Advanced ML Topics

Trainer

Hi there! I am Agil Haykal, just call me **Agil**. I am a curious guy who end up involved in data technology.

I have experienced Data Science as a trainer, consultant, and developer. I have taught +300 Data Scientist, Engineer, and Business Intelligence in total.

I handled several industry from manufacturing, banking, telecommunication, government, and Insurance. Please feel free to contact me to discuss anything about data technology.

LinkedIn: Agil Haykal

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

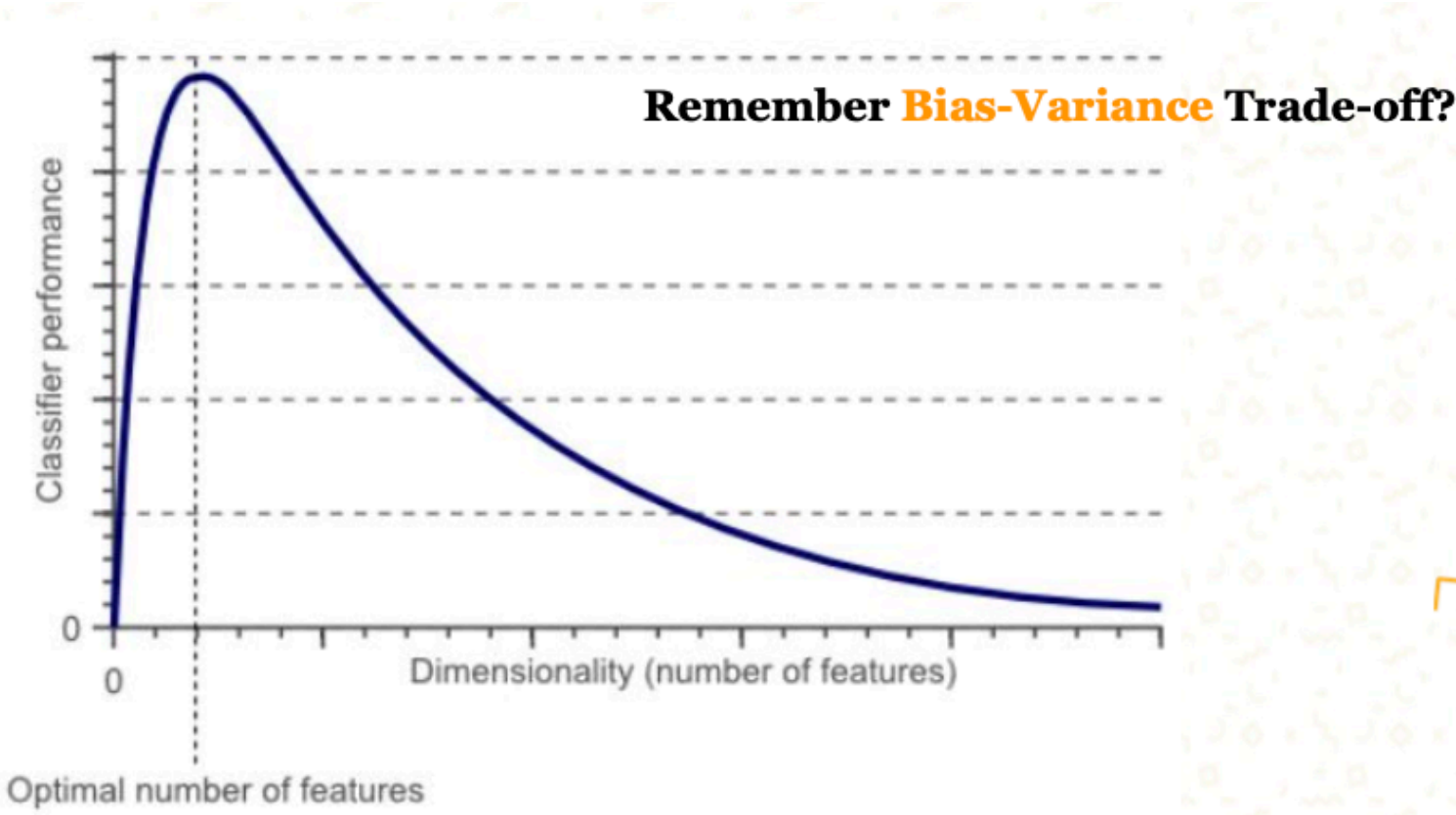
1. Feature Selection
2. Feature Importance
3. Feature Extraction
4. PCA

DigitalSkola
Uncover The World Of Digital Skills With Us

Dimensionality Curse

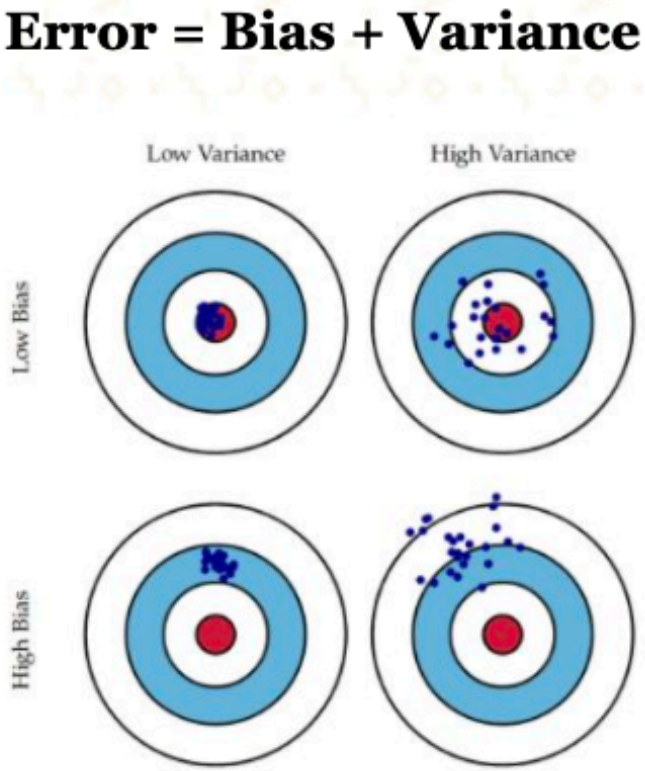
The process of reducing the dimension of your feature set / dataset

The Curse of Dimensionality Reduction



Bias Variance Trade-off

Bias & variance like throwing darts



Benefit of Dimensionality Reduction

1. Less misleading data means model accuracy improves.
2. Less dimensions mean less computing. Less means that algorithm train faster.
3. Less data means less storage space required.
4. Removes redundant features and noise.

Dimensionality Reduction

Provide plenty of variables or features for the model is a good practice to do. However, data usually are irrelevant and need to remove. There are two methods to reduce features:

1. Feature Selection
2. Feature Extraction

Feature Selection

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.

Feature Selection Methods

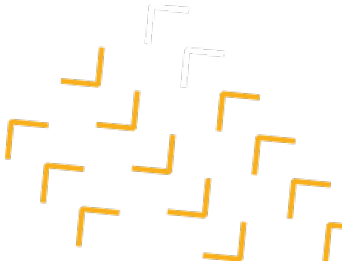
1. Univariate Selection
2. Correlation Matrix with
    1. Pearson Correlation
    2. Spearman Correlation
    3. Chi-square
3. Feature Importance

Hey, we have learned the first two methods! Let's learn the 3rd one…

Feature Importance

When we have created a ML model, it learned from the data and decide which features work best for the prediction. Not all features are used by the model, so they have different degree of importance.

Each model has their built-in feature importance calculation and we can use it.



Benefit of Feature Importance

Feature Importance has many benefit for us, those are:

1. Feature Selection
2. Model Interpretability (capable to explain why model predict into something)
3. To understand which feature has bias to the model

Permutation Importance

Permutation feature importance is a model inspection technique that can be used for any fitted estimator when the data is tabular.

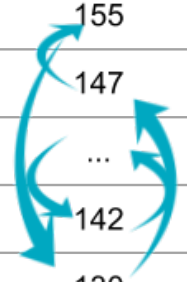
This is especially useful for non-linear or opaque estimators.

The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled.

How it works

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24



Weight	Feature
0.1750 ± 0.0848	Goal Scored
0.0500 ± 0.0637	Distance Covered (Kms)
0.0437 ± 0.0637	Yellow Card
0.0187 ± 0.0500	Off-Target
0.0187 ± 0.0637	Free Kicks
0.0187 ± 0.0637	Fouls Committed
0.0125 ± 0.0637	Pass Accuracy %
0.0125 ± 0.0306	Blocked
0.0063 ± 0.0612	Saves
0.0063 ± 0.0250	Ball Possession %
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0.0000 ± 0.0559	On-Target
-0.0063 ± 0.0729	Offsides
-0.0063 ± 0.0919	Corners
-0.0063 ± 0.0250	Goals in PSO
-0.0187 ± 0.0306	Attempts
-0.0500 ± 0.0637	Passes

Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing.

One of the algorithm is called PCA.

Principle Component Analysis (PCA)

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset.

It's often used to make data easy to explore and visualize.

Goal

1. Reduce the d-dimensions of the dataset by projecting onto k-dimensions subspace (where k < d)
2. Identify Pattern of Data
3. Detect the correlation between features

DigitalSkola
Uncover The World Of Digital Skills With Us

How does it works?

https://setosa.io/ev/principal-component-analysis/

Benefit PCA

1. Machine Learning Algorithm will train faster because of less dimension
2. Potentially improve Machine Learning performance
3. Less memory usage when training
4. Easy to visualize
5. Remove noise

When to Use

1. Model Interpretability is not the concern of the model.
2. Data have plenty of columns.
3. To Visualize multivariable dataset.

Do not Use PCA when...

1. Need to know feature importance.
2. The each variable has little correlation (Pearson < 0.3). If use PCA for this data, then the number of column would be the same.
3. Classification project has extreme imbalanced target. The features of minority target might potentially removed by the algorithm.

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU