



Curse of Dimensionality:

→ ML excels at analyzing data with many dimensions, but it becomes challenging to create meaningful models as the number of dimensions increases.

→ curse of dimensionality

↳ increasing data dimensions and its explosive tendencies

→ increase computational efforts for analyze and process the data.

Curse of DIMENSIONALITY

As the dimensionality of the features space increases, the number of configurations can grow exponentially, and thus the number of configurations covered by an observation decreases.

Chris Albon

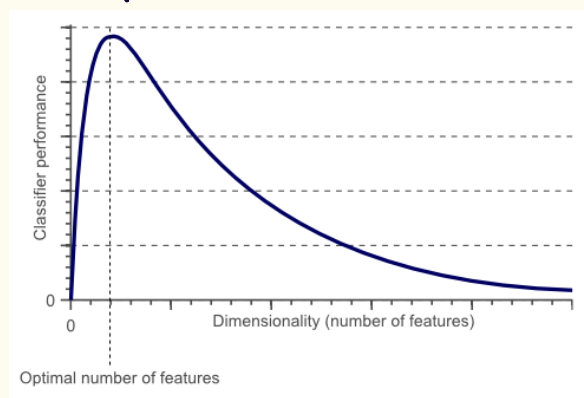
→ say, # of features = p
of data points = n

↳ COD $\Rightarrow p \gg n$



Hughes Phenomenon:

→ say that model's performance increases with the increasing number of features until we reach the optimal # of features.



→ distance = Euclidean distance

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

→ each new dimension adds a non-negative term to the sum.

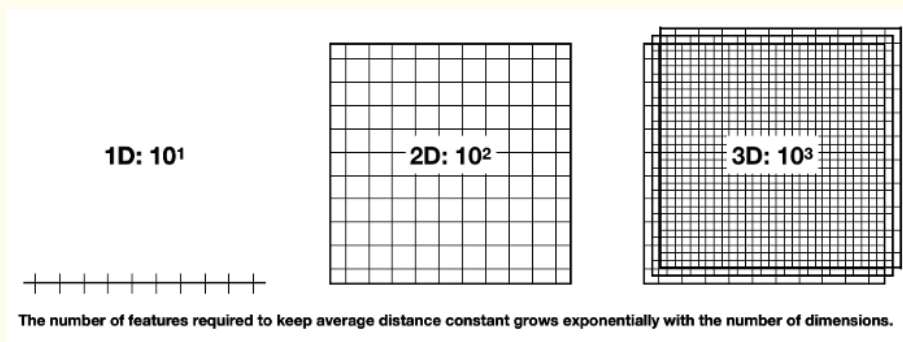
↳ distance $\uparrow \propto$ # dimension \uparrow
in feature vector

→ now with given datapoints -

as # of dimension $\uparrow \Rightarrow$ feature space becomes
less dense or emptier ← extremely sparse

→ In other words,

↳ lower data density requires more observations to keep avg distance b/w datapoints same.



Overfitting:

- variance increases as they get more opportunity to overfit to noise in more dimensions.
 - ↳ poor generalization
- KNN is very susceptible to overfitting due to CoD.
 - ↳ sparsity \uparrow with dimension \uparrow for fixed # of datapoints.
 - ↳ so the closest neighbour is also too far away in high dim space to give a good estimate.
- There is no scope of adding regularization in KNN so it suffers from CoD.
- as $p \gg n \Rightarrow \text{datapoints} \ll \text{features}$
 - ↳ overfit
 - ↳ fails in test data.



Underfit:

OCCAM'S RAZOR

When multiple hypotheses explain something equally well, we should prefer the simplest one.

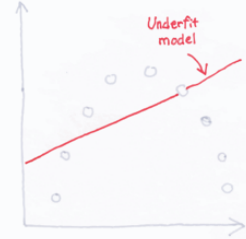
Note: This is the intuition, not the formal definition of Occam's razor.

Chris Albon

UNDERFITTING

A model is underfit when it fails to capture the pattern in the data. It suffers from high bias.

Chris Albon



→ to reduce chance of overfitting, we can prefer simple model with less parameters.

↳ But this may suffer from underfitting



solution:

→ Forward feature selection

→ Dimension Reduction

↳ PCA

↳ t-SNE

