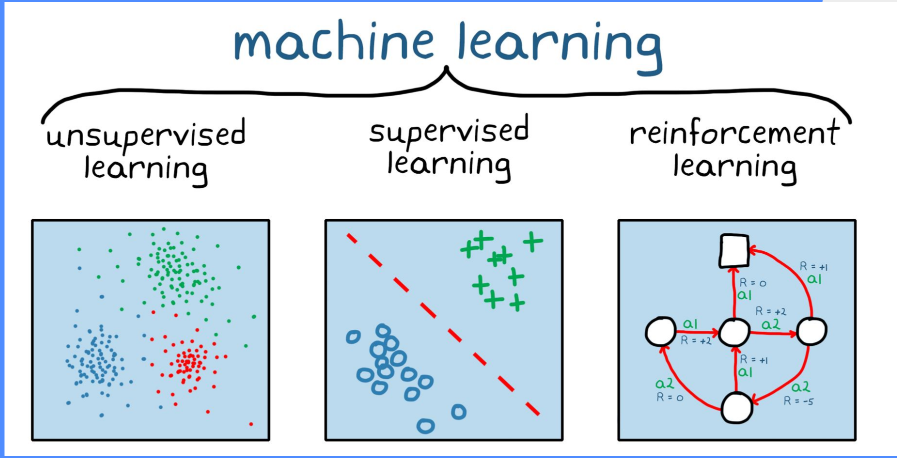
Getting Started with Python and Machine Learning

2021.02.14
Data eXperience Lab, Winter Seminar
2018312824 Ryu Chaeeun

OUTLINE

1. Categories of Machine Learning
2. Generalizing with data
3. Overfitting, underfitting and the bias-variance tradeoff
4. Techniques to avoid overfitting



Categories of Machine Learning

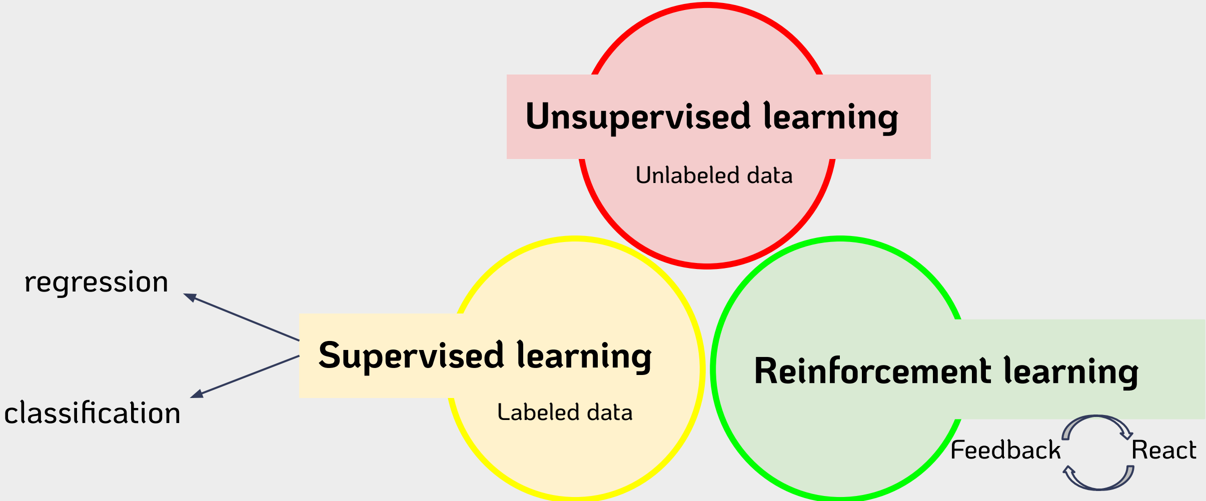
Machine Learning

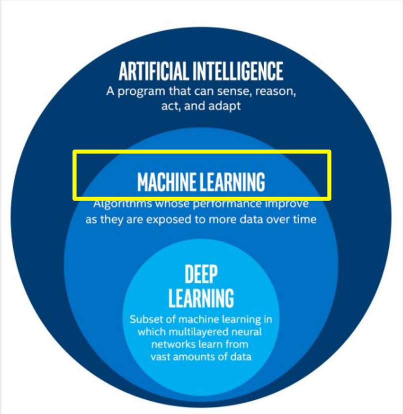
Definition

the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

Task

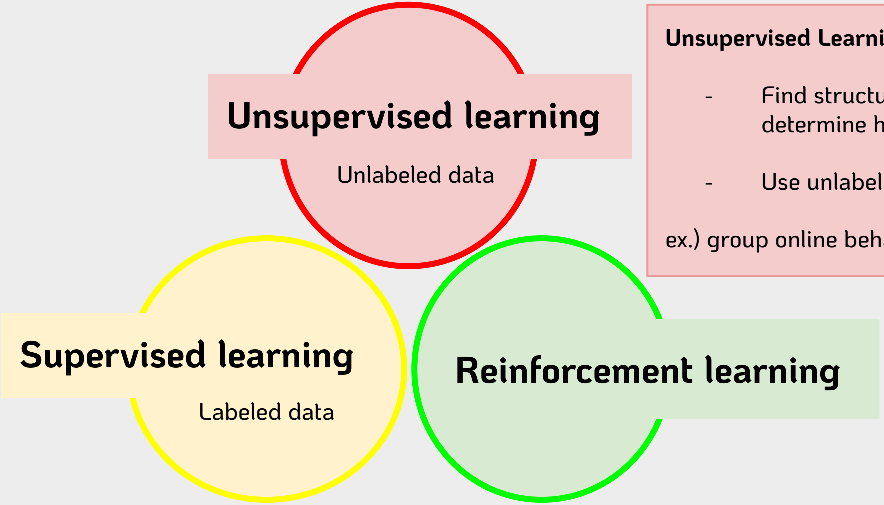
To explore and construct algorithms that can learn from historical data and make predictions on new input data.





Machine Learning

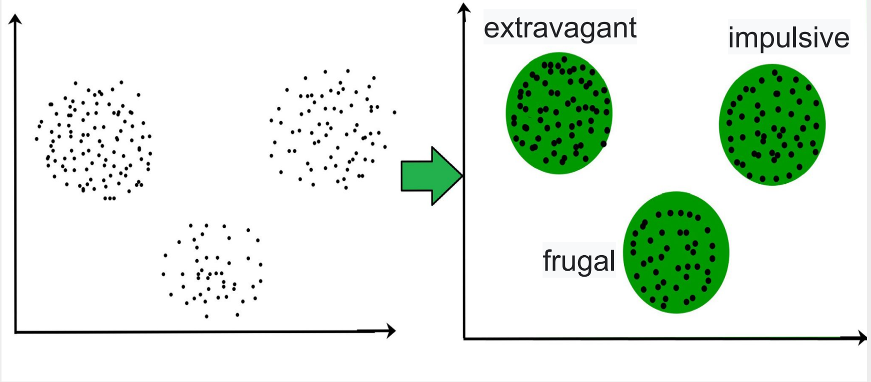
Categories of Machine Learning



Unsupervised Learning

- Find structure of the data underneath, to discover hidden information, or to determine how to describe the data.
- Use unlabeled data

ex.) group online behaviors of customers



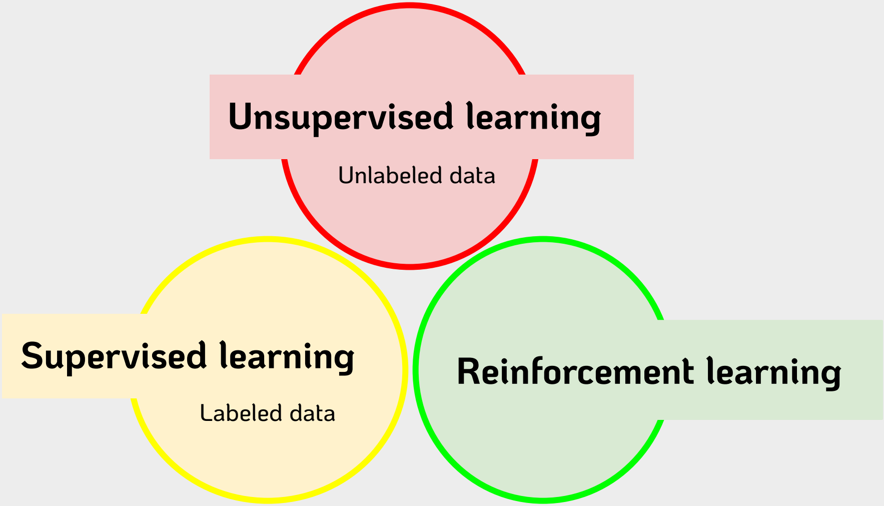
Machine Learning

Categories of Machine Learning

Supervised Learning

- Find a map that maps inputs to outputs.
- Use labeled data.

ex.)



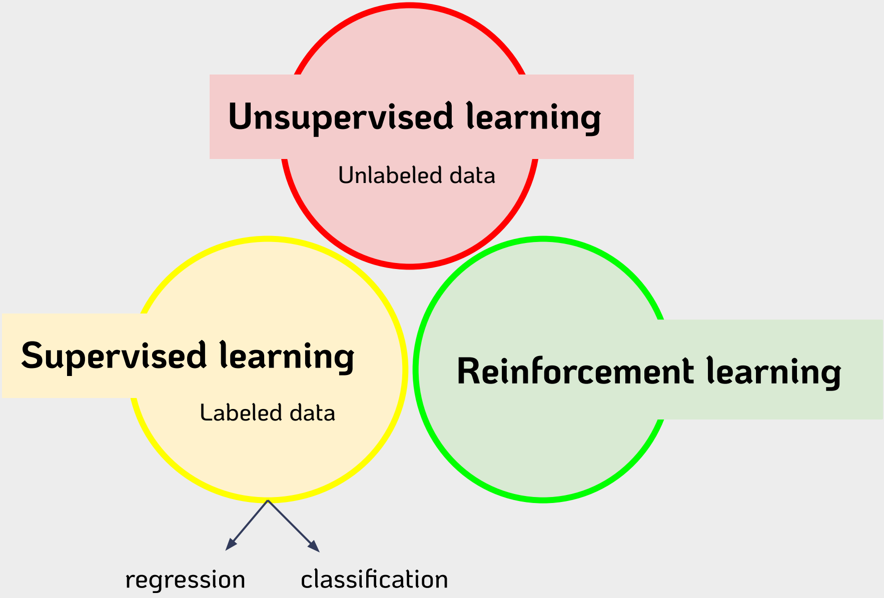
Machine Learning

Categories of Machine Learning

2 Categories of Supervised Learning

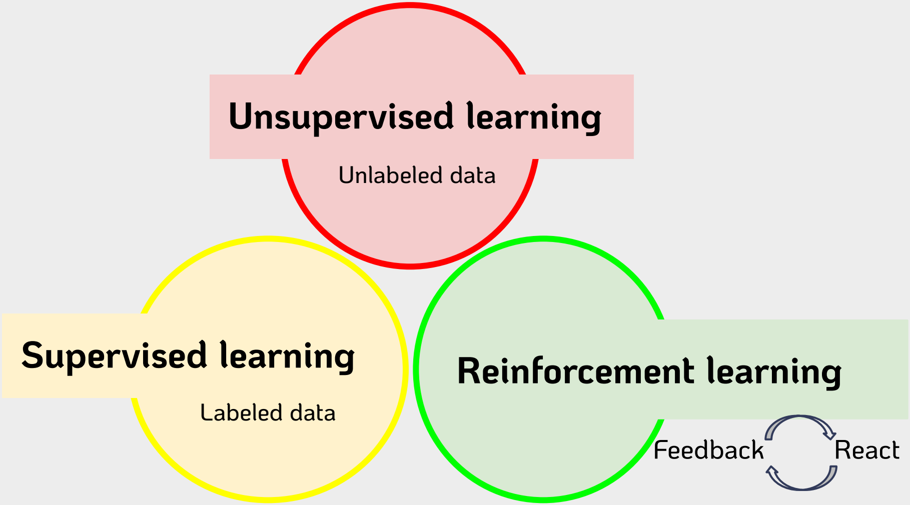
1. Regression
   Trains on & predicts a continuous-valued response.
   ex.) house prices
2. Classification
   Attempts to find the appropriate class label
   ex.) sentiment labeling (positive/negative)

Semi-supervised learning
Makes use of unlabeled data(usually a large amount) for training, besides a small amount of labeled.



Machine Learning

Categories of Machine Learning



Reinforcement learning

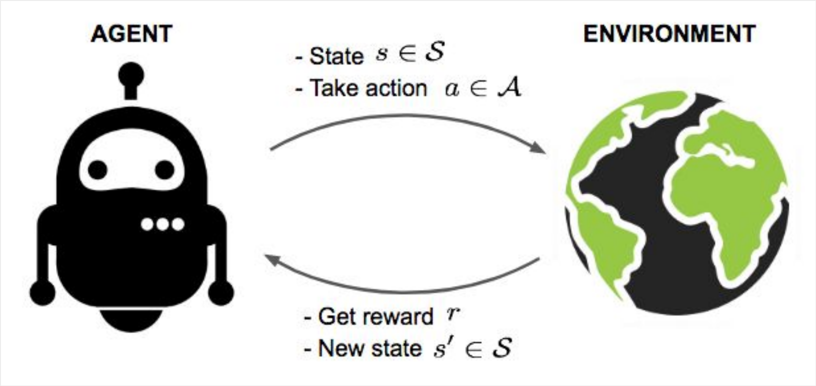
Learning data provides feedback to achieve a certain goal

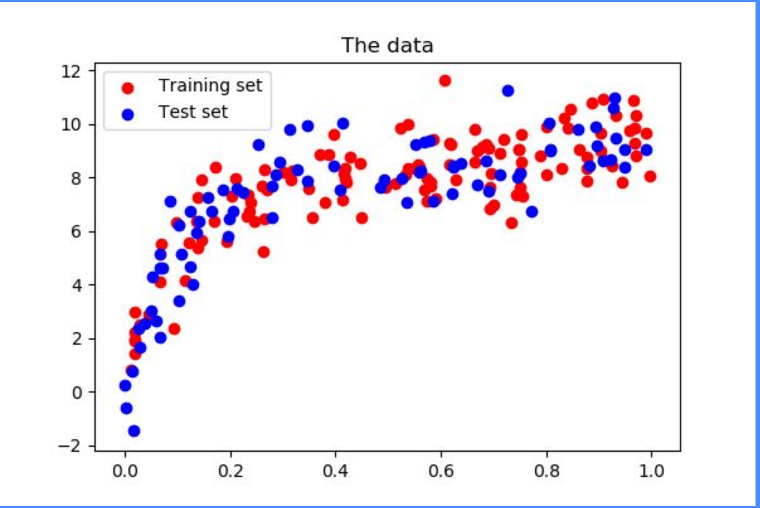
↓

Evaluates performance based on feedback

↓

Reacts according to the evaluation



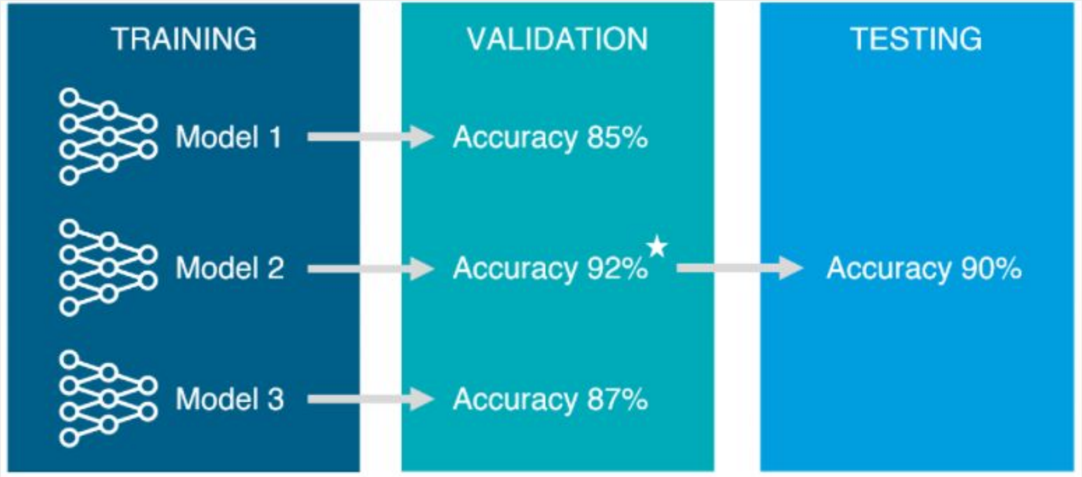


Generalizing with data

Generalizing with data

Indexing

- **Training set(samples)**: where models derive patterns from.
- **Validation set(samples)**: verify how well models will perform in a simulated setting
- **Test set(samples)**: where the models are eventually applied



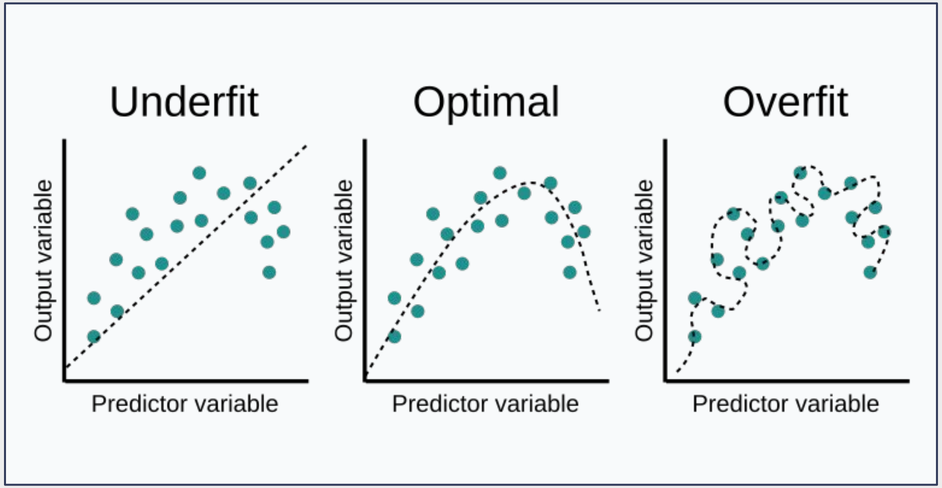
Overfitting & Underfitting

Overfitting

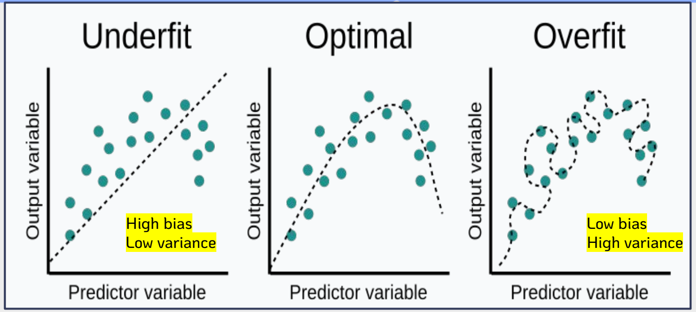
- Extracting too much information from the training set and making our model just work well with them.
  (-> occurs when the model or the algorithm fits the data too well.)
- Low bias, High variance

Underfitting

- Does not perform well on training sets
- High biase, Low variance



Bias-Variance tradeoff

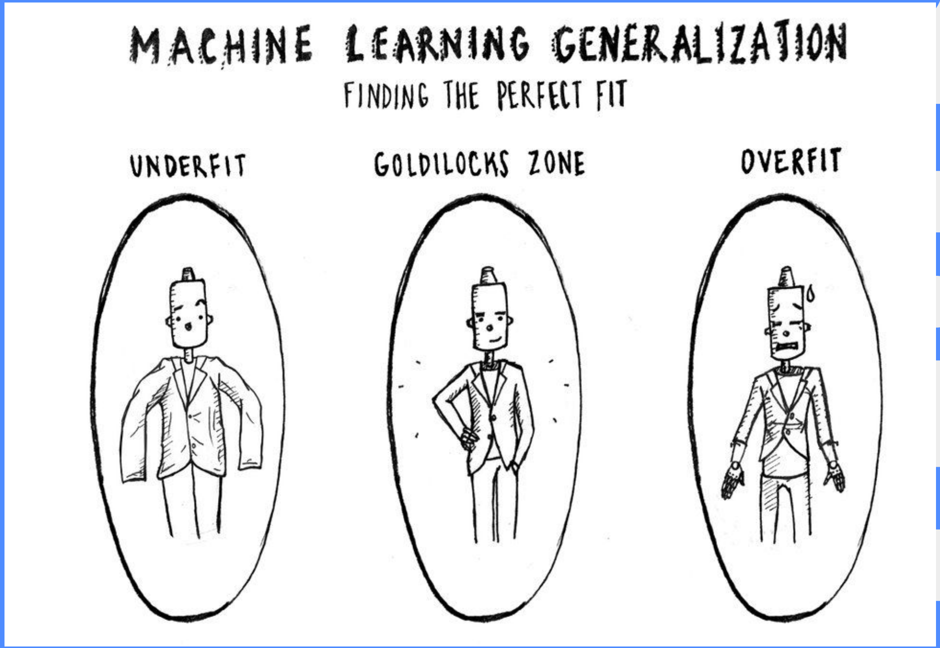


- **Bias**: error stemming from incorrect assumptions in the learning algorithm. (Underfit: High, Overfit: Low)
- **Variance**: how sensitive the model prediction is to variations in the datasets. (Underfit: Low, Overfit: High)

Best if **both** bias & variance -> **low**

But,
Bias-Variance trade-off: Decreasing one increases the other.

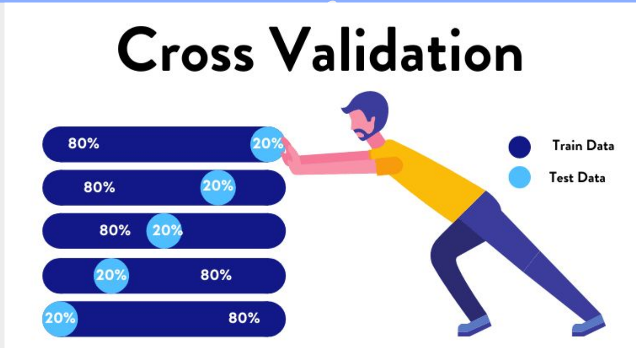
So,
Employ **cross-validation technique** to find the optimal model balancing bias and variance and to diminish overfitting.



To Avoid Overfitting

Avoid overfitting with cross-validation

- **Cross-validation** Process:
  - 1. Divide data into two subsets: Train dataset & Test dataset
  - 2. Perform multiple rounds of cross-validation, under different partitions.
  - 3. Average the testing results from all rounds.



Two Types of Cross-validation Schemes

1. **Exhaustive**: leave out a fixed number of observations in each round as testing(or validation) samples, the remaining observations as training samples. Repeat this process **until all possible different subsets of samples are used for testing once.**
   ex.) Leave-one-out-cross-validation (LOOCV)

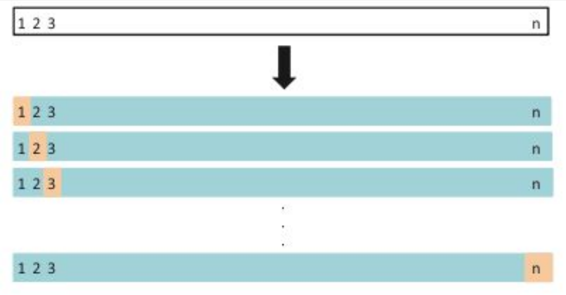
2. **Non-exhaustive**: does **not** try out **all** possible partitions.
   ex.) k-fold cross-validation

Avoid overfitting with cross-validation

Two Types of Cross-validation Schemes

1. **Exhaustive**: leave out a fixed number of observations in each round as testing(or validation) samples, the remaining observations as training samples. Repeat this process **until all possible different subsets of samples are used for testing once.**
   ex.)
   Leave-one-out-cross-validation (LOOCV):
   for a dataset of size n, LOOCV requires n rounds of cross-validation.
   (slow when n gets large)

LOOCV

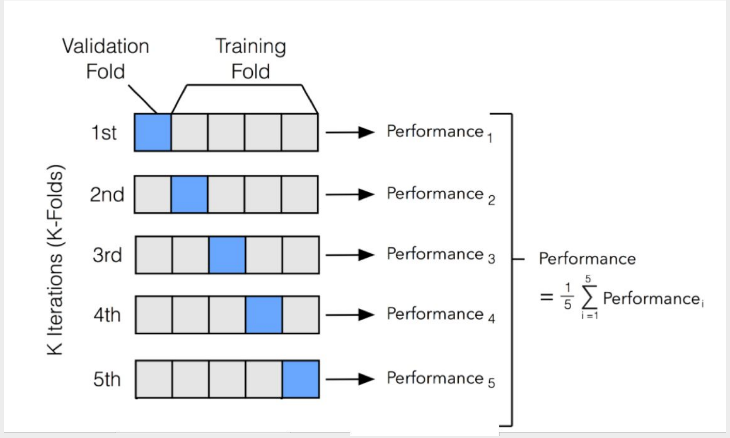


Avoid overfitting with cross-validation

2. **Non-exhaustive**: does **not** try out **all** possible partitions.

K-fold Cross-Validation
: <u>randomly</u> splits data into k equal-sized folds. In each trail, one of these folds becomes the testing set, and the rest becomes the training set. Repeat this process k times with each fold being the designated testing set once. Finally, average the k sets of test results for evaluation.

K-fold cross validation

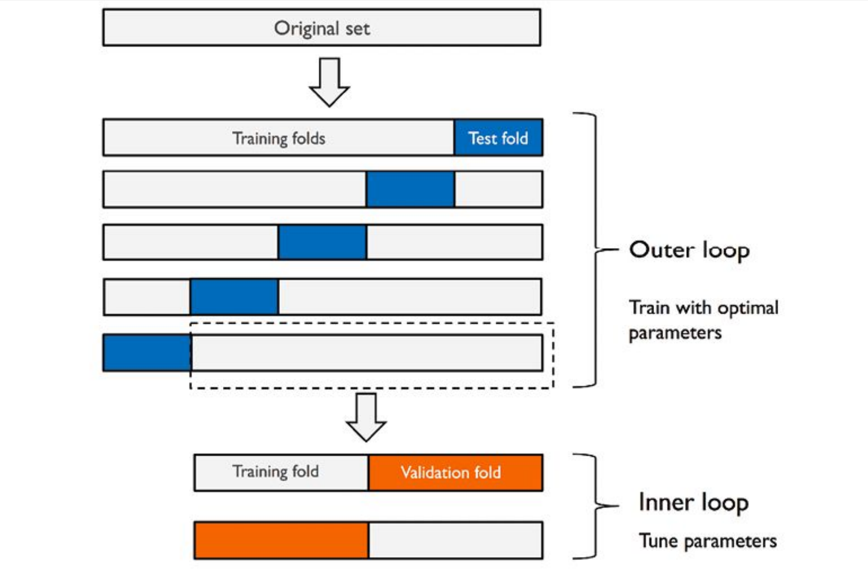


Avoid overfitting with cross-validation

Combination of Cross-validations

Nested Cross-validation: used for estimating the **generalization error of the model** along with the search of **most optimal combination of hyper parameters value**.

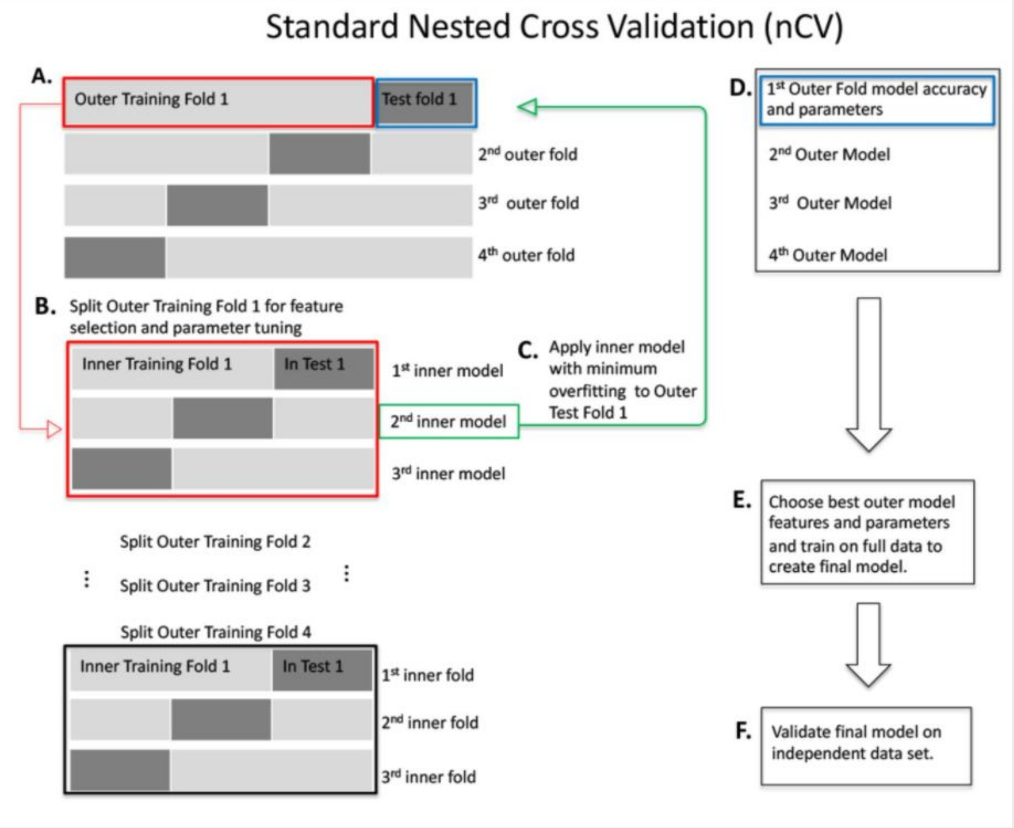
- Inner cross-validation: conducted to find the best fit, and can be implemented as a k-fold cross validation
- Outer cross-validation: used for performance evaluation and statistical analysis



Avoid overfitting with cross-validation

Combination of Cross-validations

Nested Cross-validation: used for estimating the **generalization error of the model** along with the search of **most optimal combination of hyper parameters value**.



Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

Intuitive process

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

X

target

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

target

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Predict stranger or not

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

target

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Predict stranger or not

Stranger Conditions:

- Female and Middle-aged and Average height and without glasses and dressed in black
- Male and Senior-aged and short height and with glasses and dressed in black

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

target

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Predict stranger or not

Stranger Conditions:

- Female and Middle-aged and Average height and without glasses and dressed in black
- Male and Senior-aged and short height and with glasses and dressed in black

——————————————Complex——————————————

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

target

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Predict stranger or not

Regularized stranger condition:

- Without glasses and dressed in black

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

target

Male	Young	Tall	With glasses	In grey	Friend
Female	Middle	Average	Without glasses	In black	Stranger
Male	Young	Short	With glasses	In white	Friend
Male	Senior	Short	Without glasses	In black	Stranger
Female	Young	Average	With glasses	In white	Friend
Male	Young	Short	Without glasses	In red	Friend

Predict stranger or not

Regularized stranger condition:

- Without glasses and dressed in black

Avoid overfitting with regularization

Regularization: prevents unnecessary complexity of the model -> avoid overfitting.

Early stopping:
Limit the time a model spends in learning to penalize complexity of the model