

Probabilistic View of Linear Regression

Tirtharaj Dash

Dept. of CS & IS and APPCAIR
BITS Pilani, Goa Campus

February 5, 2021

$$\underline{x} = (x_1, x_2, \dots, x_d), \quad Y$$

independent variables dependent variable

Goal: Relate Y to a linear predictor function of \underline{x}

for any given 'i' th data point:

$$\hat{y}(i) = \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d$$

Linear in parameters: $\beta = (\beta_0, \beta_1, \dots, \beta_d)$

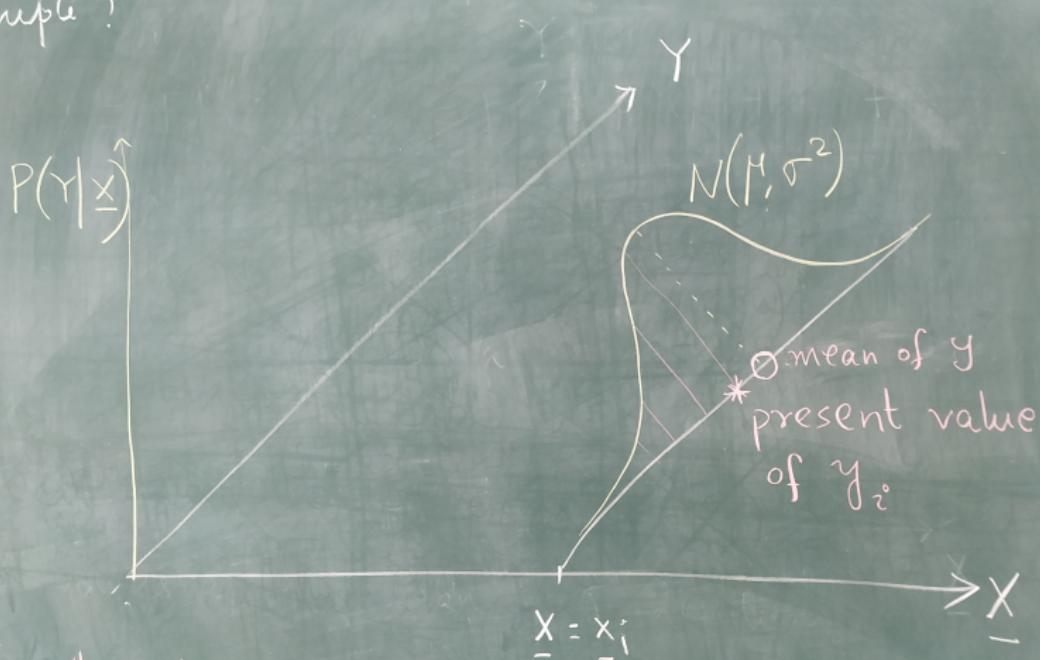
Also, possible to have linear predictor function as

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2^3 + \dots + \beta_d x_d$$

more clearly

$$\hat{y}(\underline{x}^{(i)})$$

How did we get these y 's in the present sample?



So, the dataset contains; $(x_i, \text{drawn } y_i)$ pairs



+ : Dataset 1

* : Dataset 2

any other dataset is
also a draw.

We are estimating
the means of the
Gaussians:

$P(Y|X=x_i) \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ OUR GOAL IS TO
MATCH THIS GIVEN

$P(Y|X=x_i) \sim N(\beta_0 + \beta_1 x_i^{(2)}, \sigma^2)$ THE DATASET.

How do we get these β 's?

- * Given N-data points as (x_i, y_i) pairs.

Compute the Least Square Error or MSE:

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

- * Use Partial derivatives $\frac{\partial MSE}{\partial \beta}$ and solve..

(or) Use Gradient Descent (GD).

Modelling y as a Normal distribution

A constant mean model:

$$Y \sim N(0, 1)$$

Everything is known about Y .

Nothing to estimate here !!

0 mean, μ
1 variance, σ^2

In Linear Reg:

$$y = 0 + \boxed{\epsilon}$$

Certain randomness around 0.

Let's now assume fixed $\boxed{\mu}$ and $\boxed{\sigma^2}$
Unknown

i.e. $y = \mu + \varepsilon$

and $\varepsilon \sim N(0, \sigma^2)$

or

$$Y \sim N(\mu, \sigma^2)$$

* How to get μ from data?

Sol: μ MLE of y 's in data: $\frac{1}{N} \sum_i y_i$

Maximising Likelihood: independently

N data points

identically
distributed

each is drawn in i.i.d fashion

Given μ, σ^2 , the prob. of
drawing these N points defines the
likelihood function.

$$L(\mu | y) = \prod_{i=1}^N P(y_i | \mu, \sigma^2)$$

$$\begin{aligned}
 \text{So } L(\mu | y) &= \prod_{i=1}^N P(y_i | \mu, \sigma^2) \\
 &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}}
 \end{aligned}$$

Maximise $L(\mu | y)$.

Sol: Partial derivative. w.r.t. μ and σ^2

Equate to 0, solve.

Trick: Solving is easier with Log(L).
Log likelihood

If you solve it, we get

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i$$

i.e. $E[Y] = \mu$.

Once again: Y is a random variable.

Now, we want to model the expected

value of Y as a linear function of x .

That is :

$$E\{Y|X\} = \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d.$$

Equivalently,

$$Y \sim N\left(\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d, \sigma^2\right)$$

What did we do ?

Given a data point
 \underline{x} we are trying to
estimate the mean

of the normal distro of y .

Recall again that given \underline{x} ,

y is a random variable

correspondingly $p(y)$

$$p(y) \uparrow$$



\underline{x} one point



Why we did this way?

Because, there is always noise in y :

as

$$y_i = \mu + \epsilon$$

$$= \beta_0 + \beta_1 x_1 + \dots + x_d x_d$$

Our goal: obtain μ from data

(MLE estimate)

$$L(\mu | y) = \prod_{i=1}^N P(y_i | \mu, \sigma^2)$$

$$\hat{\mu} = \arg \max_{\mu} L(\mu | y)$$

$$= \arg \max_{\mu} \log L(\mu | y)$$

$$= \arg \max_{\mu} \sum \log \left(\frac{1}{\sqrt{2\pi} \sigma^2} \right) + \left(- \frac{(y_i - \mu)^2}{2\sigma^2} \right)$$

$$= \arg \min_{\mu} \sum_i (y_i - \mu)^2$$

For our β case:

$$\underline{\beta} = \arg \min_{\beta} \sum (y_i - (\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d))^2$$

$$= \arg \min_{\beta} \sum (y_i - \hat{y}_i)^2$$

We use the notation $\hat{y}_i = E[Y | \underline{x}, \underline{\beta}]$

to denote the predicted value (expected value) of y of our model.

When predicting :-

$$\hat{y}_i = E[y_i | \underline{x}_i] = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_d x_d^{(i)}$$

↑
But, this is a point estimate

actually, it is a most likely value of y .

In reality, our model predicts a band of values of y . (mean \pm 5 deviations).