Model Selection & Training

Problem: Multivariate Regression

Preprocessing

Evaluation Model with training Data set

Model

Linear Regression

RMSE

Under fitting

Decision Tree Regression

RMSE

Over fitting

Linear Regression Goals & Evaluation Method

MSE $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

RMSE $\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$

RMSLE $\log\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$

MAE $\frac{\sum_{i=1}^{n}|Y_i-X_i|}{n}$

MAMPE $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-X_i}{Y_i}\right|$

if data set have a lot of noise value

Which method is fit for that data set?

Linear Regression Goals & Evaluation Method

MSE $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

RMSE $\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$

RMSLE $\log\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$

MAE $\frac{\sum_{i=1}^{n}|Y_i-X_i|}{n}$

MAMPE $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-X_i}{Y_i}\right|$

if data set have a lot of noise value

Which method is fit for that data set?

MAE

Linear Regression Goals & Evaluation Method

MSE $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

RMSE $\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$

RMSLE $\log\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$

MAE $\frac{\sum_{i=1}^{n}|Y_i-X_i|}{n}$

MAMPE $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-X_i}{Y_i}\right|$

if data set have a lot of noise value

Which method is fit for that data set?

MAE

WHY?

$(X+Y)^2$ vs $|X+Y|$

What's difference?

Evaluation Model with training Data set

Model

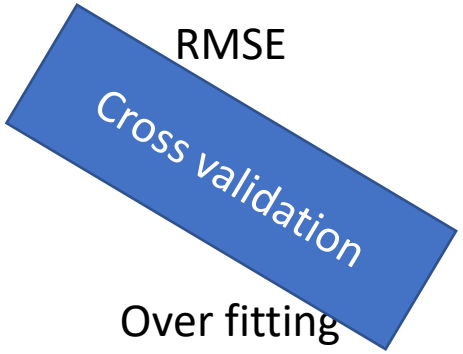
Linear Regression

RMSE

Under fitting

Decision Tree Regression

RMSE



Over fitting

Cross validation

Q. What is Goal of machine learning & statistical inference model?

Cross validation

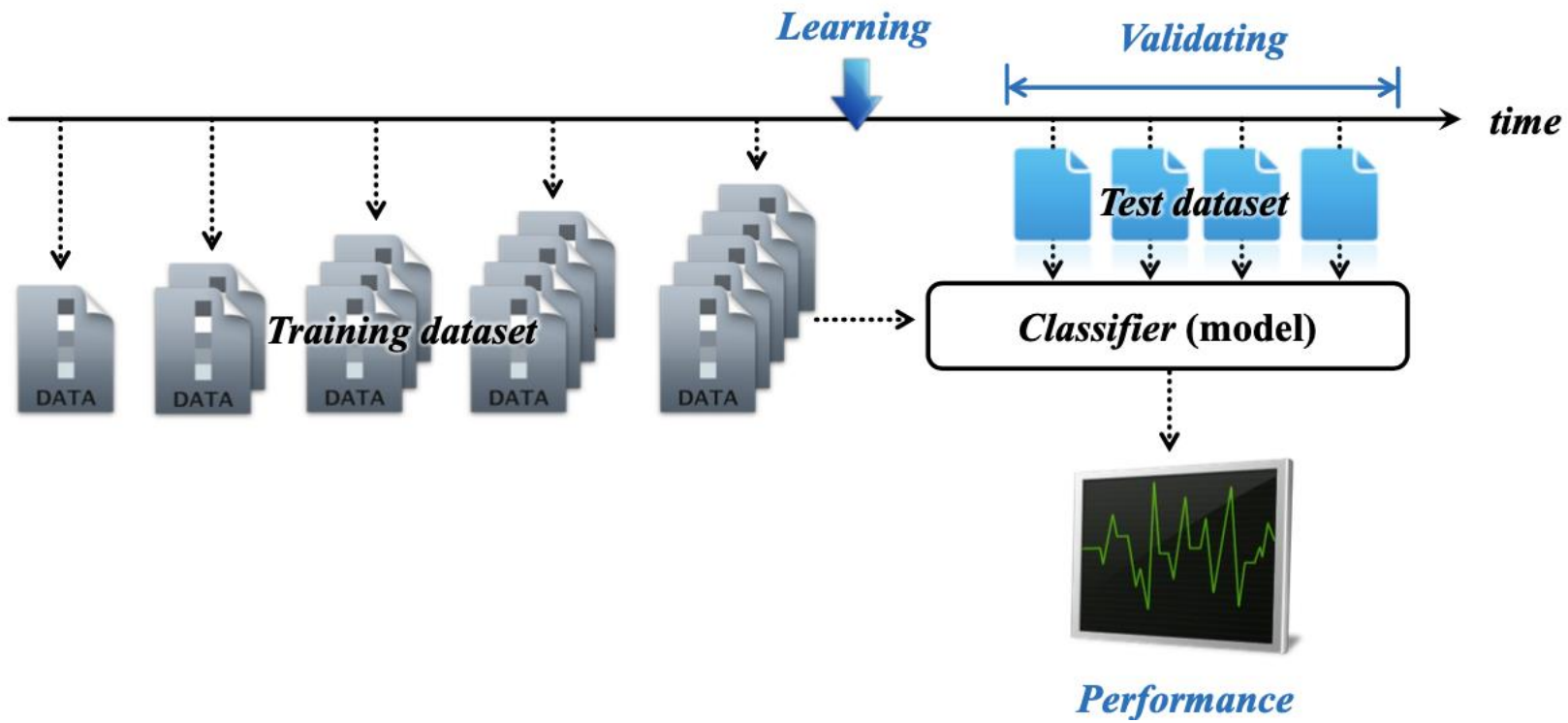
Q. What is Goal of machine learning & statistical inference model?

A. Predict the future by extracting pattern from our data(present)

Cross validation

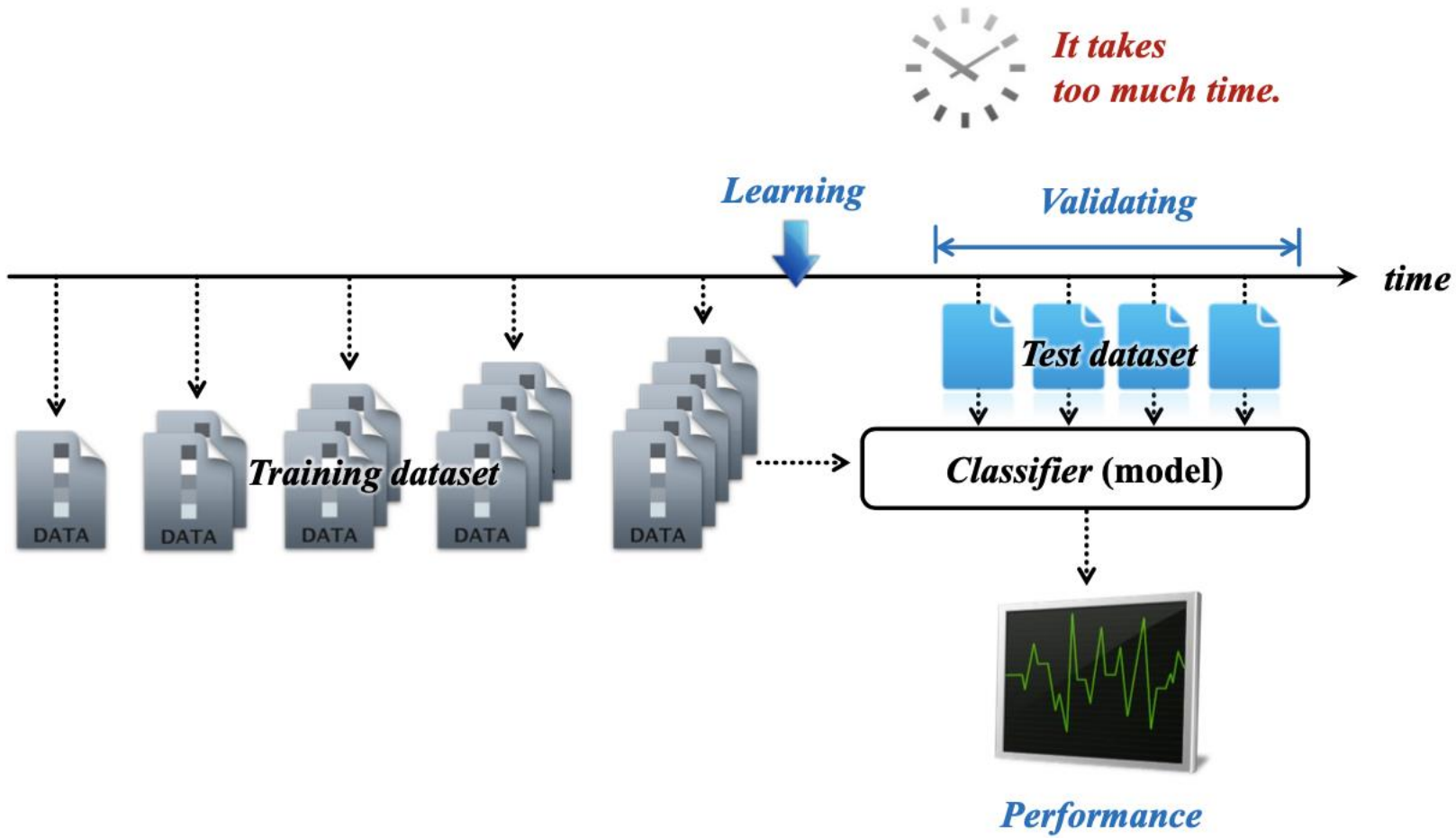
Q. What is Goal of machine learning & statistical inference model?

A. Predict the future by extracting pattern from our data(present)



Cross validation

Q. What is Goal of machine learning & statistical inference model?

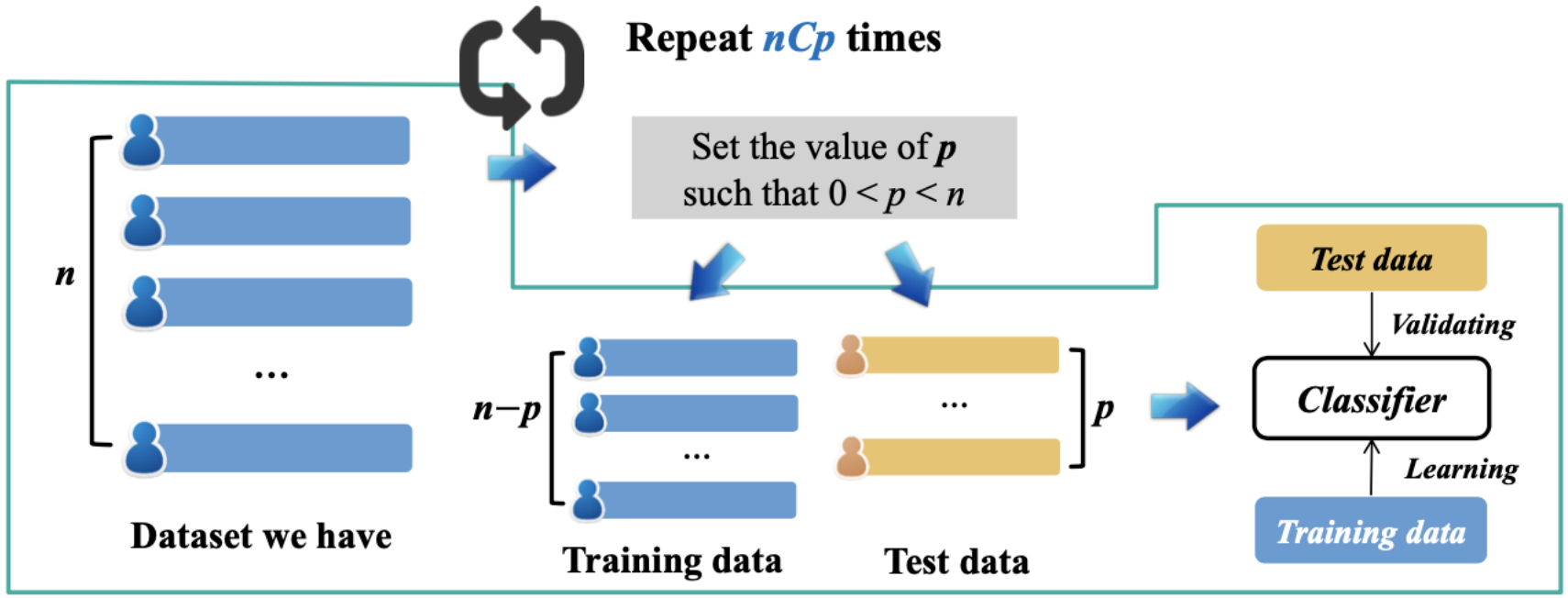


Cross validation

Leave-p-out

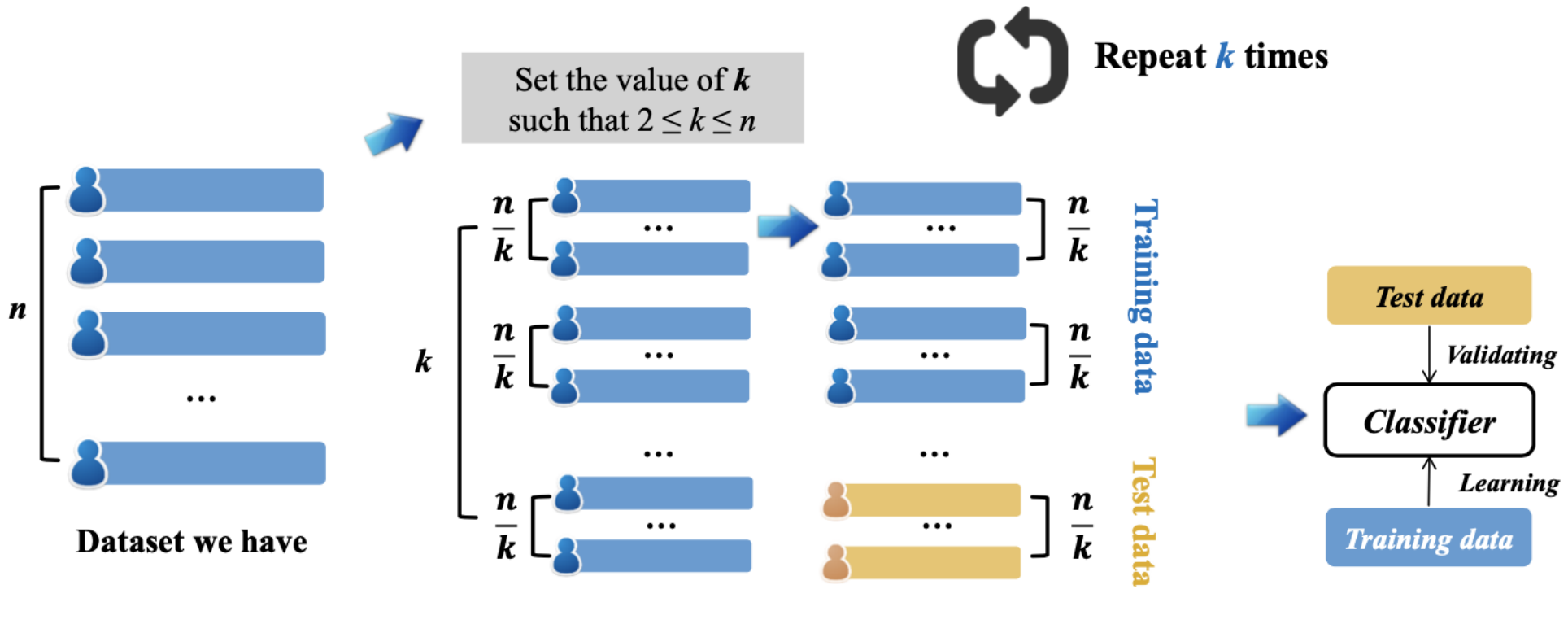
K-fold

Leave – p - out



100 1

K-fold


Set the value of k such that $2 \leq k \leq n$
Repeat k times
n
Dataset we have
k
$\frac{n}{k}$
Training data
Test data
Test data
Validating
Classifier
Learning
Training data

Fine Tuning

Model에게는 Hyper parameters 존재

Decision Tree hyperparameter

Parameters:

criterion : *{"gini", "entropy"}, default="gini"*
The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.

splitter : *{"best", "random"}, default="best"*
The strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.

max_depth : *int, default=None*
The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_split : *int or float, default=2*
The minimum number of samples required to split an internal node:

- If int, then consider `min_samples_split` as the minimum number.
- If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.

Changed in version 0.18: Added float values for fractions.

min_samples_leaf : *int or float, default=1*
The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least `min_samples_leaf` training samples in

1000

1,1,1,1
1,1,1,2
1,1,1,3
Max,max,max,max

2,3,5,1
3,5,6,1
.
.
.