Type of Regression problems

Andersen Ang

First created: 2014. Last update : 2017-Feb

1 Linear Regression

Given a training set $T = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with an unknown function relating the $\mathbf{x}$ and y : $y = f(\mathbf{x})$, the goal is to find an estimator $\hat{y}$ to approximate the true unknown $f(x)$.

In linear regression model, it is assumed the model is in the form

$$\hat{y}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{p} w_i x_i + \varepsilon$$

$\mathbf{x}$ is the datapoint vector with p dimension.
$\mathbf{w}$ is the unknown weighting coefficient to be found by using the data in T
ε is the residual error between the linear regression prediction and the true response. It is commonly assumed ε obey Gaussian distribution $\varepsilon \sim \mathcal{N}\left(\mu, \sigma^2\right)$

2 Linear regression as a parametric model of conditional probability maximization

Linear regresion is parametric model. The parametric model is an estimator that pick the value of y in a way that maximize the following conditional probability

$$\mathbb{P}\left(y \middle| \mathbf{x}, \mathbf{w}\right)$$

In Gaussian assumption, the probability can be rewritten as

$$\mathbb{P}\left(y \middle| \mathbf{x}, \mathbf{w}\right) = \mathcal{N}\left(y \middle| \mu(\mathbf{x}), \sigma^2(\mathbf{x})\right)$$

In linear regression, the $\mu(\mathbf{x})$ is a linear function $\mu(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x}$ and $\sigma^2(\mathbf{x}) = \sigma^2$ is a constant.

$$\mathbb{P}\left(y \middle| \mathbf{x}, \mathbf{w}\right) = \mathcal{N}\left(y \middle| \mathbf{w}^T\mathbf{x}, \sigma^2\right)$$

3 Basis function Expansion in linear regression

Linear regression can be used to model non-linear problem with the use of kernel $\phi(\mathbf{x})$

$$\mathbb{P}\left(y\middle|\mathbf{x},\mathbf{w}\right) = \mathcal{N}\left(y\middle|\mathbf{w}^T\phi(\mathbf{x}),\sigma^2\right)$$

4 Logistic Regression

Linear regression is used for different value of y. When there is only two class (binary classification), logistic regression is more suitable. Logistic function applies the sigmoid function

$$sigmod(t) \hat{=} \frac{1}{1+\exp(-t)}$$

Therefore, the function $\mu(\mathbf{x})$ is

$$\mu(\mathbf{x}) = sigmoid(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-\mathbf{W}^T\mathbf{X}}}$$

In terms of probability density, since it is binary class, so the density is Bernoulli distribution

$$\mathbb{P}\left(y\middle|\mathbf{x},\mathbf{w}\right) = Ber\left(y\middle|sigmod(\mathbf{w}^T\mathbf{x})\right)$$

5 The regression problem

Consider the model (in vector notation)

$$y = x^T\beta + \varepsilon$$

y (a scalar) is the response / output / measurement of the system, which is known

x ($p \times 1$ vector) is the excitation / input of the system, which is known or unknown

β ($p \times 1$ vecotr) is the coefficients / parameters of the system, which is unknown

ε (a scalar) is the noise / error of the system, Gaussian noise is assumed : $\varepsilon \sim \mathcal{N}\left(0,\sigma^2\right)$

p is the dimension of the system

Consider we have n observation

$$\begin{cases} y_1 = x_1^T\beta + \varepsilon_1 \\ y_2 = x_2^T\beta + \varepsilon_2 \\ \vdots \\ y_n = x_n^T\beta + \varepsilon_n \end{cases} \iff y_i = x_i^T\beta + \varepsilon_i \; i = 1, ..., n$$

The goal is to find β from x and y

6 The regression problem in matrix-vector form

More condensed notation can be make using vector and matrix

$$y = X\beta + \varepsilon \iff \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \longleftarrow x_1^T \longrightarrow \\ \longleftarrow x_2^T \longrightarrow \\ \vdots \\ \longleftarrow x_n^T \longrightarrow \end{bmatrix} \begin{bmatrix} \uparrow \\ \beta \\ \downarrow \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

y ($n \times 1$ vector) is the response / output / measurement of the system
X ($n \times p$ vector) is the excitation / input of the system
β ($p \times 1$ vecotr) is the coefficients / parameters of the system
ε ($n \times 1$ vector) is the noise / error of the system, Gaussian noise is assumed : $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$
p is the dimension of the system
n is the number of observation

7 Relation to Linear Algebra

Since our goal is to find β , this regression problem now becomes a matrix inverse problem

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \longleftarrow x_1^T \longrightarrow \\ \longleftarrow x_2^T \longrightarrow \\ \vdots \\ \longleftarrow x_n^T \longrightarrow \end{bmatrix} \begin{bmatrix} \uparrow \\ \beta \\ \downarrow \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In linear algebra term, n is the number of equations and p is the number of unknowns.

When $n > p$, it is a over-determined system. It means we have more #equations (information) than #unknowns. It is possible to have no solution if some equations (information) contradict with each other.

When $n < p$, it is a under-determined system. It means we have more #unknowns than #equations (information). We have less information, then there can be many possible solution. Therefore there eare infinitely many solution.

When $n = p$, it is a normal case. It means we have same #unknowns and #equations (information). Then we can only use that amount of information to solve that amount of unknowns to find one solution. Therefore if there is solution, then the solution will be unique.