1.4 Regression

1.4.1 Linear Regression

Structure of input data

Size ($feet^2$) x_1	Number of bedrooms x_2	Number of floors x_3	age of home (years) x_4	Price($1000) $h_\theta(x) = y$
2104	5	1	45	460
1416	3	2	40	232
1534	2	2	30	315
...	...	...	...	...

A data

Features (x)

Label ($h_\theta(x) = y$)

Structure of linear regression model

$$h_\theta(x) = \theta_0 + 2104\theta_1 + 5\theta_2 + \theta_3 + 45\theta_4 = 460$$
$$h_\theta(x) = \theta_0 + 1416\theta_1 + 3\theta_2 + 2\theta_3 + 40\theta_4 = 232$$
$$h_\theta(x) = \theta_0 + 1534\theta_1 + 2\theta_2 + 2\theta_3 + 30\theta_4 = 315$$

Definition: In linear regression model, a variable, called dependent variable, is assumed to be normally distributed around linear combination of other variables, called independent variables.

$$p(y|x_1, x_2, \ldots) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\|y - \boldsymbol{x}\theta\|^2}{2\sigma^2}}$$

Here y is the dependent variable and $\boldsymbol{x} = x_1, x_2, \ldots$ are independent variables. *(Multiple Linear Regression)*

We need to find θ such that this probability is maximized.

This is equivalent to minimizing $\|y - \boldsymbol{x}\,\theta\|^2$.

For multiple data points, the quantity to be minimized is $\|\boldsymbol{y} - X\,\theta\|^2$

Taking derivative with respect to θ and equating to 0 gives the solution, $\theta = (X^T X)^{-1} X^T \boldsymbol{y}$

Example 1

X: [1,0], [0,-1], [1,-2], [2,0], [0,-2]

y: 0, -1, -4, -1, -3

Example 2 (Homework)

X: 1, 0, 1, 2, 0

Y: 1.0, 2.0, 0.9, 0.0, 2.2

1.4.2 Logistic Regression

[As discussed on board.]

1.4.3 Non Linear Regression

Use linear regression after adding additional attributes derived by applying non-linear functions on original attributes. For example, x^2 and x^3 can be incorporated to use linear regression to fit cubic equation.