1(i)

Curse of dimensionality

From Wikipedia, the free encyclopedia

The **curse of dimensionality** refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. The expression was coined by Richard E. Bellman when considering problems in dynamic programming.[1][2]

Dimensionally cursed phenomena occur in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining and databases. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. In order to obtain a reliable result, the amount of data needed often grows exponentially with the dimensionality. Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data, however, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient.

Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data. *(It's just like trying to fit undersized pants!)* Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied to such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

In a nutshell, Underfitting refers to a model that can neither performs well on the training data nor generalize to new data.

Reasons for Underfitting:

1. High bias and low variance
2. The size of the training dataset used is not enough.
3. The model is too simple.
4. Training data is not cleaned and also contains noise in it.

1 (ii)

Techniques to reduce underfitting:

1. Increase model complexity
2. Increase the number of features, performing feature engineering
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results.

Overfitting: A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

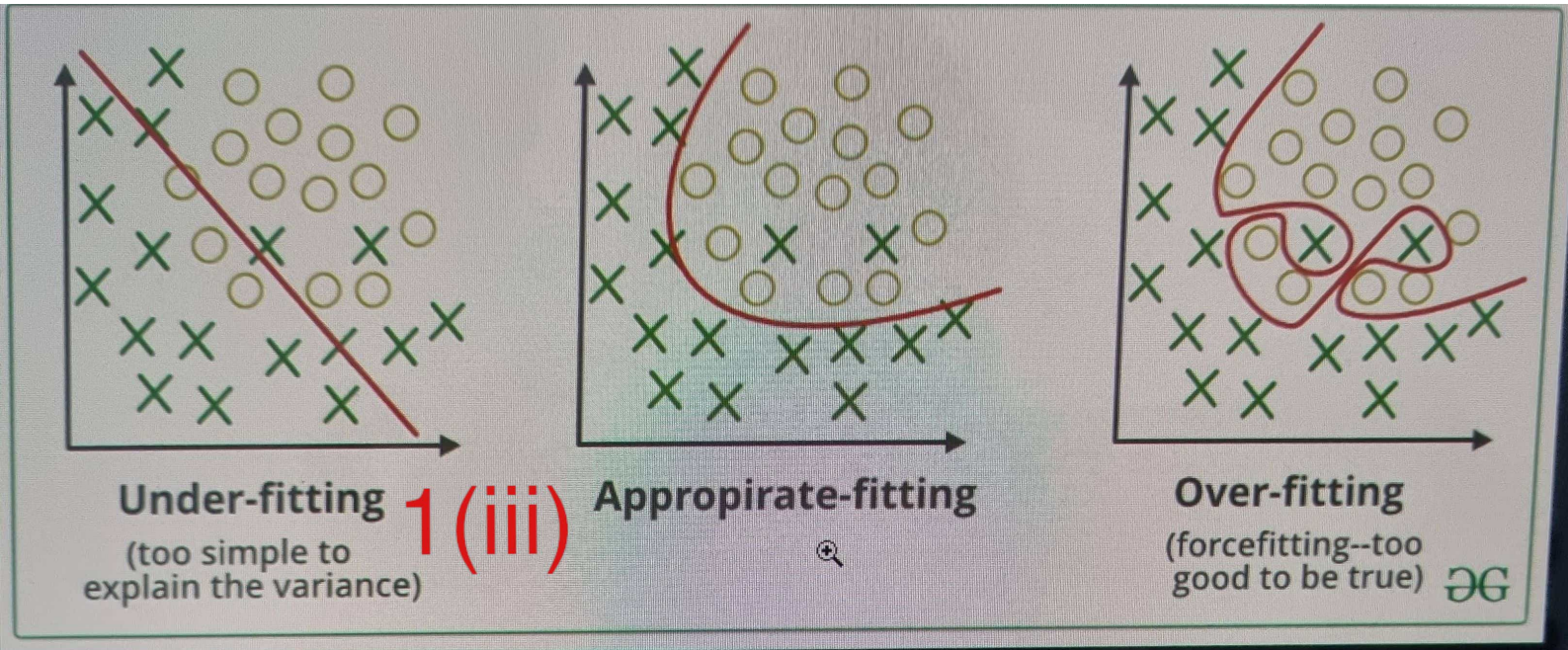
In a nutshell, Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

Reasons for Overfitting are as follows:

1 (iii)

1. High variance and low bias
2. The model is too complex
3. The size of the training data

Examples:



Techniques to reduce overfitting:

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.

Closed Itemset

1(iv)

- An itemset is closed if none of its immediate supersets has the same support as the itemset

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

2) a) $s(B,D,E) = \frac{2}{10} = 0.2$ Ans.

(we took TID = 012, 022)

b) $(B,D) \rightarrow E$

confidence = $\frac{\text{support } (B,D,E)}{\text{support } (B,D)}$

$= \frac{0.2}{0.2} = 1$ Ans

$E \rightarrow B,D$

confidence = $\frac{\text{support } (B,D,E)}{\text{support } (E)}$

$= \frac{0.2}{0,8\,4} = 0.25$ Ans

c) Using CID:-

support (B, D, E) = $\frac{4}{5}$ = 0.8 Ans

[NOTE:- create itemsets by taking unions of same Cust ID and solve].

d) No, there is no relationship. Ans

3) Info (D) = I (3, 2) = $\frac{-3}{5} \log\left(\frac{3}{5}\right) - \frac{2}{5} \log\left(\frac{2}{5}\right)$

= 0.442 + 0.528

= 0.97

Body temp:-

I (cold-blooded) = Non-mamels → 2, Mammals → 0

I (cold-blooded) = 0

I (Warm-blooded) = Non-mam → 1, Mamm → 2

I = $\frac{-1}{3} \log\left(\frac{1}{3}\right) - \frac{2}{3} \log\left(\frac{2}{3}\right)$

= 0.528 + 0.389

I (War-bld) = 0.917

Date
Page No.

$$Info_{BT} = \frac{2}{5} \times 0 + \frac{3}{5} \times 0.917$$

$$= 0.5502$$

$$Gain(BT) = 0.97 - 0.5502$$
$$= 0.4198$$

Gives Birth :-

$I(No)$ = Non-Mamm = 2, Mamm = 2

$$I = \frac{-2}{4} \log\left(\frac{2}{4}\right) - \frac{2}{4} \log\left(\frac{2}{4}\right)$$

$$= 1$$

$I(Yes)$ = Non-Mamm = 1, Mamm = 0

$$\Rightarrow 0$$

$$Info(GB) = \frac{4}{5} \times 1 + \frac{1}{5} \times 0$$

$$= 0.8$$

$$Gain(GB) = 0.97 - 0.8 = 0.17$$

DELTA Notebook

Date
Page No.

2) Four legged :-

$I(\text{Yes}) = NM = 1, M = 1$

$= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$

$= 1$

$I(\text{No}) = NM = 2, M = 1$

$= -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)$

$= 0.389 + 0.528$

$= 0.917$

$\text{Info}(FL) = \frac{2}{5} \times 1 + \frac{3}{5} \times 0.917$

$= 0.4 + 0.5502$

$= 0.9502$

$\text{Gain}(FL) = 0.97 - 0.9502 = 0.0198$

Date
Page No.

⇒ Hibernates :-

I(Yes) ⇒ NM = 1, M = 2

$$= \frac{-1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)$$

$$= 0.528 + 0.389$$
$$= 0.917$$

I(No) ⇒ NM = 2, M = 0

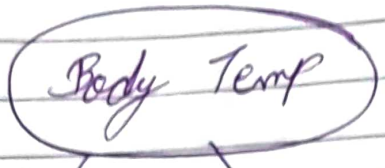
I ⇒ 0

$$\text{Info(Hib)} = \frac{3}{5} \times 0.917 + \frac{2}{5} \times 0$$

$$= 0.5502$$

$$\text{Gain(Hib)} = 0.97 - 0.5502 = 0.4198$$

Date
Page No.



Warm-blooded

Cold-blooded

GB	FL	Hiber	Class label
No	No	No	NM
No	No	Yes	M
No	Yes	Yes	M

Non-Mammal

$$\text{Info}(D) = I(1,2) = \frac{-1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\times\log\left(\frac{2}{3}\right)$$

$$= 0.528 + 0.389$$

$$= 0.917$$

GB:-

$$I(\text{No}) \Rightarrow NM = 1,\ M = 2$$

$$= \frac{-1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)$$

$$= 0.528 + 0.389$$

$$= 0.917$$

$$\text{Info}(GB) = \frac{3}{3}\times 0.917 = \underline{0.917}$$

$$\text{Gain}(GB) = 0.917 - 0.917 = 0.$$

DELTA Notebook

Date
Page No.

FL :-

I(No) ⇒ NM = 1, M = 1

⇒ 1

I(Yes) ⇒ NM = 0, M = 1

I = 0

$$\text{Info}(FL) = \frac{2}{3} \times 1 + \frac{1}{3} \times 0$$

$$= 0.667$$

Gain (FL) = 0.917 - 0.667

= 0.25

Hiber :-

I(No) ⇒ NM = 1, M = 0

I = 0

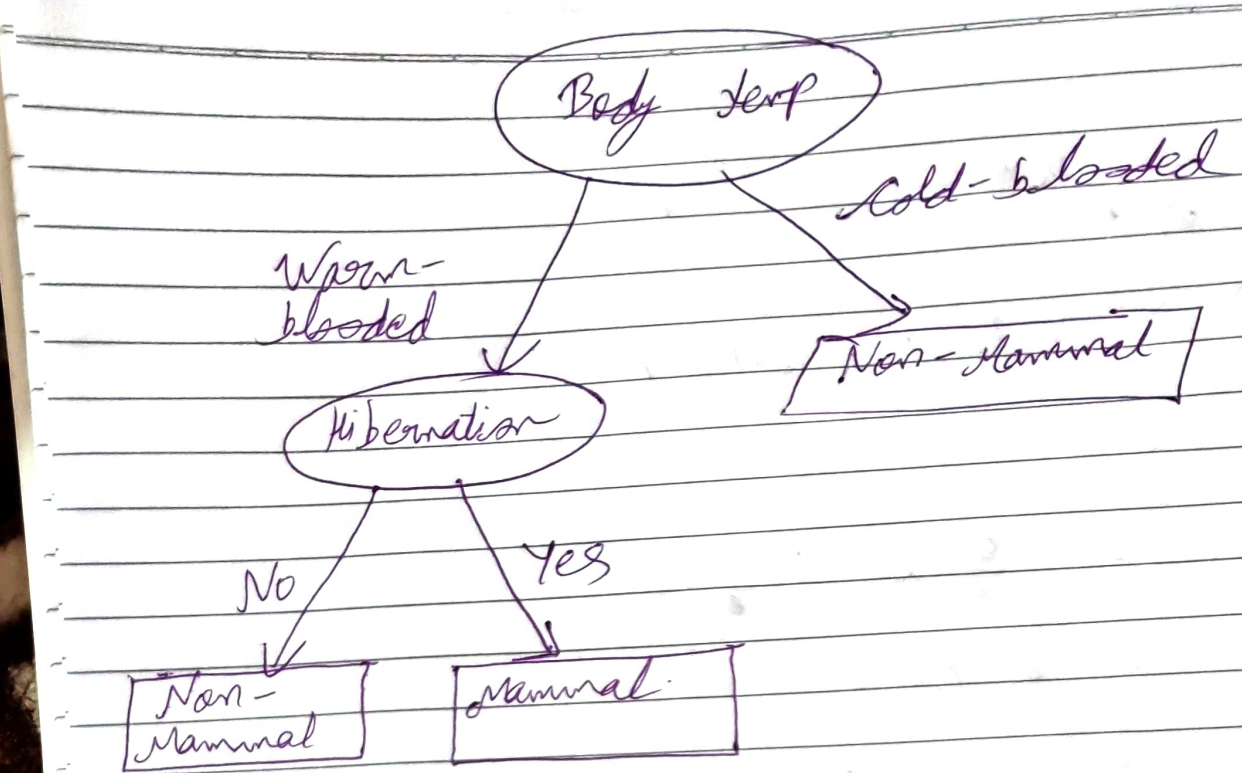
I(Yes) ⇒ NM = 0, M = 2

I = 0

Info (Hiber) = 0

Gain (Hiber) = 0.917

Date
Page No.


Body temp
Cold-blooded
Warm-blooded
Non-Mammal
Hibernation
No
Yes
Non-Mammal
Mammal