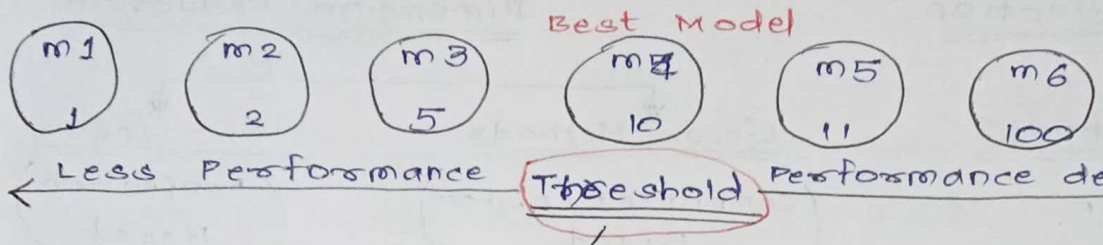
* Dimensionality Reduction *

Dimensionality :- Features / columns in dataset.

Curse of Dimensionality :-

↳ Generate models with increasing no. of features.



↳ Underfitting

↳ Less accurate

↳ May overfitting

↳ Increases complexity

↳ Decrease performance

↳ Feature not useful

↳ ↓ accuracy

↳ ↓ generalization performance

Threshold

Minimum num. of required features for the best Model.

Dimensionality Reduction :-

↳ As dimensionality increases, accuracy increases but till a certain threshold, then after that start decreasing performance.

↳ Reduce features in dataset, while retain most of important info. as possible.

↳ convert high-dim. data ⇒ lower dim. data, still preserve essence of original data.

↳ Reduce complexity & improve generalization perf.

↳ Done during pre-processing, before building model to improve performance.

↳ Sometimes can discard useful info as well, so have to be careful.

Dimensionality Reduction Techniques

(Find subset of original set of variables)

Feature Selection

- Filters Methods:
  - Correlation
  - Chi-Square Test
  - ANOVA
  - Information Gain
- Wrapper Methods:
  - Forward Selection
  - Backward Selection
  - Bi-directional Elimination
- Embedded Methods:
  - LASSO
  - Elastic Net
  - Ridge Regression

Dimension Reduction

Linear Methods (Component / Factor Based)

- PCA
- ICA
- LDA
- Factor Analysis

Non-Linear (Projection Based)

- ISOMAP
- t-SNE
- UMAP
- Auto-encoders
- Kernel PCA

- Filters Methods: Filters dataset & keep relevant features.
- Wrapper Methods: Use model to evaluate better performance features. More accurate than filtering, complex.
- Embedded Methods: check training iterations & evaluate each feature importance.

PCA (Principal Component Analysis)

LDA (Linear Discriminant Analysis)

ICA (Independent Component Analysis)

ISOMAP (Isometric Mapping)

t-SNE (t-distributed stochastic Neighbor Embedding)

UMAP (Uniform Manifold Approximation & Mapping)

2 components of Dimensionality Reduction:-

① Feature selection: Find subset of original set of variables/ features to use for training.

- ↳ Filter
- ↳ Wrapper
- ↳ Embedded

② Feature Extraction: Reduce data in high-dim space to low-dim space, with lesser no. of dimensions.

- Advantages of Dimensionality Reduction:

① Data compression ⇒ Reduce storage space.
② Reduce computation time.
③ Remove any redundant feature.
④ Easy visualize low-dim. data.
⑤ Reduce overfitting, Reduce complexity.
⑥ Improve model performance.
⑦ Reduce noise & irrelevant info.

- Limitations of Dimensionality Reduction:

① May loss som important data.
② Find linear correlations, sometimes undesirable.
③ May difficult understand relationship in original & reduced dimensions.
④ Sometimes may lead to overfitting.
⑤ Can be computational expensive, especially large datasets.