32} \end{pmatrix}$$

[n, 3] [?, ?] [n, 2]

$$H(X) = XW$$

# 다중 선형 회귀 분석

matrix를 사용한 Hypothesis(n output)

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix} = \begin{pmatrix} x_{11}w_{11} + x_{12}w_{21} + x_{13}w_{31} & x_{11}w_{12} + x_{12}w_{22} + x_{13}w_{32} \\ x_{21}w_{11} + x_{22}w_{21} + x_{23}w_{31} & x_{21}w_{12} + x_{22}w_{22} + x_{23}w_{32} \\ x_{31}w_{11} + x_{32}w_{21} + x_{33}w_{31} & x_{31}w_{12} + x_{32}w_{22} + x_{33}w_{32} \\ x_{41}w_{11} + x_{42}w_{21} + x_{43}w_{31} & x_{41}w_{12} + x_{42}w_{22} + x_{43}w_{32} \\ x_{51}w_{11} + x_{52}w_{21} + x_{53}w_{31} & x_{51}w_{12} + x_{52}w_{22} + x_{53}w_{32} \end{pmatrix}$$

[n, 3] **[3, 2]** [n, 2]

$$H(X) = XW$$

# 다중 선형 회귀 분석

## WX vs XW

- 이론

$$H(x) = Wx + b$$

- 구현(TensorFlow)

$$H(X) = XW$$

# 다중 선형 회귀 분석

## Matrix

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_1 w_1 + x_2 w_2 + x_3 w_3 \end{pmatrix} \qquad H(X) = XW$$

```
x_data = [[73., 80., 75.], [93., 88., 93.],
          [89., 91., 90.], [96., 98., 100.], [73., 66., 70.]]
y_data = [[152.], [185.], [180.], [196.], [142.]]


# placeholders for a tensor that will be always fed.
X = tf.placeholder(tf.float32, shape=[None, 3])
Y = tf.placeholder(tf.float32, shape=[None, 1])

W = tf.Variable(tf.random_normal([3, 1]), name='weight')
b = tf.Variable(tf.random_normal([1]), name='bias')

# Hypothesis
hypothesis = tf.matmul(X, W) + b
```

# 다중 선형 회귀 분석

## Loading data from file

**data-01-test-score.csv**

```
# EXAM1,EXAM2,EXAM3,FINAL
73,80,75,152
93,88,93,185
89,91,90,180
96,98,100,196
73,66,70,142
53,46,55,101
```

```python
import numpy as np

xy = np.loadtxt('data-01-test-score.csv', delimiter=',', dtype=np.float32)
x_data = xy[:, 0:-1]
y_data = xy[:, [-1]]

# Make sure the shape and data are OK
print(x_data.shape, x_data, len(x_data))
print(y_data.shape, y_data)
```