K6312 Information Mining & Analysis

Chen Zhenghua & Zhao Rui

Linear Regression

Supervised Learning

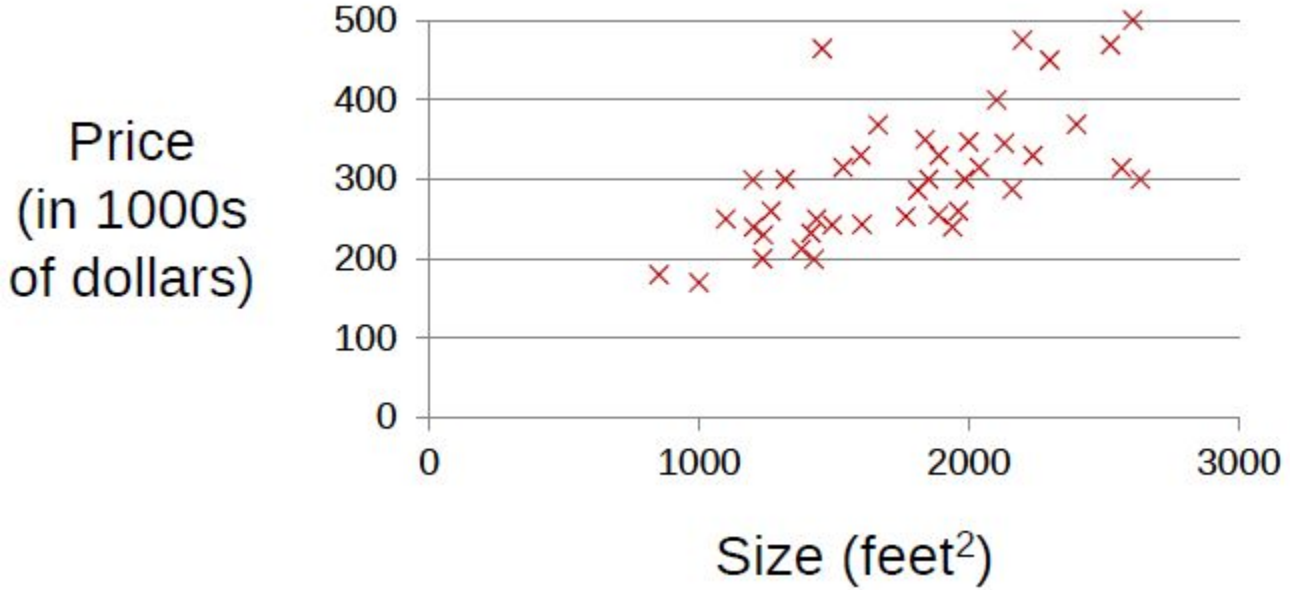
- Formalization
  - Input: $\mathbf{x} \in \mathcal{X} \quad \mathbb{R}^n$
  - Output: $y \in \mathcal{Y} \begin{cases} \mathbb{R} & \text{regression} \\ \{+1, -1\} & \text{binary classification} \\ \{1, 2, \ldots, K\} & \text{multi-class classification} \end{cases}$
  - Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ (unknown)
  - Training Data: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$
  - Hypothesis: $h : \mathcal{X} \rightarrow \mathcal{Y} \quad h \approx f$
  - Hypothesis space: $h \in \mathcal{H}$

Explanatory and Target Variables

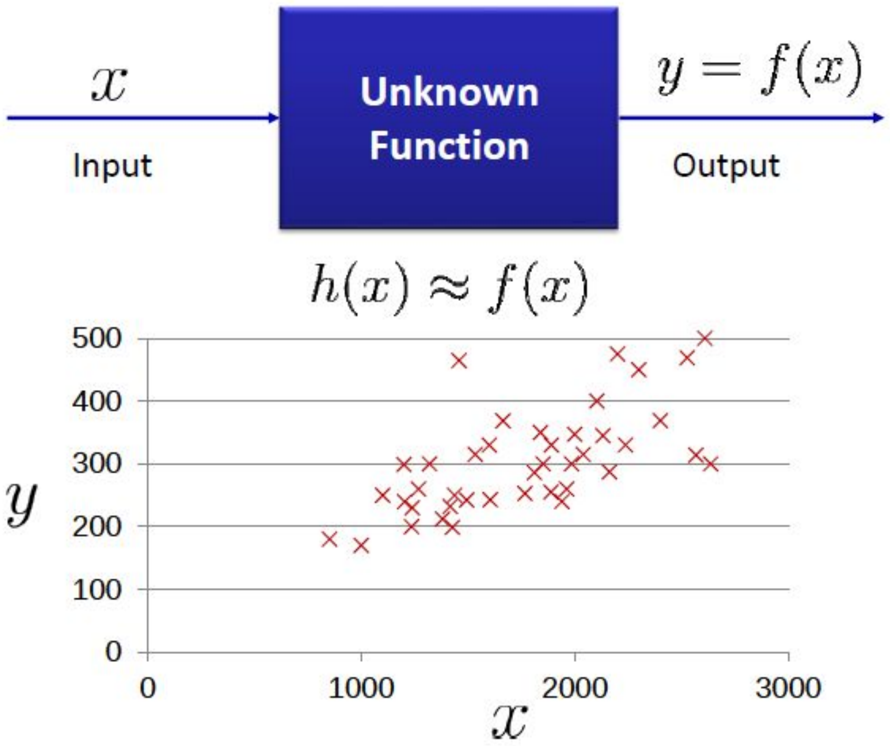
Size in $feet^2$ (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

- ***x*** = input variable / explanatory variable
- ***y*** = output variable / target variable

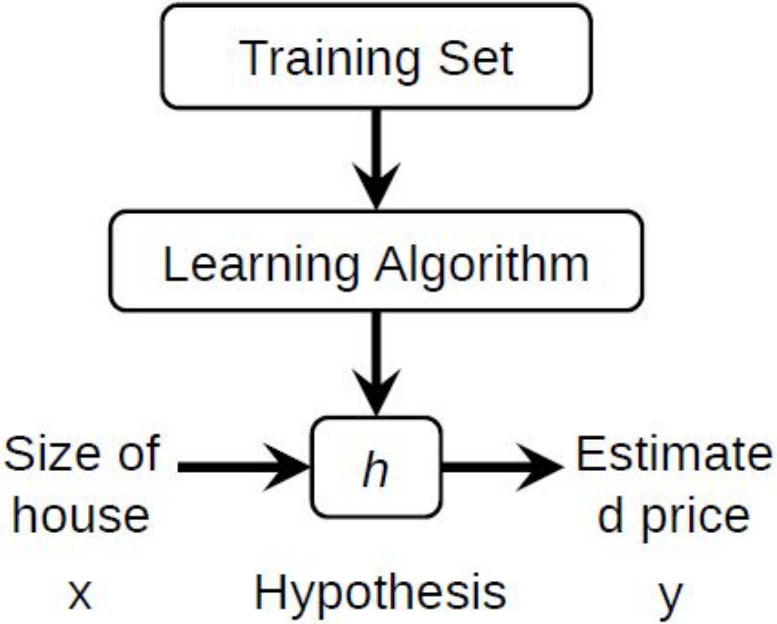
Target Functions



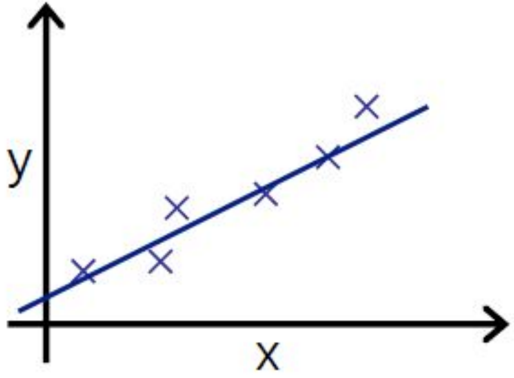
A Learning Problem



Model Representation



How do we represent *h* ?



$$h(x) = w_0 + w_1 x$$

Linear regression with one variable.
"**Univariate Linear Regression**"

Formulation: Cost Function

Hypothesis:

$$h(x) = w_0 + w_1 x$$

Parameters:

$$w_0, w_1$$

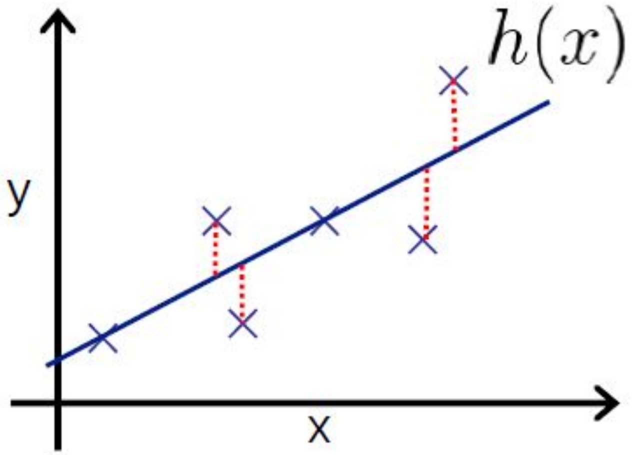
Cost Function:

Mean Squared Error (MSE)

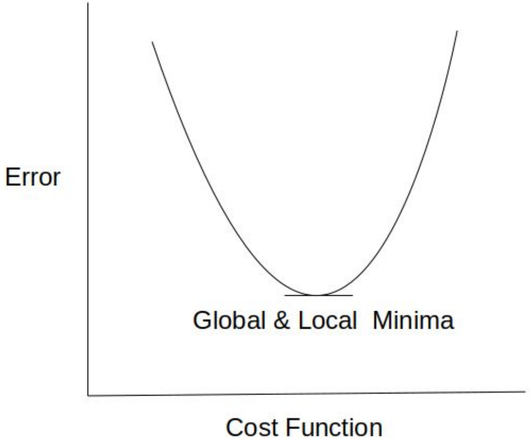
$$J(w_0, w_1) = \frac{1}{m} \sum_{i=1}^{m} [h(x_i) - y_i]^2$$

Goal:

$$\min_{w_0, w_1} J(w_0, w_1)$$



Normal Equation: Least Square Fit



$$\frac{\partial J(w_0, w_1)}{\partial w_0} = \frac{2}{m}\sum_{i=1}^{m}(w_0 + w_1 x_i - y_i) = 0$$

$$\frac{\partial J(w_0, w_1)}{\partial w_1} = \frac{2}{m}\sum_{i=1}^{m}x_i(w_0 + w_1 x_i - y_i) = 0$$

$$w_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1\bar{x}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{m} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i$ are the samples means

Assessing the Overall Accuracy of the Model

- Mean Square Error

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

- Mean absolute Error

$$\text{MAE} = \frac{1}{m}|y_i - \hat{y}_i|$$

$\bar{y}$: mean of y_i
$\hat{y}_i$: prediction of y_i

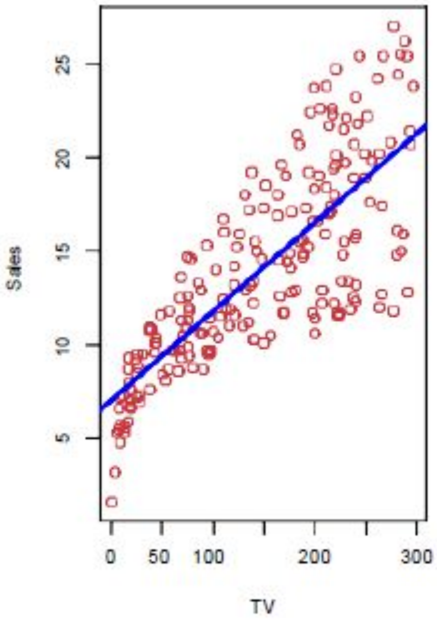
Linear regression for the advertising data

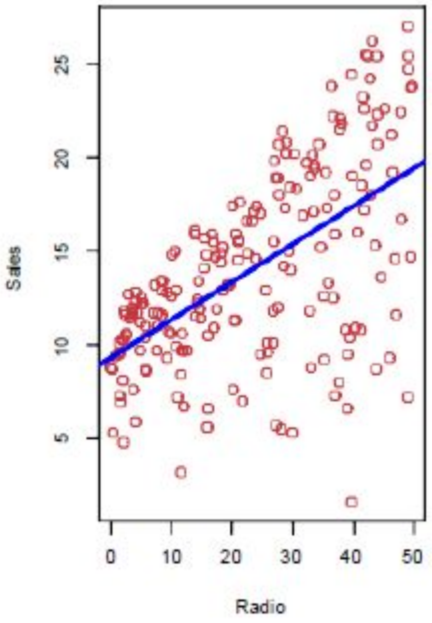
Consider the advertising data shown on the next slide.

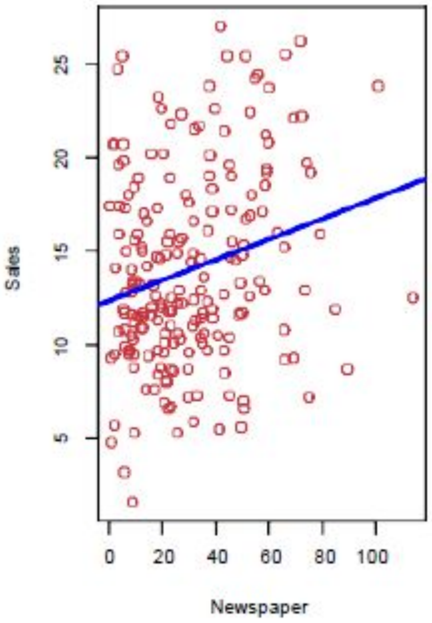
Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?

Advertising data



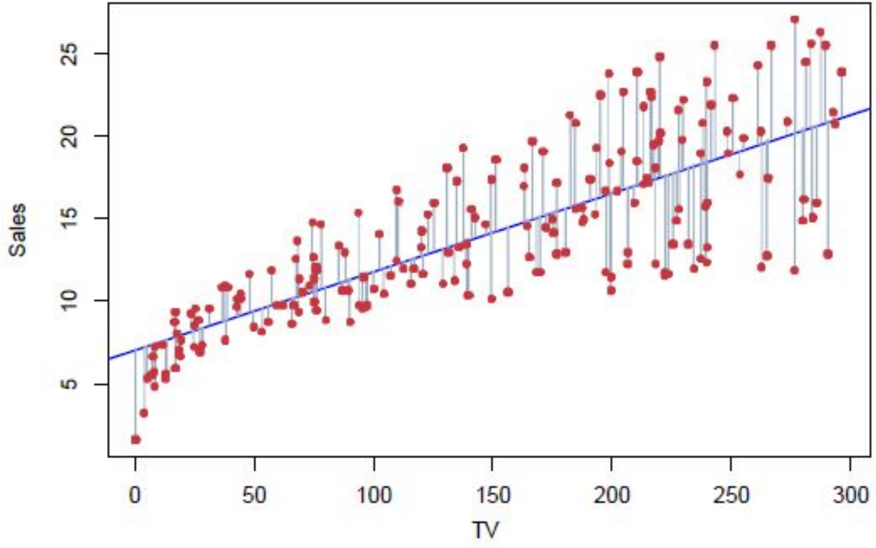




Example: advertising data

- The least squares fit for the regression of *sales* onto *TV*.
- In this case, a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

$$w_0 = 7.03 \text{ and } w_1 = 0.0475$$



Multivariate Linear Regression

Multiple features (variables).

Size ($feet^2$)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

Notation:

n = number of features

$\mathbf{x}_i$ = input (features) of i^{th} training example.

x_{ij} = value of feature j in i^{th} training example.

Multivariate Linear Regression

Hypothesis:

Previously: $h(x) = w_0 + w_1 x$

$\vec{x} \in \mathbb{R}^n \quad h(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$

For convenience of notation, define $x_0 = 1$.

$$h(\vec{x}) = \sum_{j=0}^{n} w_j x_j = \vec{w}^\intercal \vec{x} = \langle \vec{w}, \vec{x} \rangle$$

$$\vec{x} \in \mathbb{R}^{n+1} \quad \vec{w} \in \mathbb{R}^{n+1}$$

Normal Equation

$$J(\vec{w}) = \frac{1}{m}\sum_{i=1}^{m}(\vec{w}^{\intercal}\vec{x}_i - y_i)^2 = \frac{1}{m}(\mathbf{X}\vec{w} - \vec{y})^{\intercal}(\mathbf{X}\vec{w} - \vec{y})$$

$$(\vec{w}^{\intercal}\vec{x}_1 - y_1)$$

$$\mathbf{X}\vec{w} - \vec{y} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1n} \\ x_{20} & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m0} & x_{m1} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

$$m \times (n+1) \qquad (n+1) \times 1 \qquad m \times 1$$

Normal Equation

- Matrix-vector formulation

$$J(\vec{w}) = \frac{1}{m}(\mathbf{X}\vec{w} - \vec{y})^\intercal(\mathbf{X}\vec{w} - \vec{y})$$

$$\begin{aligned} \nabla J(\vec{w}) &= \nabla_w \frac{1}{m}(\mathbf{X}\vec{w} - \vec{y})^\intercal(\mathbf{X}\vec{w} - \vec{y}) \\ &= \mathbf{X}^\intercal\mathbf{X}\vec{w} - \mathbf{X}^\intercal\vec{y} \end{aligned}$$

$$\mathbf{X}^\intercal\mathbf{X}\vec{w} = \mathbf{X}^\intercal\vec{y}$$

- Analytical solution:

$$\vec{w} = ((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\vec{y} = \mathbf{X}^\dagger\vec{y}$$

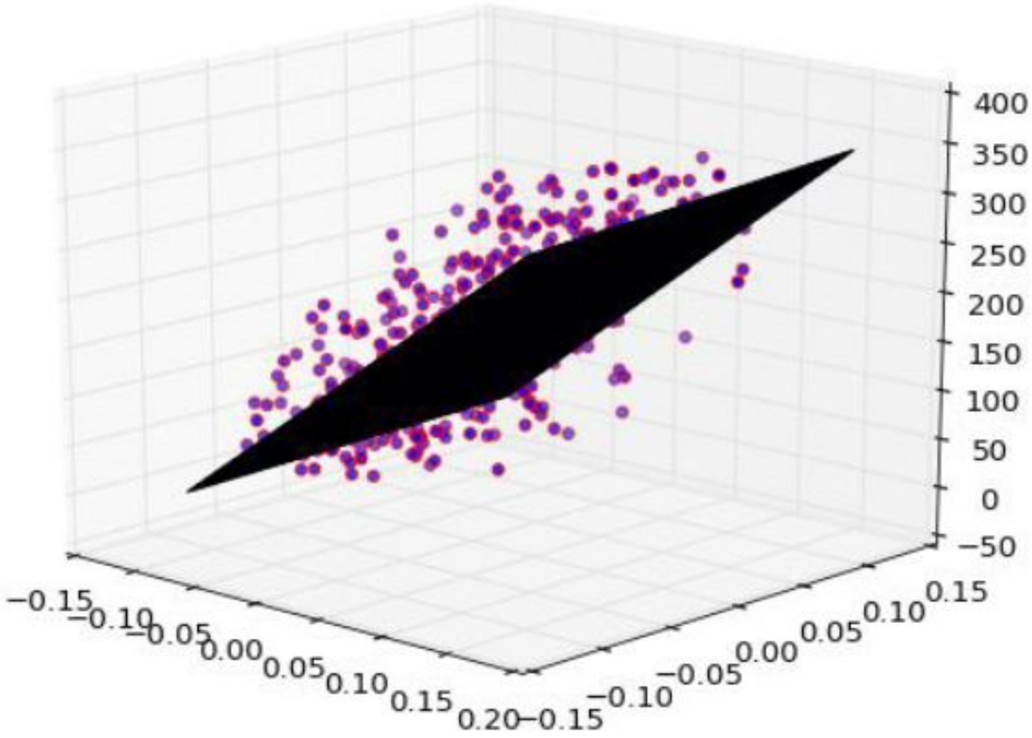
where $\mathbf{X}^\dagger = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$

Advertising Example

In the advertising example, the model becomes

$$\text{sales} = w_0 + w_1 \times \text{TV} + w_2 \times \text{radio} + w_3 \times \text{newspaper}$$

Linear Regression Visualization



Twenty Minutes Break