Probabilistic View of Linear Regression

Tirtharaj Dash

$\underline{X} = (x_1, x_2, \ldots x_d)$, Y

independent variables ; dependent variable

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

for any given 'i'th data point:

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

Linear in parameters : $\underline{\beta} = (\beta_0, \beta_1, \ldots \beta_d)$

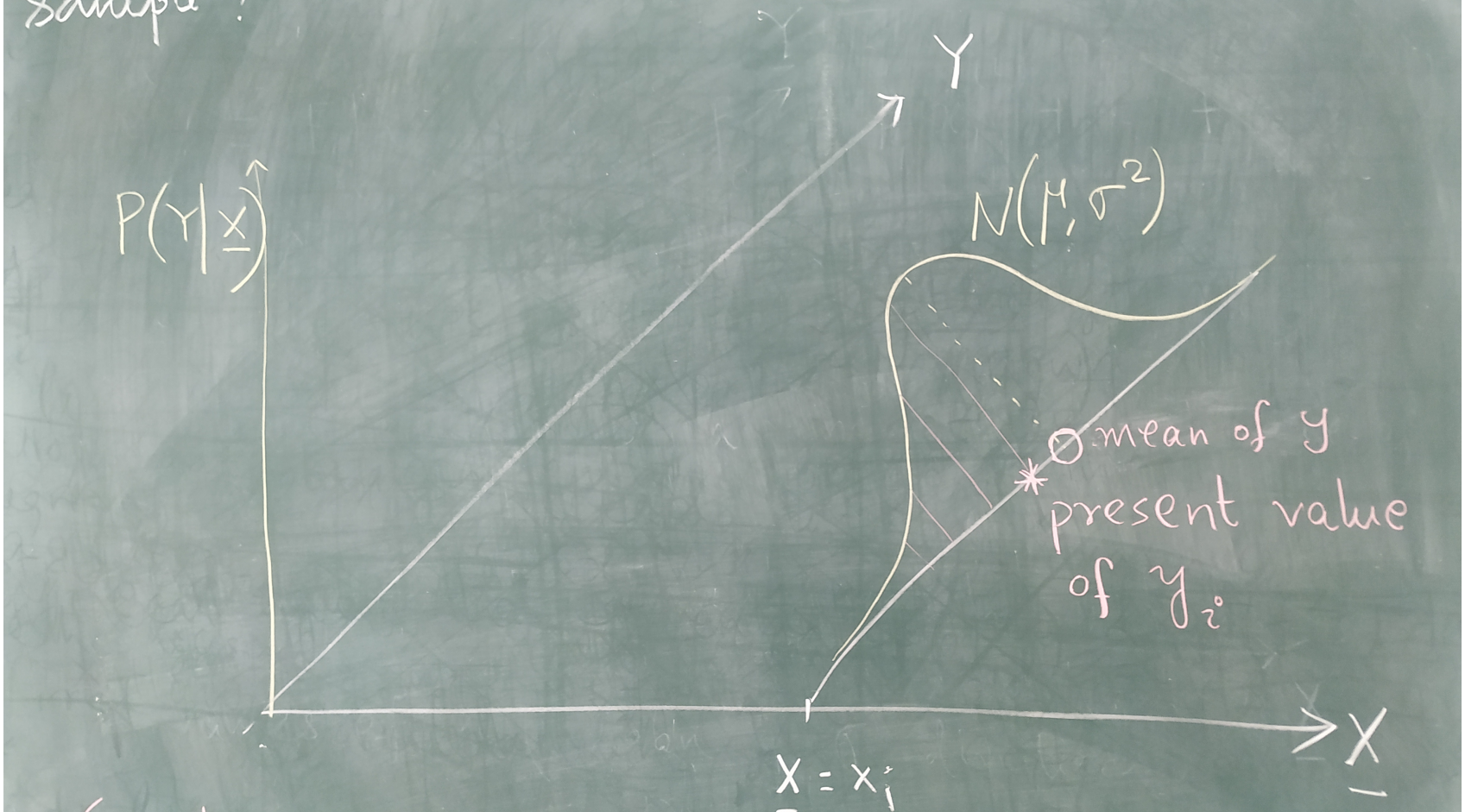
Also, possible to have linear predictor function as

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

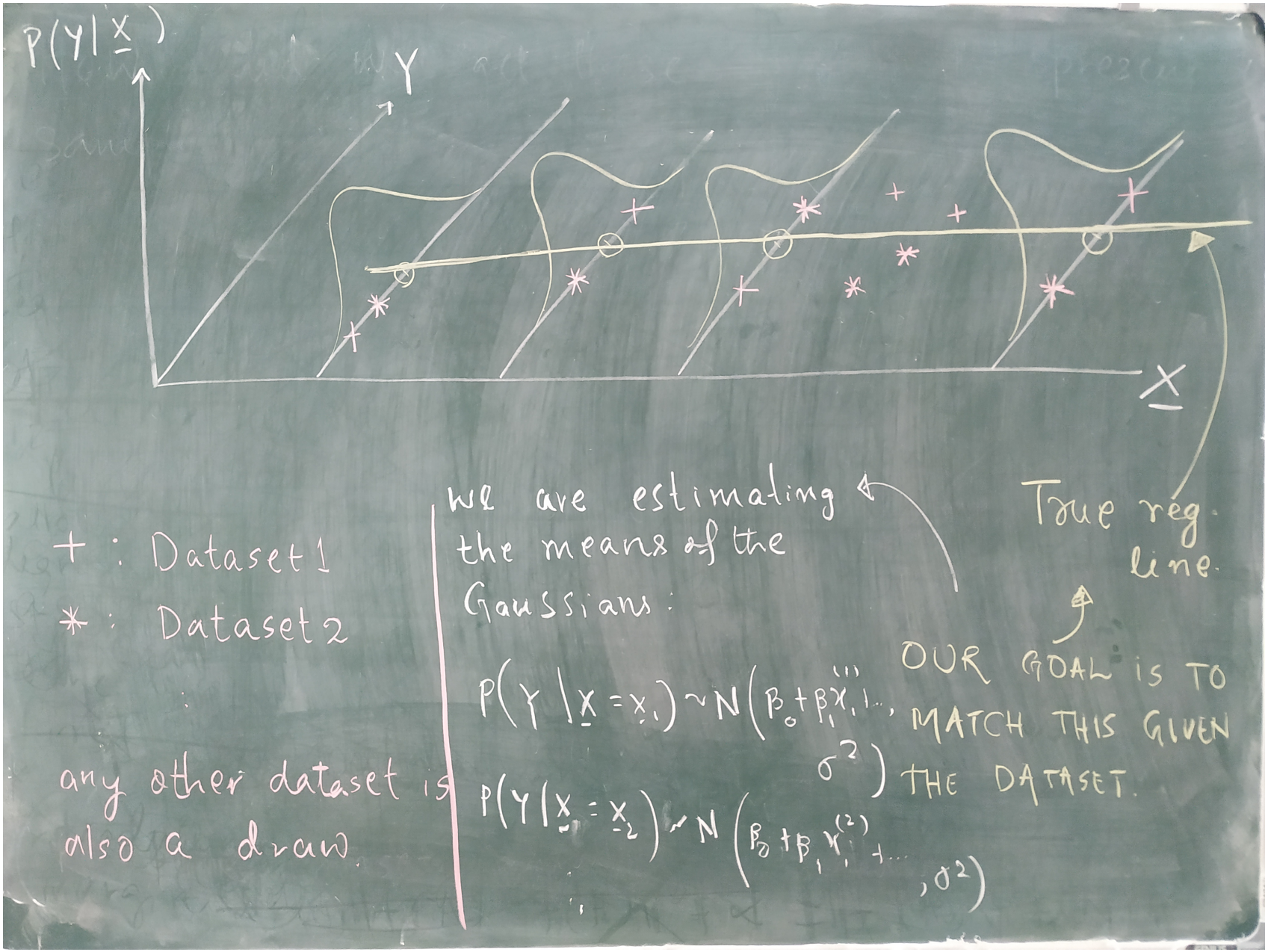
more clearly

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

How did we get these ys in the present sample?



So, the dataset contains: (x_i, drawn y_i) pairs


P(Y|X)
Y
X
+ : Dataset 1
* : Dataset 2
:
any other dataset is also a draw.
We are estimating the means of the Gaussians:
$P(Y|X=x_1) \sim N(\beta_0+\beta_1 x_1^{(1)}+\ldots, \sigma^2)$
$P(Y|X=x_2) \sim N(\beta_0+\beta_1 x_1^{(2)}+\ldots, \sigma^2)$
True reg. line.
OUR GOAL is TO MATCH THIS GIVEN THE DATASET.

How do we get these $\underline{\beta}$'

* Given N-data points as $(\underline{X}_i, y_i)$ pairs.

  Compute the Least Square Error or MSE :

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

* Use partial derivatives $\frac{\partial MSE}{\partial \underline{\beta}}$ and Solve..

(or) Use Gradient Descent (GD).

Modelling y as a Normal distro :-

A constant mean model :

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

Everything is known about Y.

0 mean, μ

1 variance, σ^2

Nothing to estimate here !!

In Linear Reg :

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

← Certain randomness around 0.

Let's now assume fixed μ and σ^2

Unknown

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

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

or $\boxed{Y \sim N(\mu, \sigma^2)}$

* How to get μ from data ?

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

Maximising Likelihood:

N data points

each x_i is drawn in i.i.d fashion

(i.i.d: independently identically distributed)

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

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

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

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

Maximise $L(\mu|y)$.

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

Equate to 0, solve.

Trick: Solving is easier with $\text{Log}(L)$.

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:

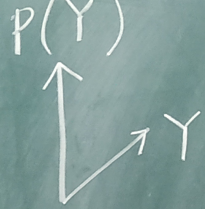
$$E[Y \mid \underline{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d.$$

Equivalently,

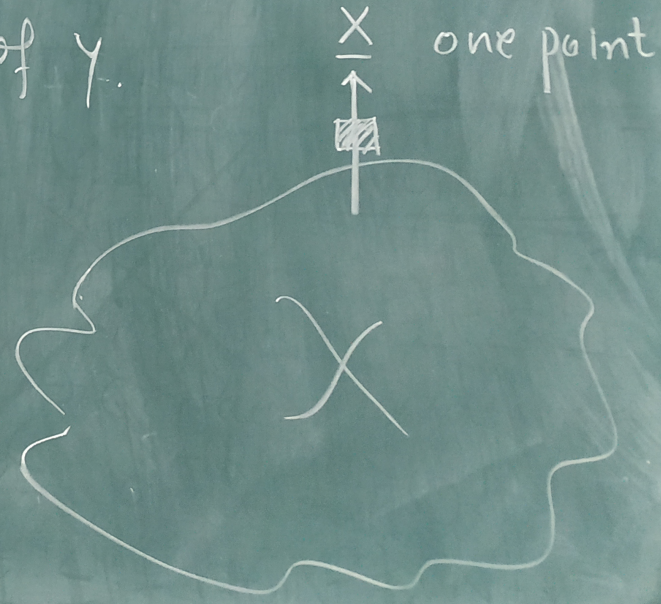
$$Y \sim N\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \,,\, \sigma^2\right)$$

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)



Why we did this way?

Because, there is always noise in y_i:

as $$y_i = \mu + \varepsilon$$

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

Our goal: obtain μ from data.

(MLE estimate)

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

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

$$= \arg\max_{\mu} \text{Log}\, \mathcal{L}(\mu|y)$$

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

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

For our β case:

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

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

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 $\pm$ σ diviations).