The Principle of Maximum Likelihood

Linear Regression : The Probabilistic Perspective

We can go a long way in Machine Learning without having to deal with the scary probabilistic expressions. But, in order to be able to read the bleeding edge research as soon as it is published, and not from a blog, months later, we need to able to speak the language of probability. Probability is a great way to express the relationship between the data and the model, concisely. We are familiar with how Linear Regression works from Andrew Ng's course (https://www.coursera.org/learn/machine-learning). We know the loss function just tries to minimize the quadratic distance between data points and the model (line). In this post, we will revisit Linear Regression from a probabilistic perspective, using a method known as the Maximum Likelihood estimation. We could apply this knowledge to any Neural Network based architecture.

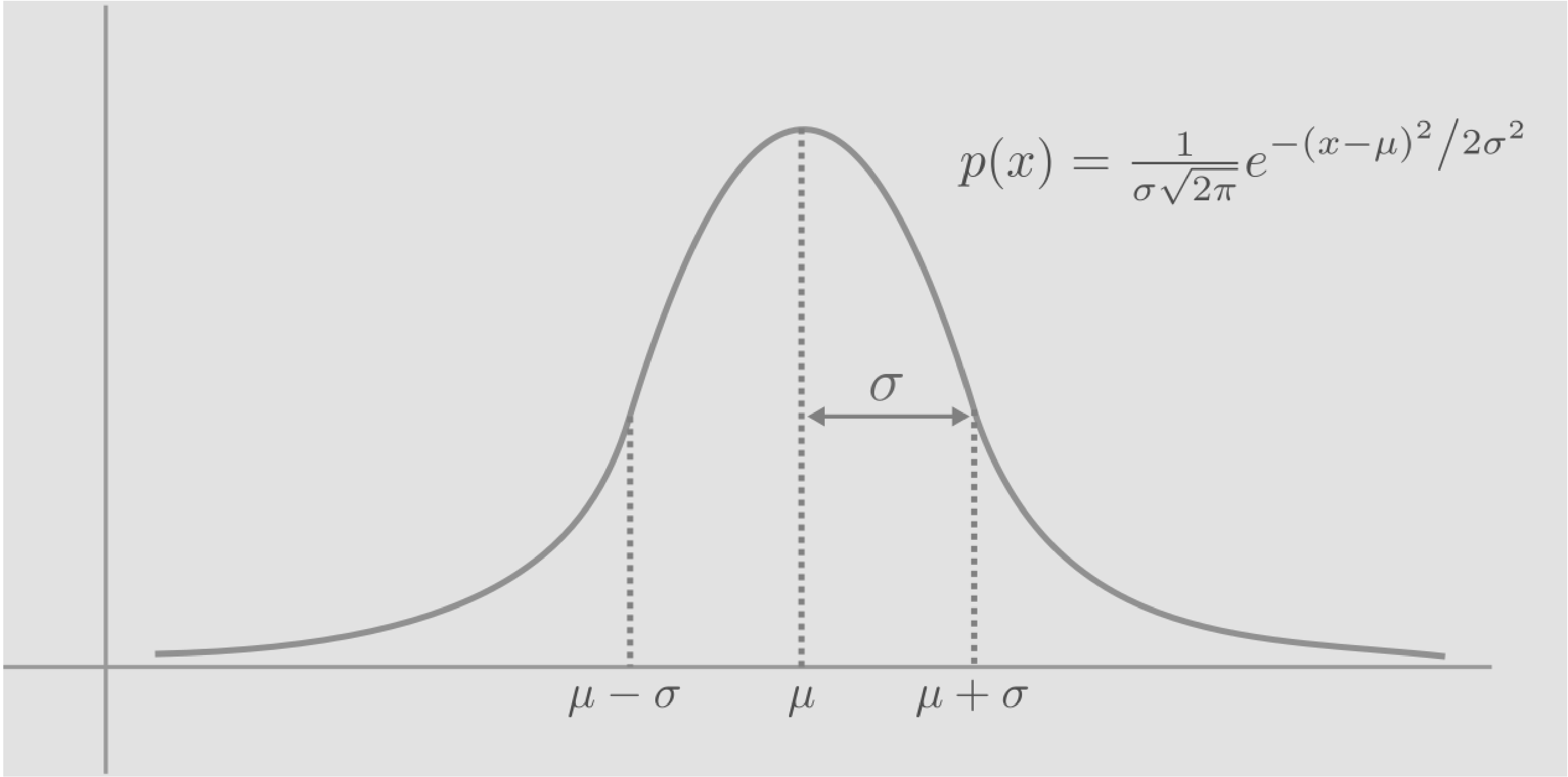
The principle of Maximum Likelihood is at the heart of Machine Learning. It guides us to find the best model in a search space of all models. In simple terms, Maximum Likelihood Estimation or MLE lets us choose a model (parameters) that explains the data (training set) better than all other models. For any given neural network architecture, the objective function can be derived based on the principle of Maximum Likelihood.

MLE is a tool based on probability. There are a few concepts in probability, that should be understood before diving into MLE. Probability is a framework for meauring and managing uncertainty. In machine learning, every inference we make, has some degree of uncertainty associated with it. It is essential for us to quantify this uncertainty. Let us start with the Gaussian Distribution.

Gaussian Distribution

A probability distribution is a function that provides us the probabilities of all possible outcomes of a stochastic process. It can be thought of, as a description of the stochastic process, in terms of the probabilities of events. The most commonly occurring distribution is the Gaussian Distribution or the Normal Distribution.

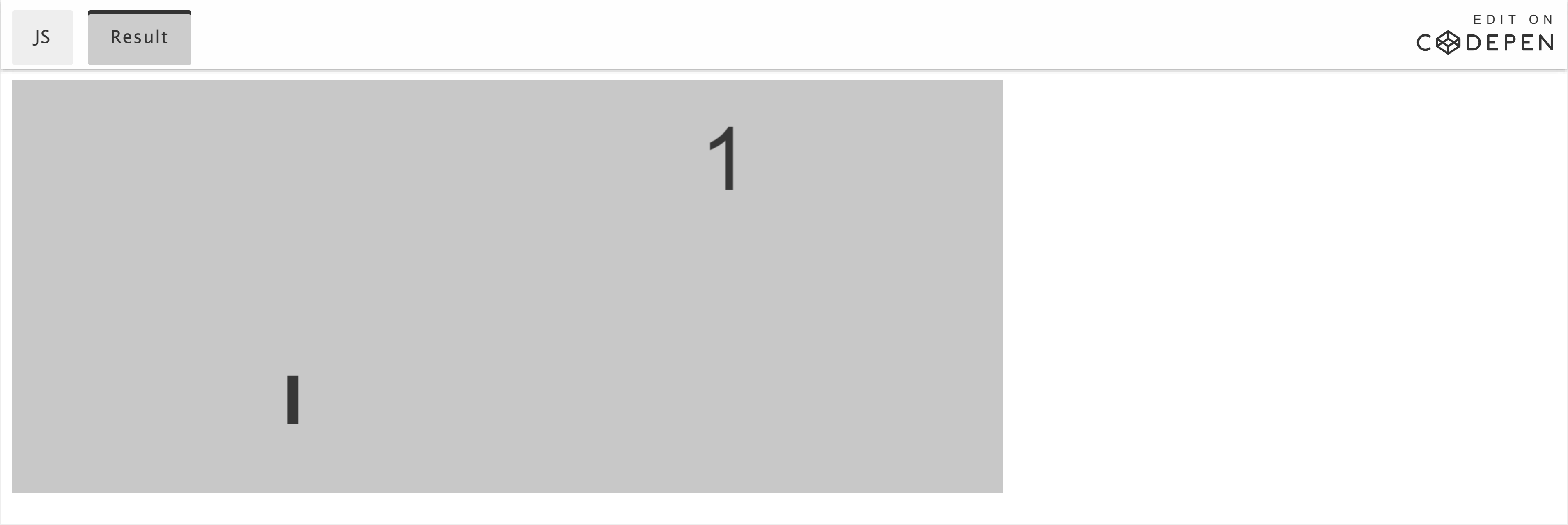
The gaussian distribution is a means to measure the uncertainty of a variable that is continuous between $-\infty$ and $+\infty$. The distribution is centered at mean, μ. The width depends on the parameter σ, the standard deviation (variance, σ^2). Naturally, area under the curve equals 1.


$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
σ
$\mu - \sigma$
μ
$\mu + \sigma$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

Let us take the example of coin toss, to understand the normal distribution. We take an unbiased coin and flip it ‘n’ times. We calculate the probability of occurance of 1 to ‘n’ heads for each ‘n’ value. In the animation below, each frame is an experiment and the number on right top corner denotes the number of flips in that experiment. Each experiment involves flipping the coin ‘n’ times. We observe the probability of getting 1, 2,…‘n’ heads for each experiment, and plot it.

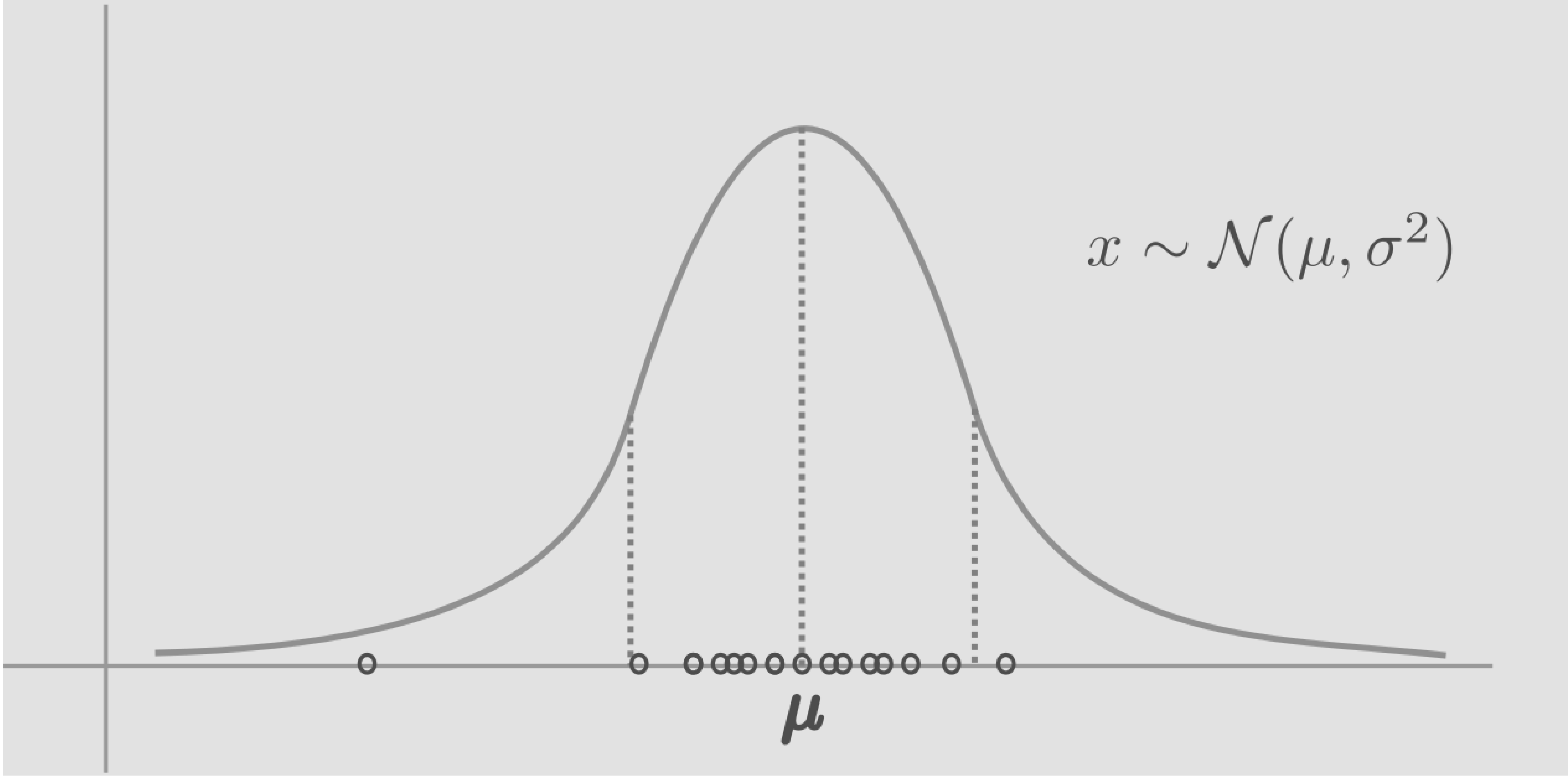


As you can see, with increase in number of flips (experiments), the curve of probability distribution tends to assume the shape of a normal distribution, represented by the equation above. In each of the experiments, the peak probability happens at half the number of heads and the probability density tends to decay on both sides. This is basically due to the fact that, there are more possible ways for the results to be close to half heads and half tails, compared to number of heads dominating the number of tail or vice versa.

That was fun to watch but how is this relevant to linear regression or machine learning? The data points in the training set, do not accurately represent the original data generating distribution or process. Hence we consider the process stochastic and build our model to accomodate a certain level of uncertainty. Every data point can be considered a random variable sampled from the data generating distribution which we assume to be gaussian. By that logic, learning or training is basically recreating the original distribution, that generated the training data.

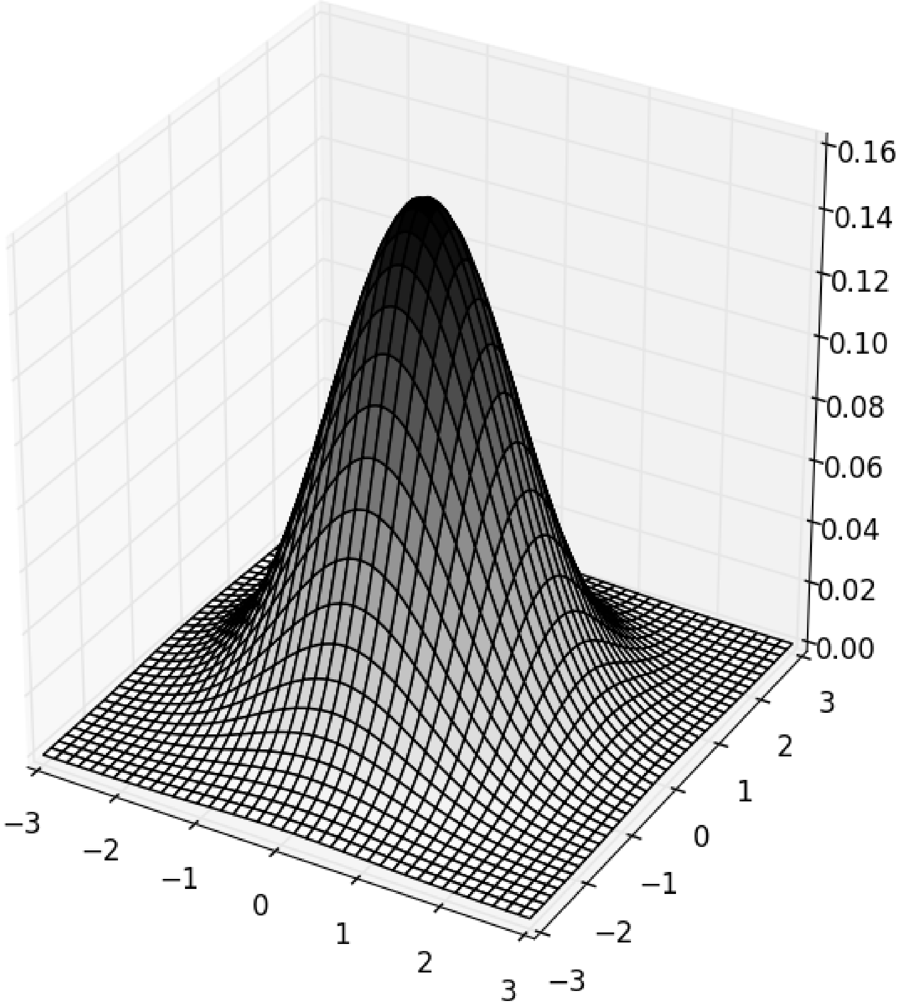
Random Sampling

The process of sampling from a normal distribution is expressed as, $x \sim \mathcal{N}(\mu, \sigma^2)$. ‘x’ is a random variable sampled or generated or simulated from the guassian distribution. As we sample from this distribution, most samples will fall around the center, near the mean, because of higher probability density in the center.



Multivariate Gaussian

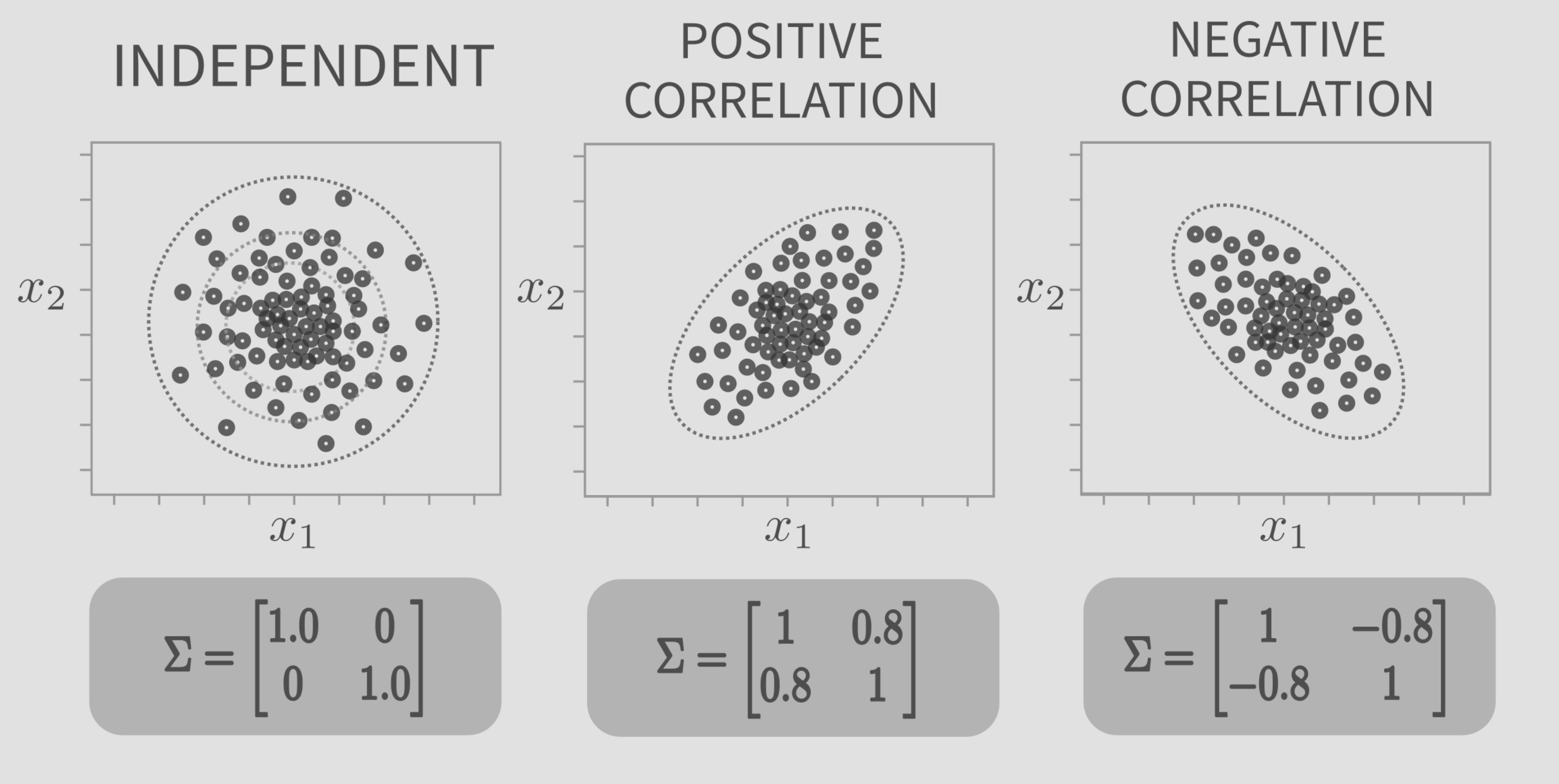
The 2D gaussian distribution or bivariate distribution, consists of 2 random variables x_1 and x_2. It can have many different shapes. The shape depends on the covariance matrix, Σ. The Multivariate Gaussian Distribution is a generalization of bivariate distribution, for ‘n’ dimensions.



$$p(y) = |2\pi\Sigma|^{-1/2} e^{-1/2(y-\mu)^T \Sigma^{-1}(y-\mu)}$$

In the above expression, **y** and μ are vectors of ‘n’ dimensions, and Σ is a matrix of shape ‘nxn’.

The figure below presents the top view of bivariate gaussian. The smaller circles denote the data points sampled from the distribution. Notice the variation in the shape of the gaussian with Σ. The mean, given by x_1 and x_2 (μ_1 and μ_2) determine the center of the gaussian, while Σ determines the shape.



Maximum Likelihood

I am borrowing this amazing toy example from Nando de Fretais's lecture (https://www.youtube.com/watch?v=voN8omBe2r4), to illustrate the principle of maximum likelihood. Consider 3 data points, $y_1 = 1, y_2 = 0.5, y_3 = 1.5$, which are independent and drawn from a guassian with unknown mean θ and variance 1. Let's say we have two choices for θ : {1, 2.5}. Which would you choose? Which model (θ) would explain the data better?

In general, any data point drawn from a gaussian with mean θ and and variance 1, can be written as,

$$y_i \sim \mathcal{N}(\theta, 1) = \theta + \mathcal{N}(0, 1)$$

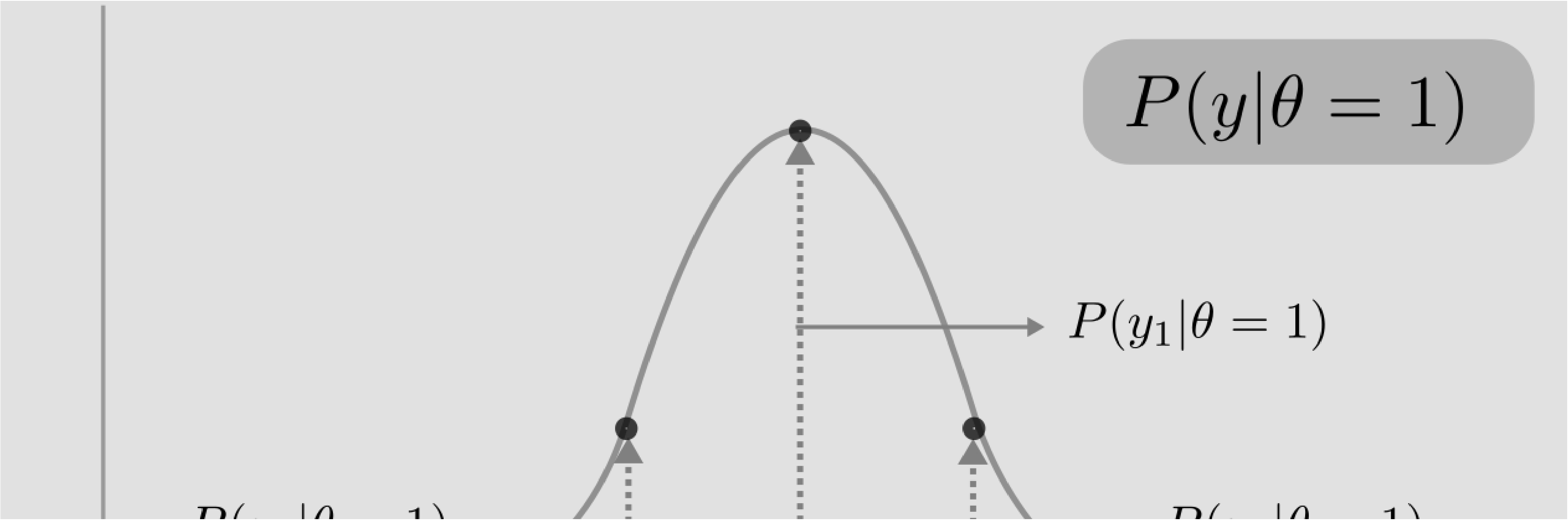
θ, the mean, shifts the center of the standard normal distribution ($\mu = 0$ and $\sigma^2 = 1$).

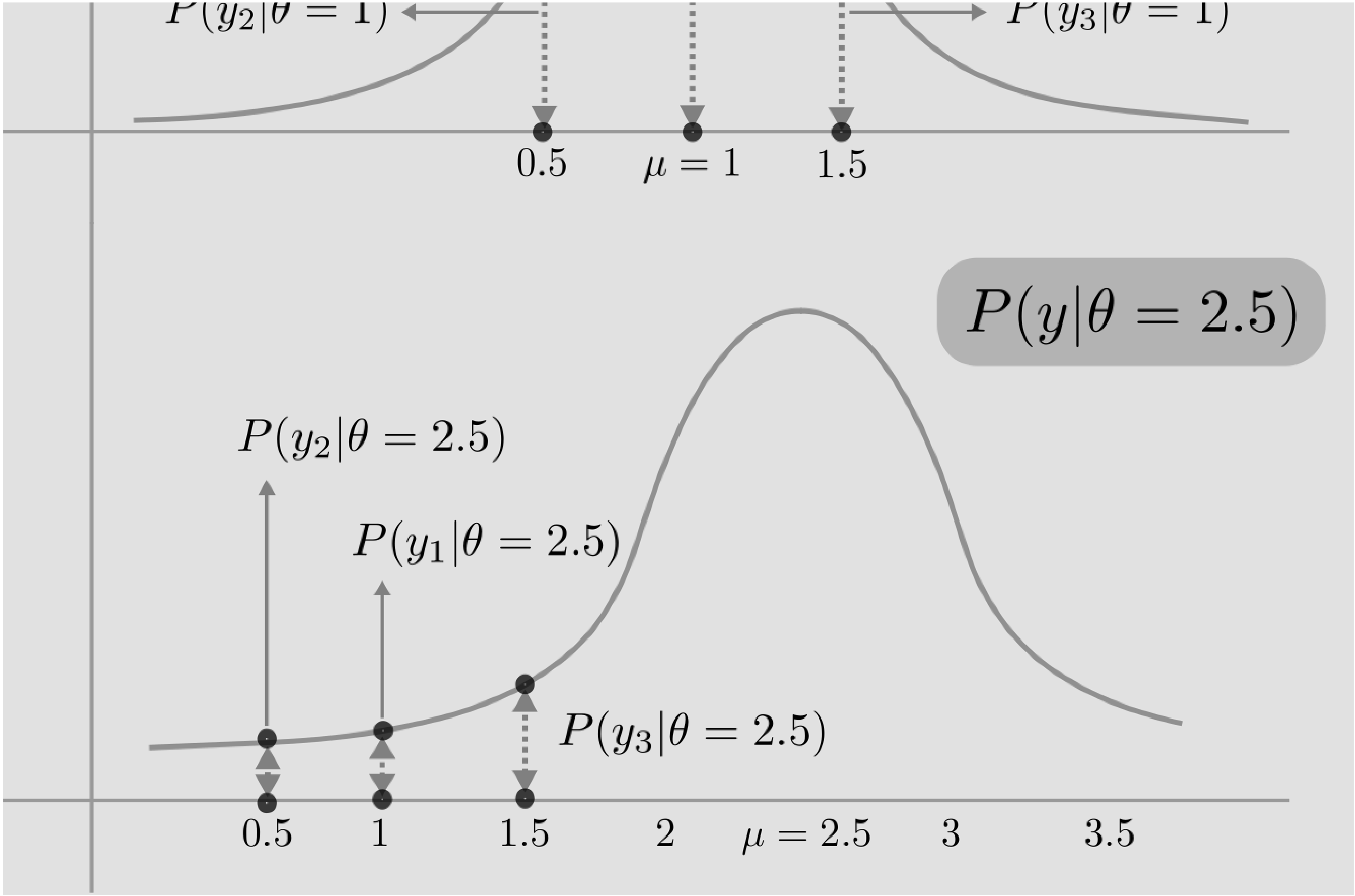
The likelihood of data (y_1, y_2, y_3) having been drawn from $\mathcal{N}(\theta, 1)$, can be defined as,

$$P(y1, y2, y3|\theta) = P(y1|\theta)P(y2|\theta)P(y3|\theta)$$

as y_1, y_2, y_3 are independent.

Now, we have two normal distributions defined by θ = 1 and θ = 2.5. Let us draw both and plot the data points. In the figure below, notice the dotted lines that connect the bell curve to the data points. Consider the point $y_2 = 0.5$ in the first distribution($\mathcal{N}(\mu = 1, \sigma^2 = 1)$). The length of the dotted line gives the probability of the $y_2 = 0.5$ being drawn from $\mathcal{N}(\mu = 1, \sigma^2 = 1)$.





The likelihood of data (y_1, y_2, y_3) having been drawn from $\mathcal{N}(\mu = 1, \sigma^2 = 1)$, is given by,

$$P(y1, y2, y3|\theta = 1) = P(y1|\theta = 1)P(y2|\theta = 1)P(y3|\theta = 1)$$

The individual probabilities in the equation above, are equal to the heights of corresponding dotted lines in the figure. We see that the likelihood, given by the product of individual probabilities of data points given model, is basically the product of lengths of dotted lines. It is obvious that the likelihood of model $\theta = 1$ is higher. We choose the model ($\theta = 1$), that maximizes the likelihood.

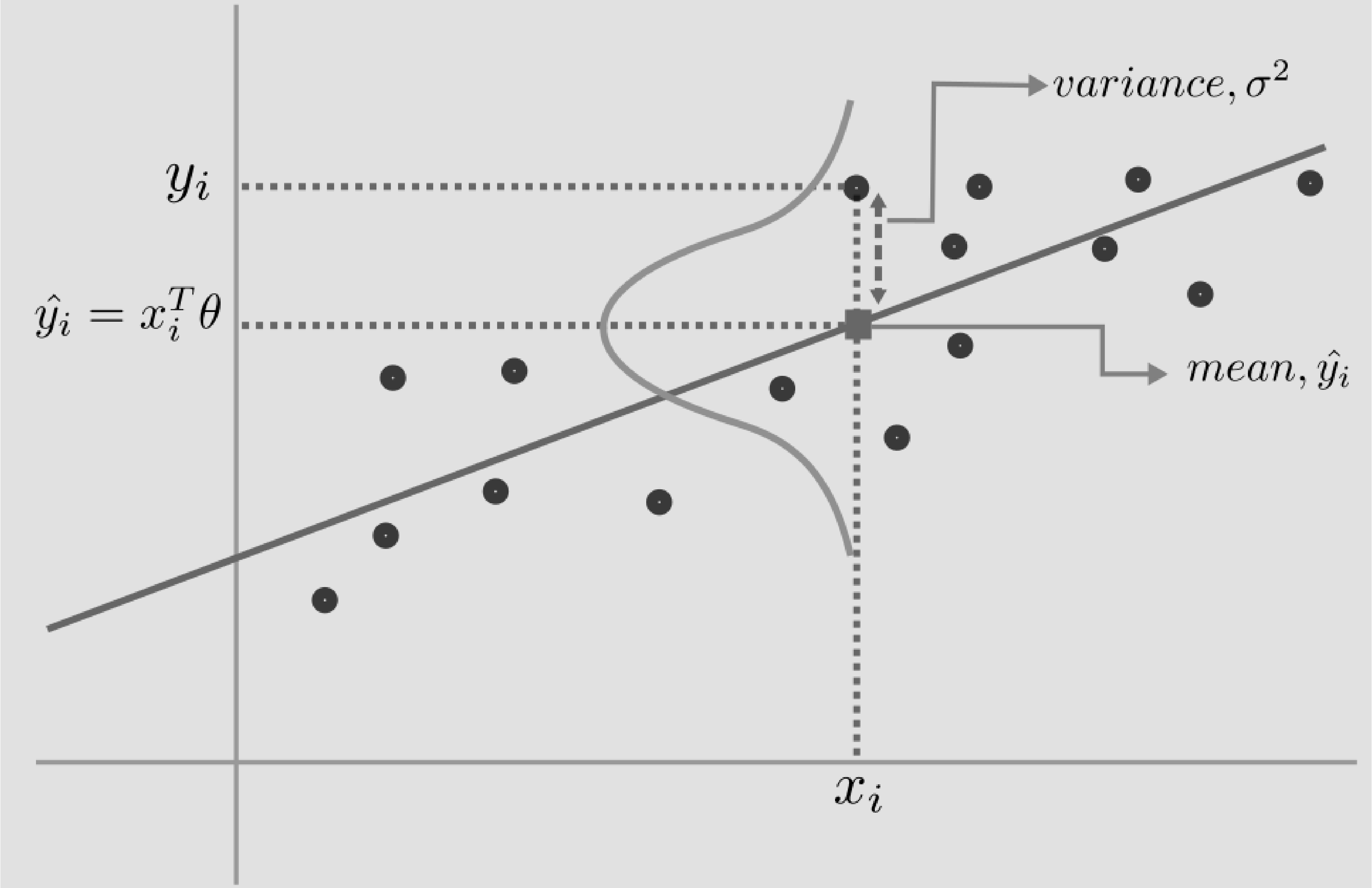
Linear Regression

We are ready to apply MLE to linear regression, where the objective is to find the best line that fits the data points. But first, let us make some assumptions. We assume each label, y_i, is gaussian distributed with mean, $x_i^T\theta$ and variance, σ^2, given by

$$y_i = \mathcal{N}(x_i^T\theta, \sigma^2) = x_i^T\theta + \mathcal{N}(0, \sigma^2)$$
$$prediction, \hat{y}_i = x_i^T\theta$$

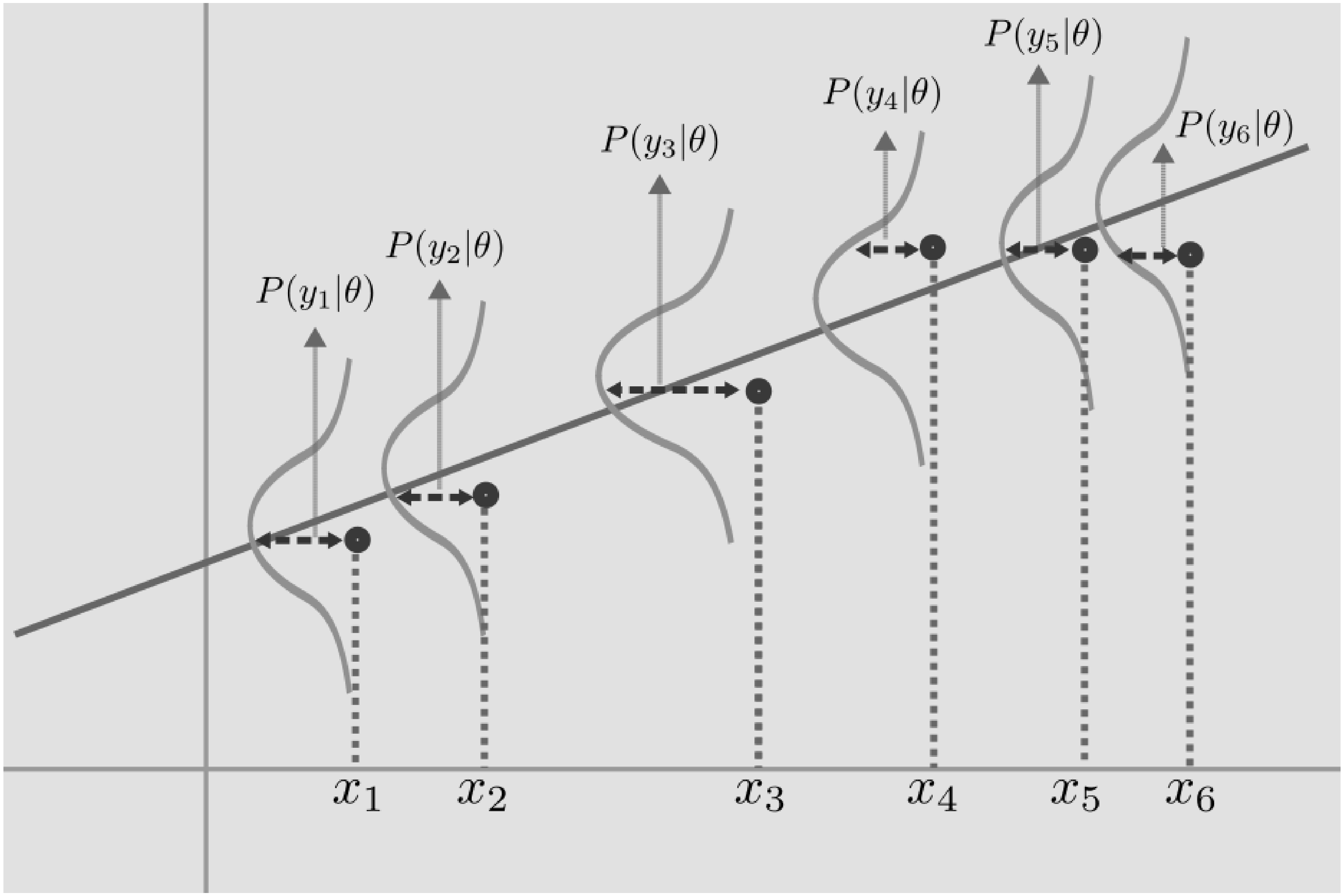
where each x_i is a vector of form ($x_i^1 = 1, x_i^2$).

The mean, $x_i^T\theta$ represents the best fit line. The data points will vary about the line, and the second term, captures this variance, $\mathcal{N}(0, \sigma^2)$ (see figure below).



Learning

If we assume that each point y_i is gaussian distributed, the process of learning becomes the process of maximizing the product of the individual probabilities, which is equivalent to maximizing the log likelihood. We switch to log space, as it is more convenient and it removes the exponential in the gaussian distribution.



As the data points are independent, we can write the joint probability distribution of y, θ, σ as,

$$p(y|X, \theta, \sigma) = \prod_{i=1}^{n} p(y_i|x_i, \theta, \sigma)$$

$$p(y|X, \theta, \sigma) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\theta)^2}$$

rewriting in vector form,

$$p(y|X,\theta,\sigma) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y-X\theta)^T(y-X\theta)}$$

Log likelihood,

$$l(\theta) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta)$$

The first term is a constant and the second term is a parabola, the peak (maxima) of which can be found by equating the derivative of $l(\theta)$ to zero. Equating first derivative to zero, we get,

$$\frac{dl(\theta)}{d\theta} = 0 = -\frac{1}{2\sigma^2}(0 - 2X^TY + X^TX\theta)$$

we get,

$$\hat{\theta_{ML}} = (X^TX)^{-1}X^TY$$

Finally, we reach our goal of finding the best model for linear regression. This equation is commonly known as the normal equation. The same equation can be derived using the least squares method (perhaps in another post).

Similarly, we can get the maximum likelihood of variance, σ^2, by differentiating log likelihood with respect to σ and equating to zero.

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(Y - X\theta)^T(Y - X\theta) = \frac{1}{n}\sum_{i=1} n(y_i - x_i\theta)^2$$

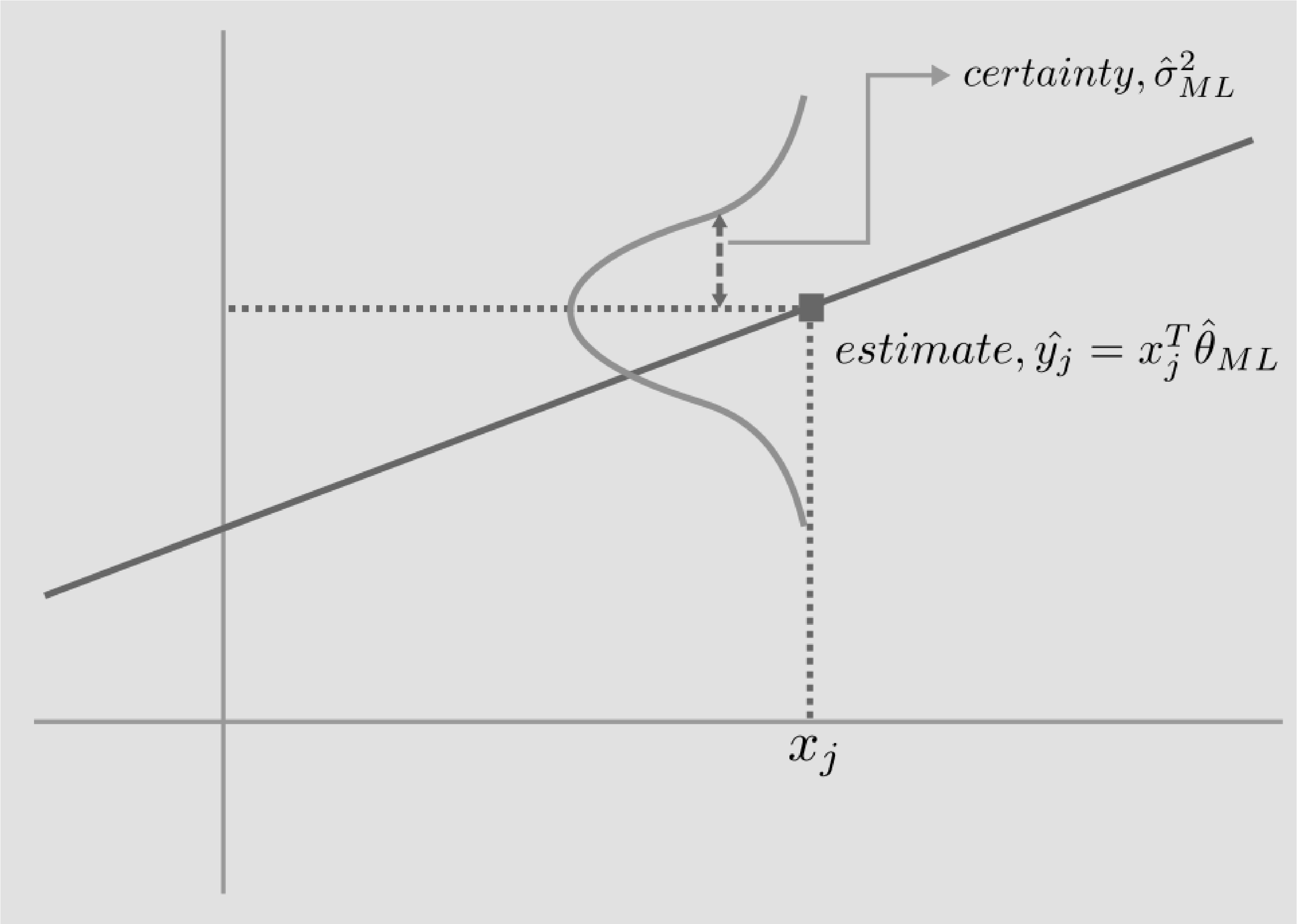
This gives us the standard estimate of variance in the training data.

Inference

So, we have learned the parameters of our model, $\hat{\theta}_{ML}$ and $\hat{\sigma}^2_{ML}$. Based on these learned parameters, we can make inferences for new unseen data. For a new data point x_j, our best estimate is given by,

$$estimate, \hat{y}_j = x_j^T \hat{\theta}_{ML}$$

The degree of uncertainty in our estimate, is given by the variance, $\hat{\sigma}^2_{ML}$.


$certainty, \hat{\sigma}^2_{ML}$
$estimate, \hat{y_j} = x_j^T \hat{\theta}_{ML}$
x_j

Reference

1. Andrew Ng's Machine Learning course (https://www.coursera.org/learn/machine-learning)
2. Nando de Freitas's Machine Learning course (https://www.youtube.com/playlist?list=PLE6Wd9FR--EdyJ5lbFl8UuGjecvVw66F6)
3. Maximum Likelihood and Linear Regression (https://www.youtube.com/watch?v=voN8omBe2r4&index=3&list=PLE6Wd9FR--EdyJ5lbFl8UuGjecvVw66F6)

4. Coin Toss Experiment (https://www.fourmilab.ch/rpkp/experiments/statistics.html)
5. Bivariate Gaussian in matplotlib (http://mpastell.com/2013/05/02/matplotlib_colormaps/)

I hope to do a follow up post very soon, applying maximum likelihood to more complex neural networks, and also introduce the concept of Bayesian Learning and compare it with the Frequentist approach.

Feel free to drop a comment.

4 Comments Scientia est Potentia 1 Login

♡ Recommend 1 Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Park Ji Ho • 10 months ago

Hi, thanks for the post. i find it truly valuable!

I was wondering in the gaussian curve under the heading "Maximum Likelihood"

shouldn't it be P(y1 = 1 | theta), P(y2 = 0.5 | theta), P(y3 = 1.5 | theta) instead of P(y1| theta=1), P(y1| theta=0.5), P(y1| theta=1.5)?

and the x-axis should be y rather than /mu?

I just want to ask if I am misunderstanding anything.

1 • Reply • Share ›

Suriyadeepan Ramamoorthy Mod → Park Ji Ho • 10 months ago

I'm glad you find it useful. And yes, you are right about the probabilities. I have fixed the figure. The x-axis is indeed, in terms of 'y'. \mu marks a point in the x-axis, which corresponds to the mean (\theta) of the gaussian curve.

• Reply • Share ›

cosmobird • a year ago

neat...
little bit curve is still needed for learners in statistics, ML...
but i could see your perseverance in communicating the interpretation of MLE.

but I could see your perseverance in communicating the interpretation of MLE.
expecting a follow up post on bayesian... and its interpretation.
:)

1 ^ | ˅ • Reply • Share ›

deathmetal124 • a month ago

Amazing post man. Thank you so much.

I have a comment on this line - "The individual probabilities in the equation above, are equal to the heights of corresponding dotted lines in the figure."
My understanding is that in continuous distributions, the 'probability' of any particular 'x' is 0 and it can only be understood as P(X>x) and not P(X=x). Because you can have steep probability distributions, but it does not mean that the probability at a particular point is high (like 100) since probability can only be between 0 and 1.

^ | ˅ • Reply • Share ›

ALSO ON **SCIENTIA EST POTENTIA**

Chatbots with Seq2Seq

15 comments • a year ago

Avatar **Victor Mishin** — Yes, with your config TensorFlow will use 6GB GPU memory and then Crash with Out-Of-Memory Errors. How I know? I have ...

Unfolding RNNs

4 comments • a year ago

Avatar **Soumen Madan** — This is cool. You should also have a look at http://r2rt.com/ The tutorials at this website are very comprehensive

Get Started with Machine Learning

2 comments • 2 years ago

Avatar **Sang-Kil Park** — Thanks for your great summary! But there are some errors in multiplication.First, `*` in Numpy is element-wise multiplication. ...

Practical seq2seq

30 comments • a year ago

Avatar **Pratik Kataria** — How can I test the model for custom input string and get an output ?

✉ Subscribe Ⓓ Add Disqus to your siteAdd DisqusAdd 🔒 Privacy DISQUS

● (https://www.facebook.com/suriya.geek.deepan) ● (https://github.com/suriyadeepan) ● (https://twitter.com/suriya_cosmist)
● (mailto:suriyadeepan.r@gmail.com)

Suriyadeepan Ramamoorthy • 2017 • suriyadeepan.github.io (http://suriyadeepan.github.io)

Theme by beautiful-jekyll (http://deanattali.com/beautiful-jekyll/)