2

11 January 2026 16:39

- Data contains information, and we assume it is generated from some unknown underlying distribution (this is why we need ML).
- Machine Learning = learning patterns of this unknown distribution from data.
- Information = surprise:
  - If a data point was fully expected → low surprise → low information
  - If it reveals something new → high surprise → high information
- Seeing a data point and learning something new from it = gaining information.
- This idea of "surprise = information" is the foundation for information theory concepts like entropy (average surprise).
- The same idea applies to any data type: images, text, audio, etc.
- Many ML terms later (entropy, loss, likelihood) are built on this basic idea

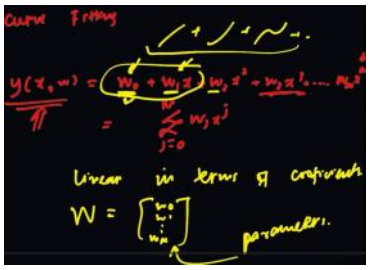
Real world data has : Pattern and Noise, and pattern is not direclty inferrable becaue there is random noise . Our goal is to capture pattern.

In real life we don't know how data is genreated, for understanding purpose will assume we know true process, so that we can see what model is supposed to learn, for example: lets take sin function and add noise to it, (synthetic data)

Link: code and explaination for sine prediction

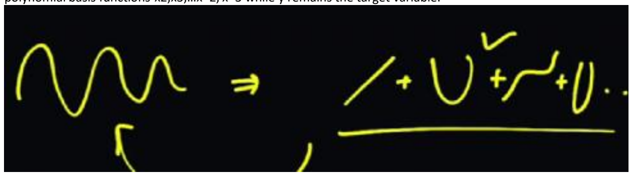
In reality, data comes from an unknown and potentially infinite population/distribution, but we only observe a finite dataset.
From this limited data, we must generalize to unseen data, not simply memorize the given samples.
This is inherently difficult because finite samples may not fully represent the true distribution.
Sampling data also involves uncertainty, and probability theory provides a mathematical framework to quantify this uncertainty.
Once we obtain a probabilistic representation of the data (i.e., uncertainty is modeled), we still need to make predictions or decisions.
Decision theory tells us how to make optimal predictions under uncertainty.

Curve fitting means:
Finding a function that best fits (approximates) the given data points.

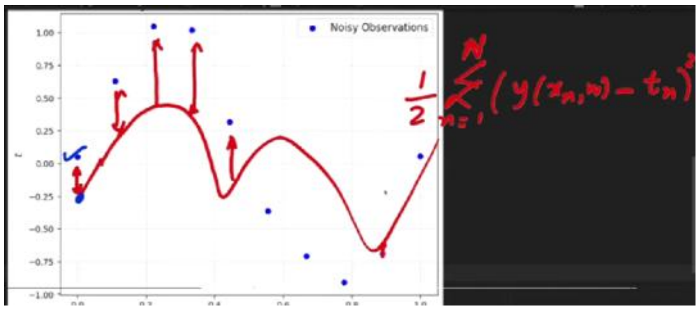


Curve Fitting
$y(x,w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \dots w_M x^M$
$= \sum_{j=0}^{M} w_j x^j$
Linear in terms of coefficients
$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}$
parameters.

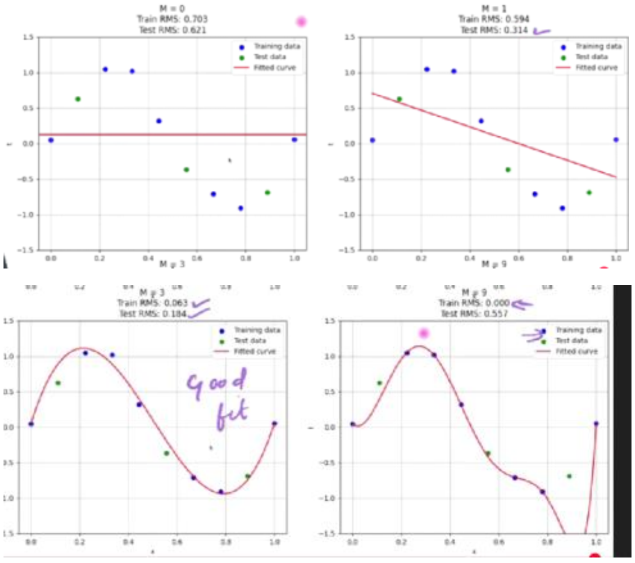
there is only one original feature xxx; higher-dimensional feature space is created by applying polynomial basis functions x2,x3,…x^2, x^3 while y remains the target variable.



How do we know our curve is good?
we get W* , set of W for which error function is minized.
it should not memorize (less error during training, but sucks in testing), it should generalize



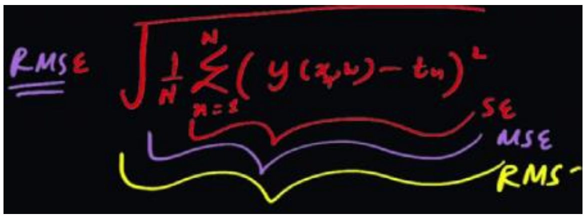
code

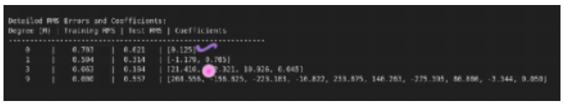


We can see, if we increase the number of parameters, its overfitting, so shall we reduce the number of paramters but it would limit th learning cabaplity

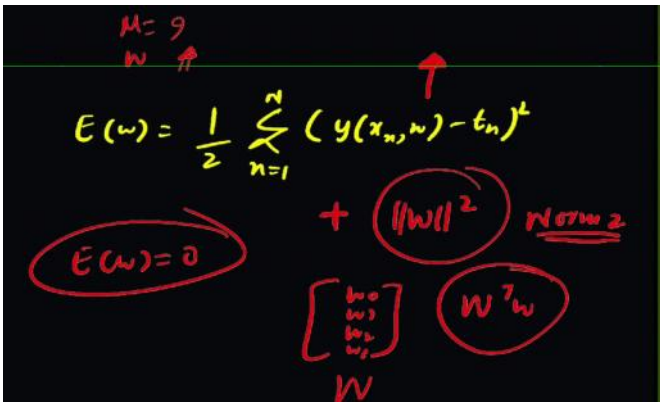
but isnt higher degree polynomial superset of of lower degree polynomial?
- ➢ It might fit the noise too.
- ➢ Goal is not find best fit in training , but to generalize

(the reason is its having very high coefficient values, so if we can maintain or penalize it to have lower coeffiient values we are good with high parameters)





```
Detailed RMS Errors and Coefficients:
Degree (M) | Training RMS | Test RMS | Coefficients
---------------------------------------------------------------
    0      |    0.703     |  0.621   | [0.125]
    1      |    0.594     |  0.314   | [-1.179, 0.705]
    3      |    0.063     |  0.184   | [21.410, [illegible]2.321, 10.926, 0.045]
    9      |    0.000     |  0.557   | [288.556, -158.825, -223.183, -16.822, 233.875, 146.283, -275.395, 86.886, -3.344, 0.058]
```

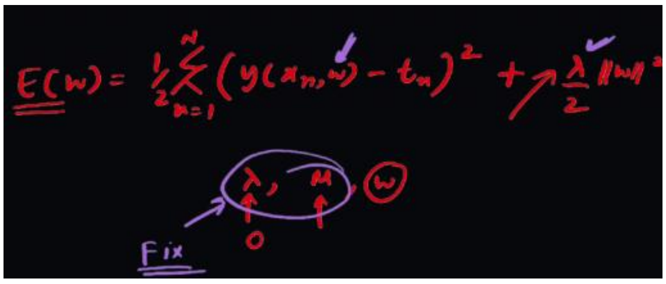


M= 9
w ↑
$E(w) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)^2$
$+ \|w\|^2$
Norm 2
$E(w) = 0$
$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$
W
$w^T w$

This is called regularization, M = no. of degree is hyperparameter, fixed during training

Regularization says: "Don't use large weights unless really necessary."

λ controls how much we punish complexity.



$$\underline{\underline{E(w)}} = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, w) - t_n\right)^2 + \frac{\lambda}{2}\|w\|^2$$

$$\lambda, \; M, \; \boxed{w}$$

$\lambda \uparrow 0$, $M \uparrow$

Fix