Parshvanath Charitable Trust's

A. P. SHAH INSTITUTE OF TECHNOLOGY

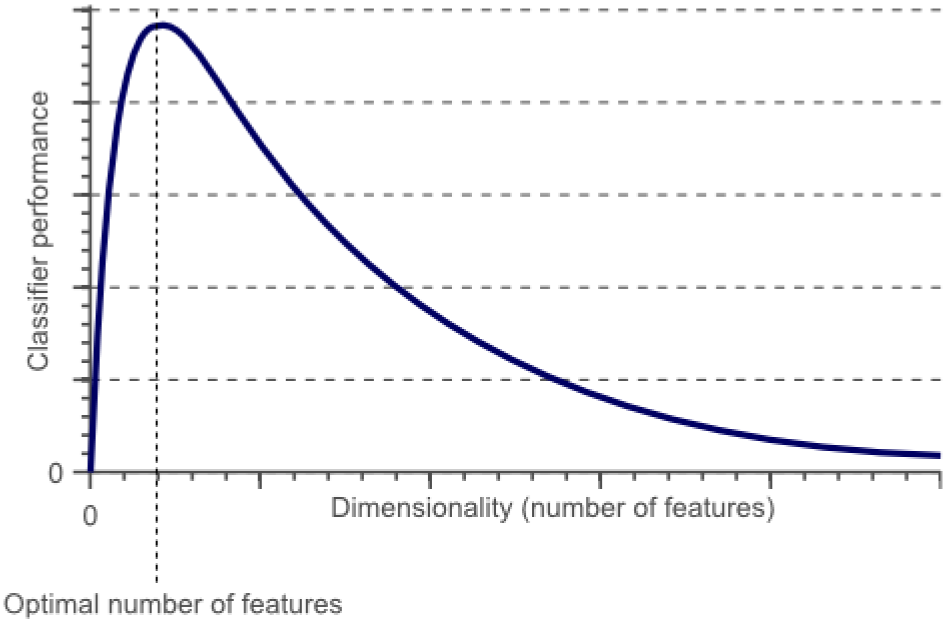
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)

Curse of Dimensionality

Regarding the curse of dimensionality — also known as the Hughes Phenomenon — there are two things to consider. On the one hand, ML excels at analyzing data with many dimensions. Humans are not good at finding patterns that may be spread out across so many dimensions, especially if those dimensions are interrelated in counter-intuitive ways. On the other hand, as we add more dimensions we also increase the processing power we need to analyze the data, and we also increase the amount of training data required to make meaningful data models.

The Hughes Phenomenon shows that as the number of features increases, the classifier's performance increases as well until we reach the optimal number of features. Adding more features based on the same size as the training set will then degrade the classifier's performance.



Let us understand this peculiarity with an example, suppose we are building several machine learning models to analyze the performance of a Formula One (F1) driver. Consider the following cases:

i) Model_1 consists of only two features say the circuit name and the country name.

ii) Model_2 consists of 4 features say weather and max speed of the car including the above two.

iii) Model_3 consists of 8 features say driver's experience, number of wins, car condition, and driver's physical fitness including all the above features.

iv) Model_4 consists of 16 features say driver's age, latitude, longitude, driver's height, hair color, car color, the car company, and driver's marital status including all the above features.

v) Model_5 consists of 32 features.

vi) Model_6 consists of 64 features.

vii) Model_7 consists of 128 features.

viii) Model_8 consists of 256 features.

ix) Model_9 consists of 512 features.

x) Model_10 consists of 1024 features.

Assuming the training data remains constant, it is observed that on increasing the number of features the accuracy tends to increase until a certain threshold value and after that, it starts to decrease. From the above example the accuracy of Model_1 < accuracy of Model_2 < accuracy of Model_3 but if we try to extrapolate this trend it doesn't hold true for all the models having more than 8 features. Now you might wonder if we are providing some extra information for the model to learn why is it so that the performance starts to degrade. This is nothing but **curse of dimensionality** is.

If we think logically some of the features provided to Model_4 don’t actually contribute anything towards analyzing the performance of the F1 driver. For example, the driver’s height, hair color, car color, car company, and the driver’s marital status is giving useless information for the model to learn, hence the model gets confused with all this extra information, and the accuracy starts to go down.

Curse of dimensionality in various domains:

There are several domains where we can see the effect of this phenomenon. Machine Learning is one such domain. Other domains include numerical analysis, sampling, combinatorics, data mining, and databases.

What problems does curse of dimensionality cause?

- **Data Sparsity:** Data points become increasingly spread out, making it hard to find patterns or relationships.
- **Computational Complexi**ty: The computational burden of algorithms increases exponentially.
- **Overfitting:** Models become more likely to memorize the training data without generalizing well.
- **Distortion of Distance Metrics**: Traditional distance metrics become less reliable in measuring proximity.
- **Visualization Challenge**s: Projecting high-dimensional data onto lower dimensions leads to loss of information.
- **Data Preprocessin**g: Identifying relevant features and reducing dimensionality is crucial for effective analysis.
- **Algorithmic Efficienc**y: Algorithms need to be scalable and efficient to handle the complexity of high-dimensional spaces.
- **Domain-Specific Challenges:** Each domain faces unique challenges in high-dimensional spaces, requiring tailored approaches.

- **Interpretability Issues**: Understanding the decision-making process of high-dimensional models becomes increasingly difficult.
- **Data Storage Requirement**s: Efficient data storage and retrieval strategies are essential for managing large volumes of high-dimensional data.

To mitigate the curse of dimensionality, there are two common approaches: feature selection and feature extraction.