Recap

$(x_1, t_1), (x_2, t_2) \ldots (x_N, t_N)$

$y = w_0 + w_1 x$

RMSE

find rmse of all polynomial fits of degree $0, 1, 2 \ldots 9$
take min [this is for testing btw]

$\Rightarrow$ Next : Overfitting
if data fits training very well, but fuck up testing

Solution: Regularization
↳ Bound the weights
make it an unconstrained optimization problem,
You get the model to depend on a regularization parameter λ, you (find out testing error & keep experimenting)

Question: Why? we can just find best degree no?

Ans: Only with good number of examples can we find the best degree!

We may definitely end up overfitting, even with the "correct" degree due to less number of examples

(from domain expert)
eg: 10 examples, 9 degree polynomial required will give overfit!!

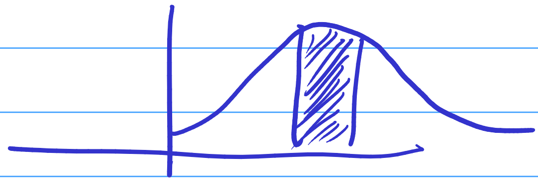
So that testing may be better

So if we use 3 degrees instead, we may end up " training overfitting'
So wee need to use regularized polynomial of 9 degrees

This is one way of solving a regression problem

Using Bayesian regression model

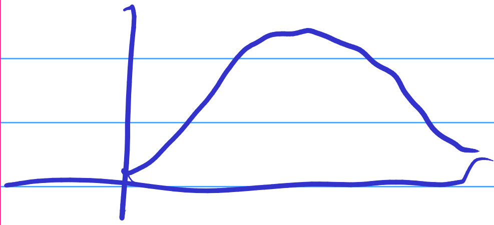
Gaussian distribution

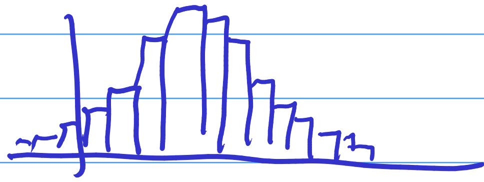


We can see data is distributed in a probability distribution

So for our sake, if we can get an idea on the distribution type, and get a sample of data

Then determine the parameters of the distributions





μ, σ^2 can only be estimated, they cannot be exact

How do we get estimates??

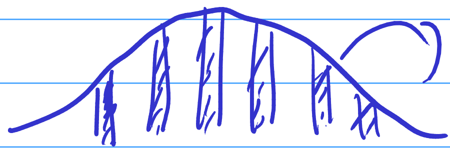
Assuming that your training data follows gaussian

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

How do we find μ, σ^2 (estimates $\hat{\mu}$ $\hat{\sigma}^2$)

Consider 2 distributions!

for 1 of them, find: $p(x_1 \in (x_1-\delta, x+\delta),\ x_2 \in (x_2-\delta, x_2+\delta), \dots x_n \in (x_n-\delta, x_n+\delta))$



find the probability of these intervals

$\Rightarrow$ find wrt to both distributions and the higher one is better

so let's generalize!

Find this for all possible μ & σ^2

OR maximize the probability (MLE)

$$= \prod_{n=1}^{N} p(x \in (x_n-\delta, x_n+\delta))$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \int_{x_n-\delta}^{x_n+\delta} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx$$

$\delta = 10^{-20}$

$\approx$ height x width $= 2\delta \times f(x)$

$$\approx \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, 2\delta$$

we need the μ, σ to maximise.

$$\max\ (2\delta)^n \frac{1}{(\sqrt{2\pi}\,\sigma)^N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{n}(x_n-\mu)^2}$$

So just apply MLE!

$$= \max \log \left(\cdots e^{-\frac{1}{2\sigma^2}} \cdots \right)$$

$$= \max_{\mu, \sigma^2} -\frac{N}{2} \log 2\pi - N \log \sigma - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

$$= \max_{\mu, \sigma^2} LL(\mu, \sigma^2) \quad \text{log-likelihood}$$

$$\frac{\partial L}{\partial \mu} = 0 \rightarrow \frac{-1}{2\sigma^2} \frac{\partial}{\partial \mu} \left(\sum_{n=1}^{N} (x_n - \mu)^2 \right) = 0$$

$\boxed{\text{and}}$

$$\frac{\partial L}{\partial \sigma} = 0 \qquad \Rightarrow \sum_{n=1}^{N} (x_n - \mu) = 0$$

$$-\frac{N}{\sigma} - \frac{2\sum (x_n - \mu)^2}{2\sigma^3} = 0 \qquad \Rightarrow \left(\sum_{n=1}^{N} x_n \right) - n\mu = 0$$

$$\sigma^{-2} (1 + \sigma^{-3}) \sum \cdots = 0 \qquad \Rightarrow \frac{1}{N} \sum_{n=1}^{N} x_n = \mu$$

$$\underset{\sigma^2}{\text{Variance}} \propto \frac{1}{\underset{1/\beta}{\text{precision}}} \qquad \frac{1}{\sqrt{2\pi} \frac{1}{\sqrt{\beta}}} e^{-\frac{\beta}{2}(x-\mu)^2}$$

$$t \sim N(y(x, \omega), \beta)$$

↳ not exactly

Now we will solve a regression problem with a probabilistic approach to find x, ω, where our target attributes having a normal dist. each, with mean as our prediction!

will have same minimization function as MSE too!