Python을 활용한 데이터 분석 강의

Model Evaluation & Improvement

HYUNDAI MOBIS

Generalization

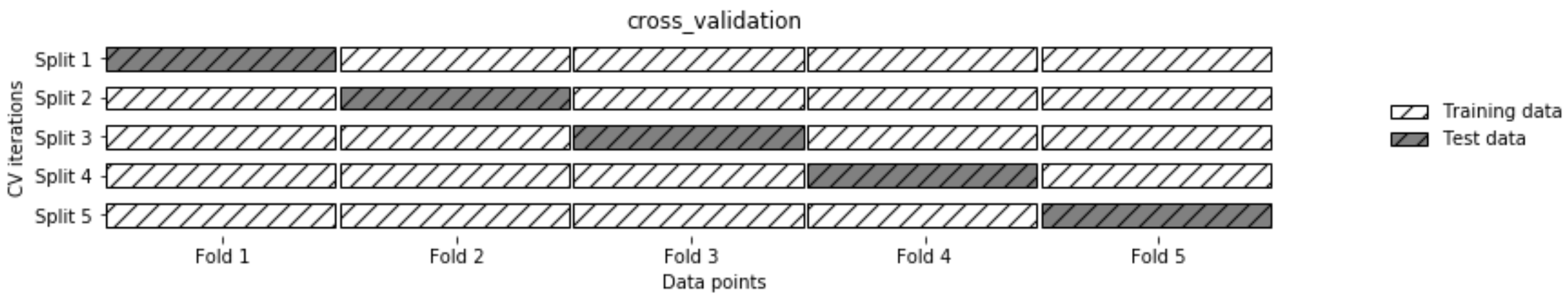
- training set에 얼마나 잘 fit 되는지는 중요하지 않음
- 새로운 데이터에서의 generalization이 중요!

- Cross-Validation
- Grid Search

HYUNDAI MOBIS

Cross-Validation

- generalization의 성능을 평가하는 방법
- k-fold cross-validation
  - (default) k = 3
  - 각 fold마다의 accuracy를 계산(test set)
  - cross_val_score

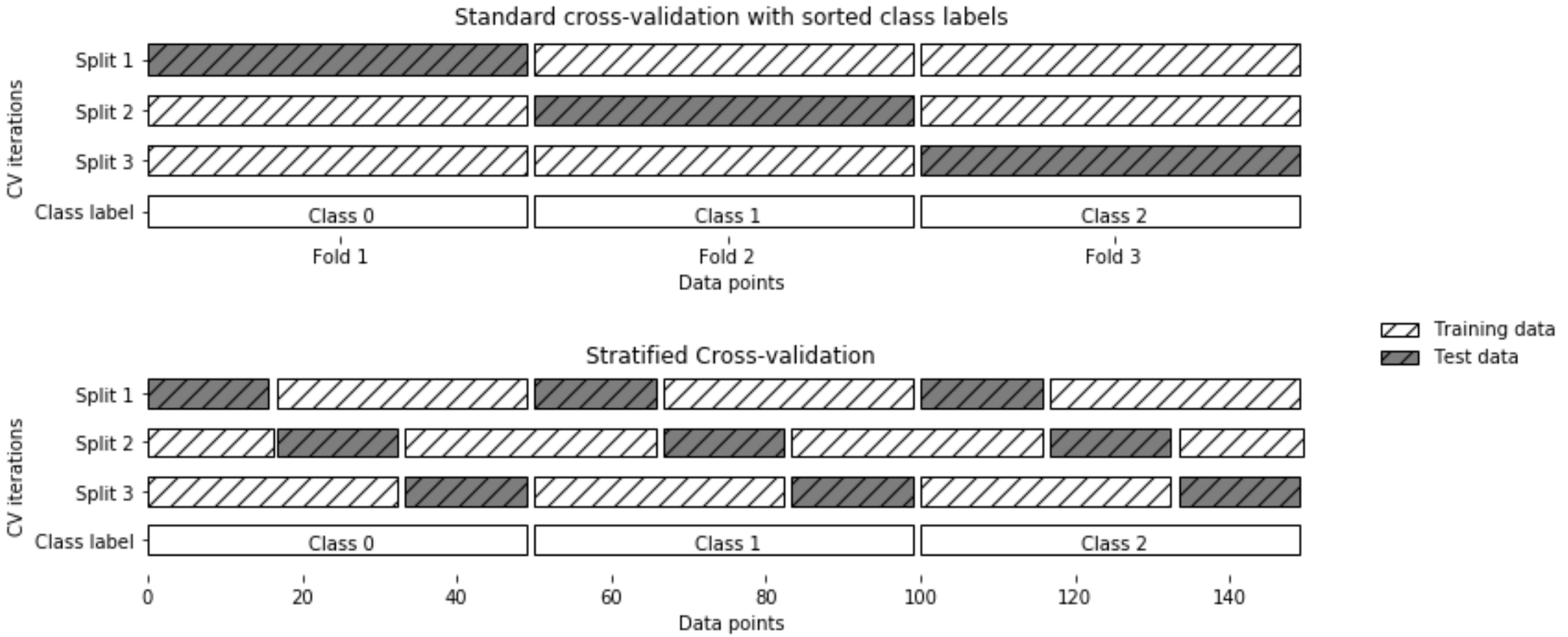


HYUNDAI MOBIS

Stratified k-Fold Cross-Validation

```
print("Iris labels:\n{}".format(iris.target))
```

```
Iris labels:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
```



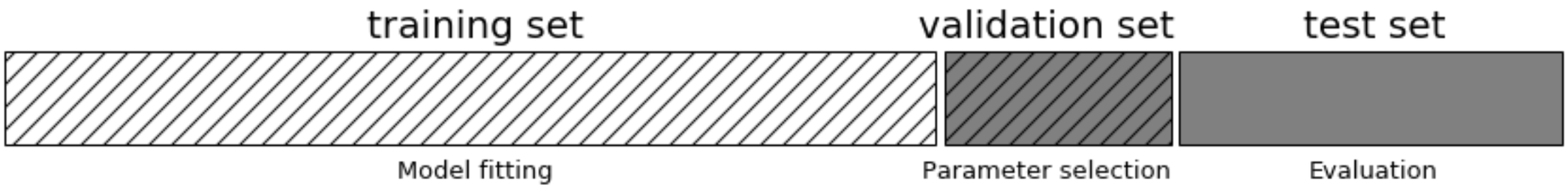
HYUNDAI MOBIS

Grid Search

- tuning parameters

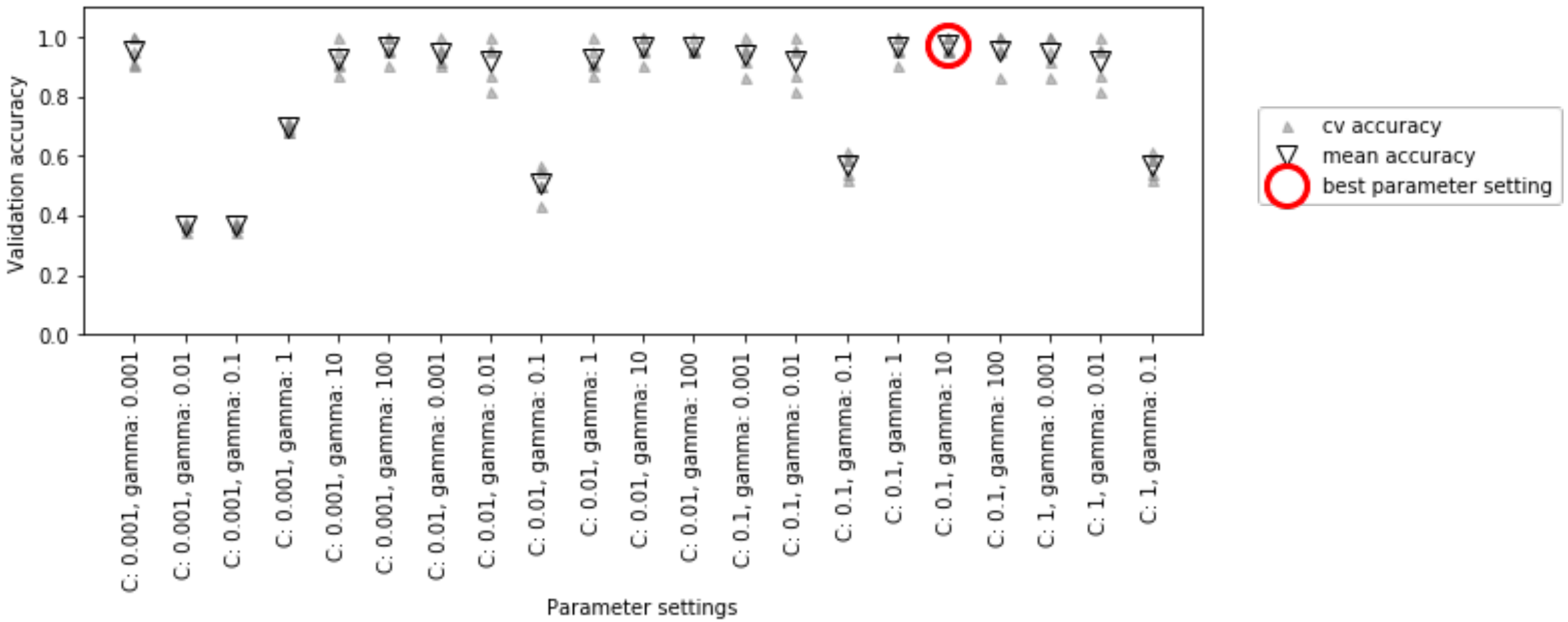
- 중첩 for문으로 parameter 값을 조정해가면서 가장 높은 score를 산출
  - overly optimistic

- data를 한 번 더 split
  - training set / validation set // test set 으로 분할



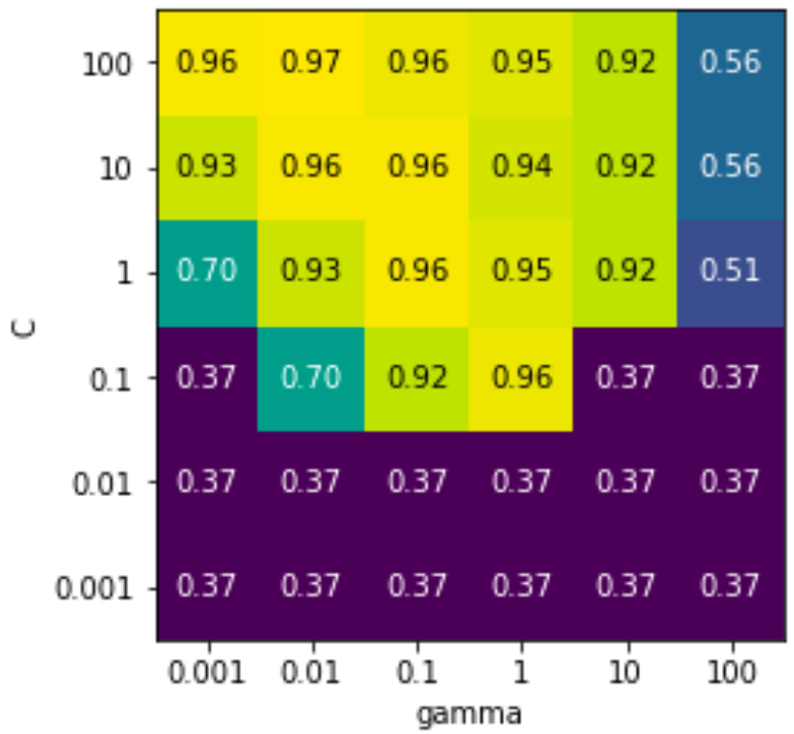
HYUNDAI MOBIS

Grid Search w/ Cross-Validation



Grid Search w/ Cross-Validation

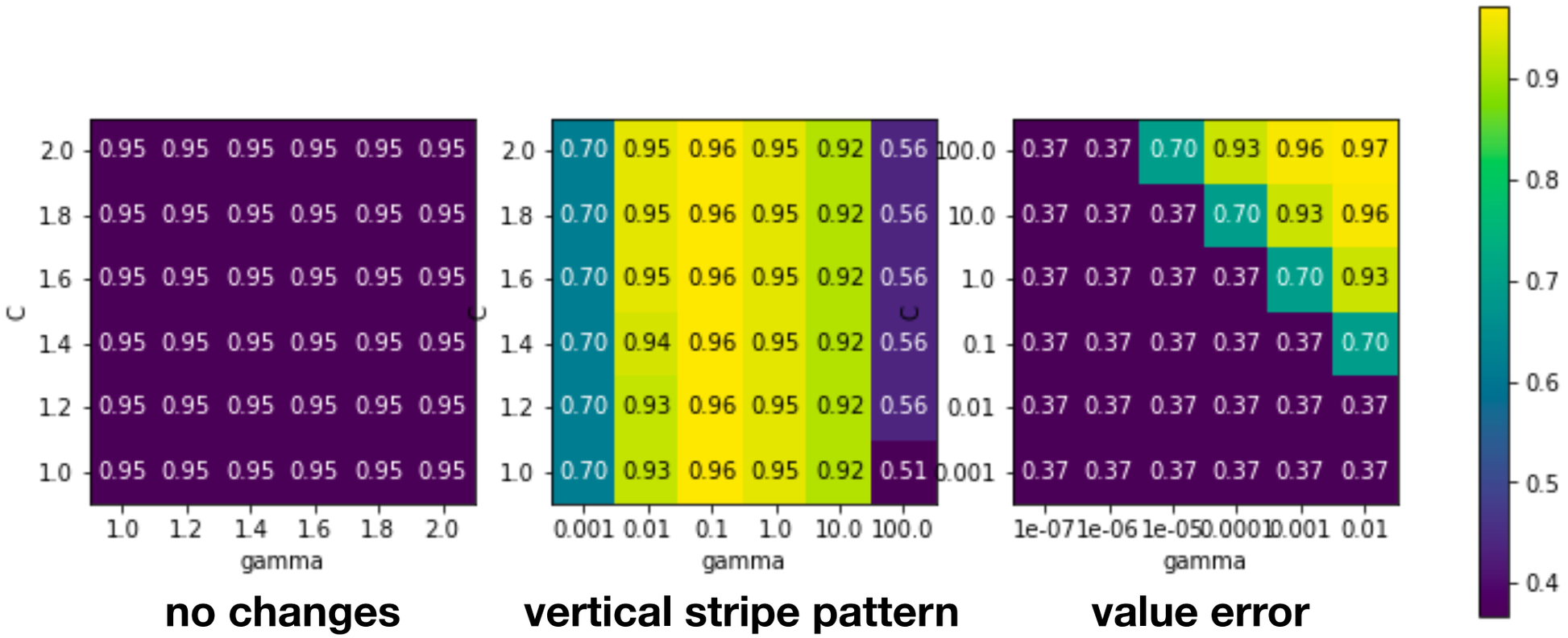
- **.score: test set score**
- .best_params_
- .best_score_
- .cv_results_: DataFrame 형태로 저장되어 있음



HYUNDAI MOBIS

Grid Search w/ Cross-Validation

- grid가 잘못 지정되었을 때



no changes **vertical stripe pattern** **value error**

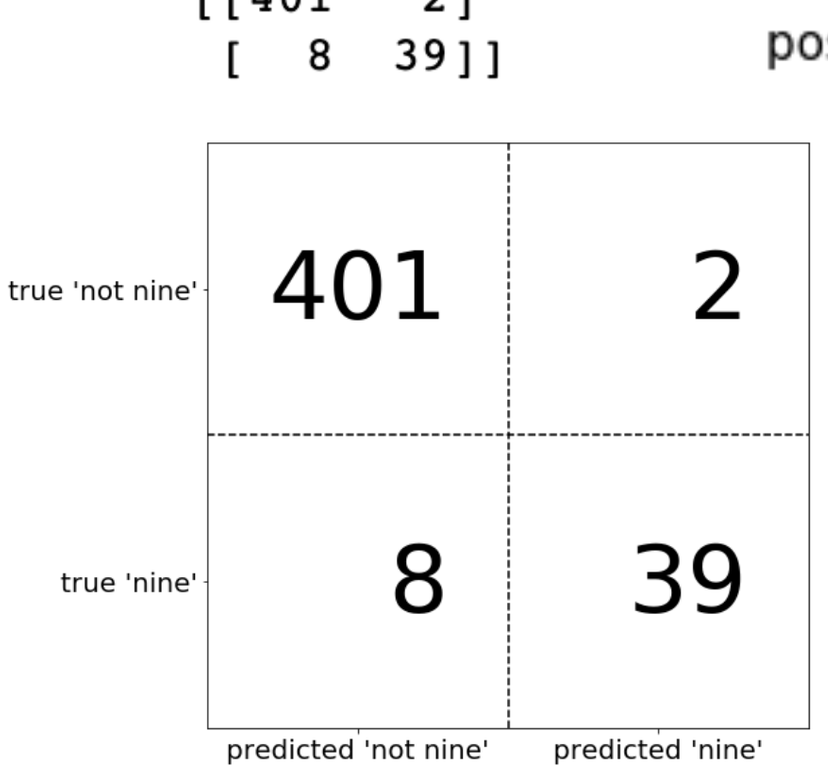
Evaluation

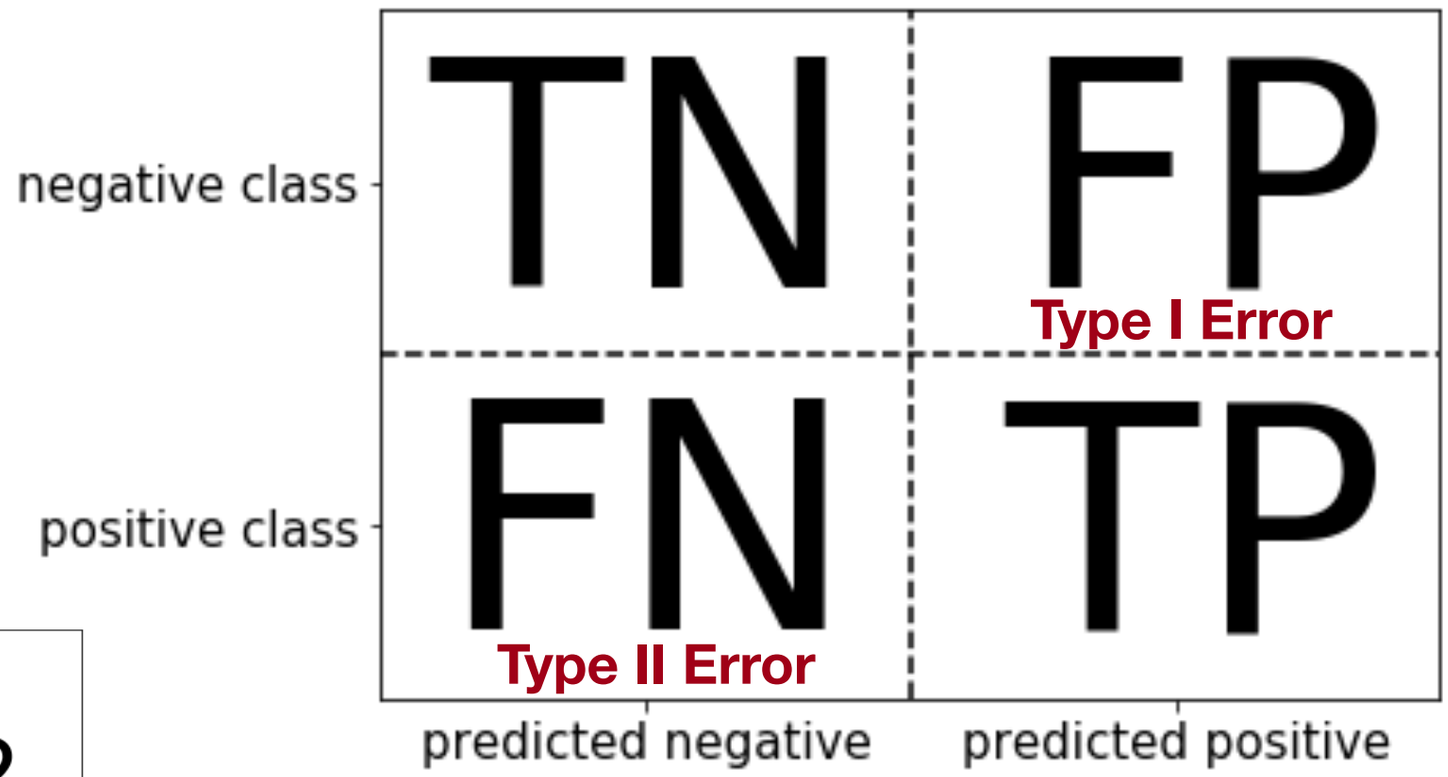
- accuracy는 좋은 평가 기준이 아니다
  - ex. early detection of cancer
    - 암이 있는 사람을 없다고 판단하는 것
    - 암이 없는 사람을 있다고 판단하는 것

> 같은 error로 판단하면 안돼!

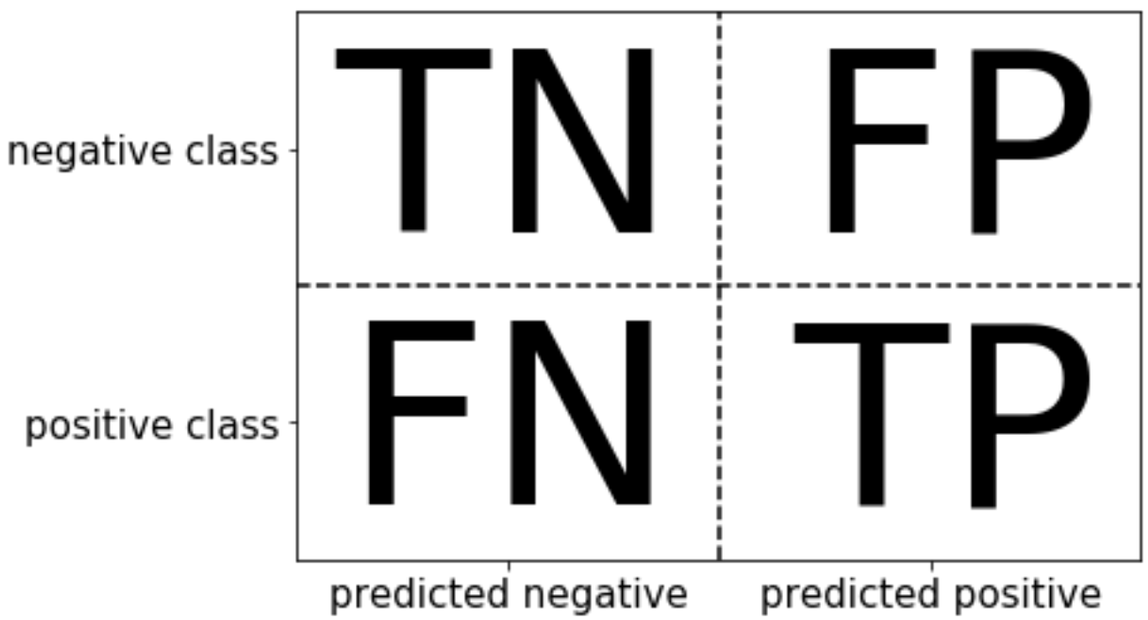
Confusion Matrix

```
Confusion matrix:
[[401   2]
 [  8  39]]
```





Evaluation



$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

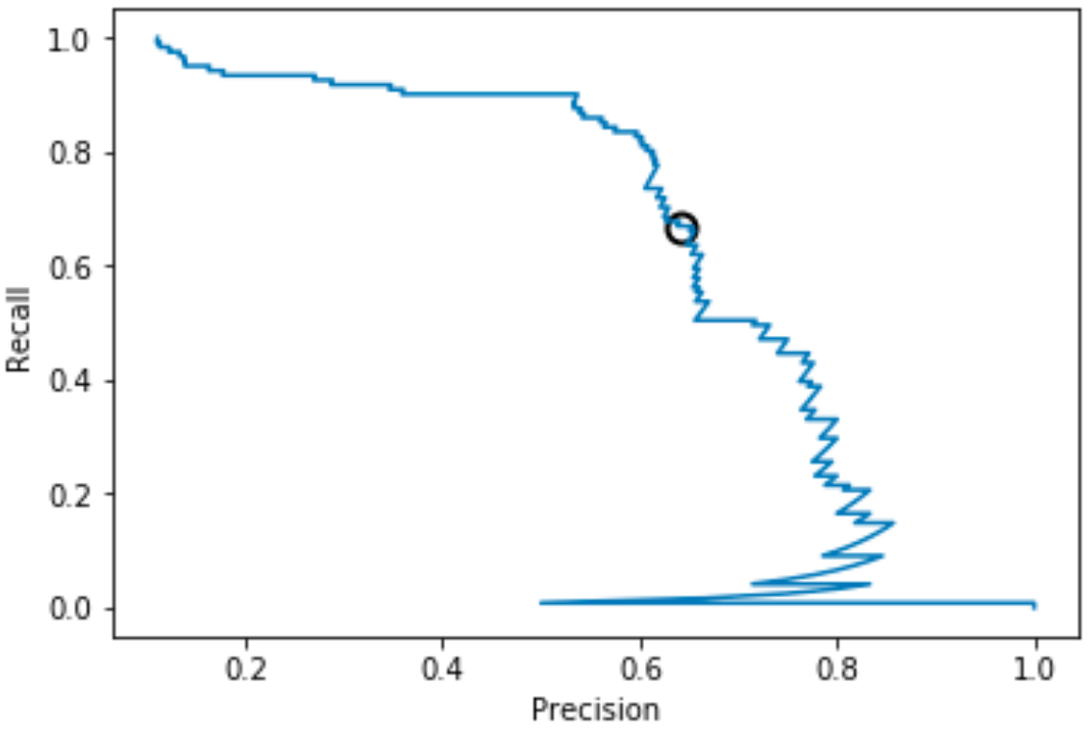
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}$$

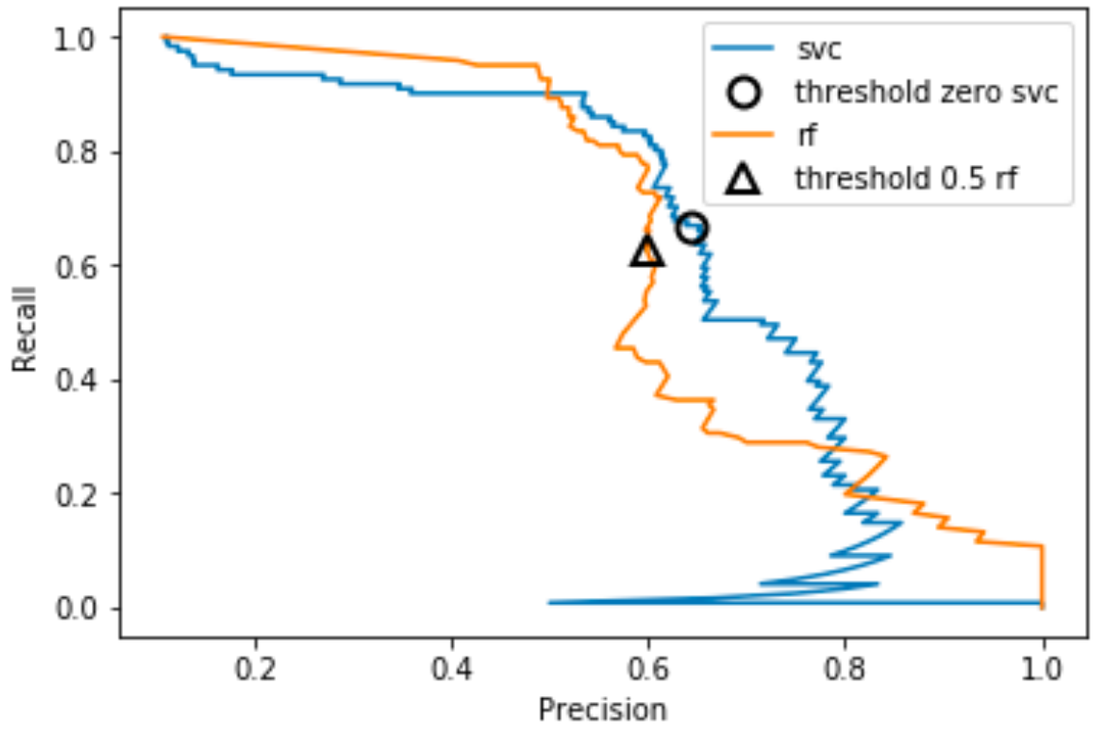
HYUNDAI MOBIS

Evaluation

- Precision
  - 아닌 것을 맞다고 예측하는 것을 피하고 싶음
    - ex. 신약 임상실험: 정말 맞다고 판단될 때에만 실험을 진행
- Recall
  - 맞는 것을 아니라고 예측하는 것을 피하고 싶음
    - ex. 암 진단: 암을 놓치면 안된다
- F-score(f1-score)
  - summarize

Precision-recall Curves





HYUNDAI MOBIS

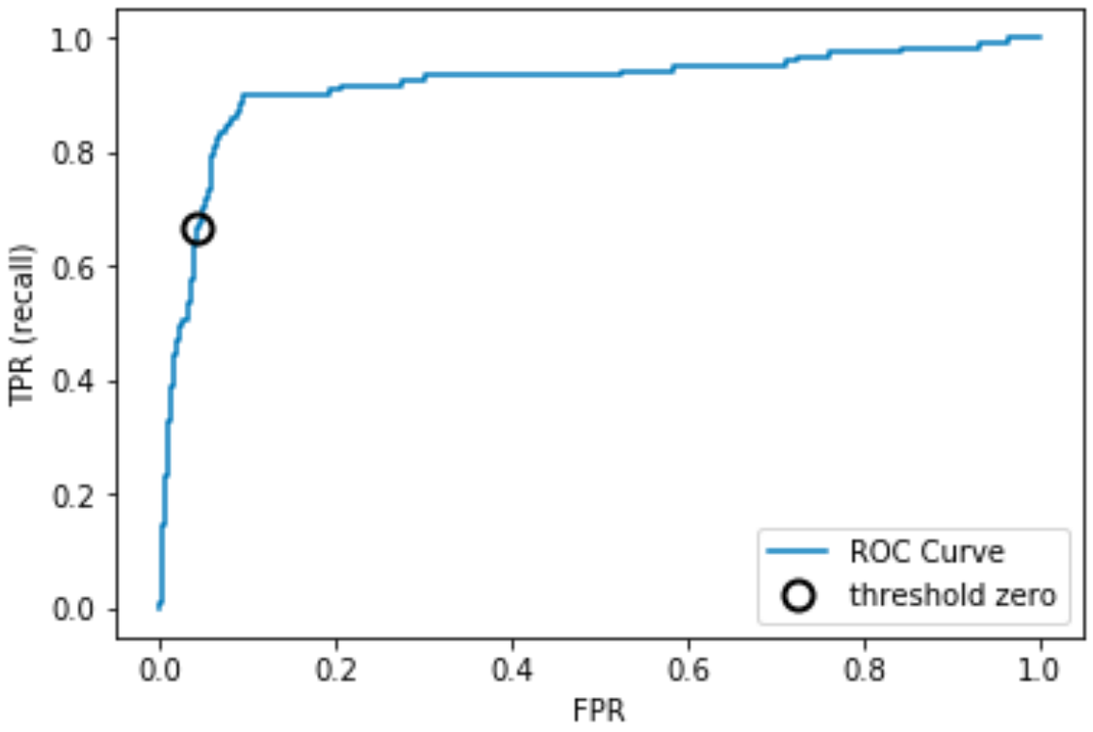
ROC & AUC

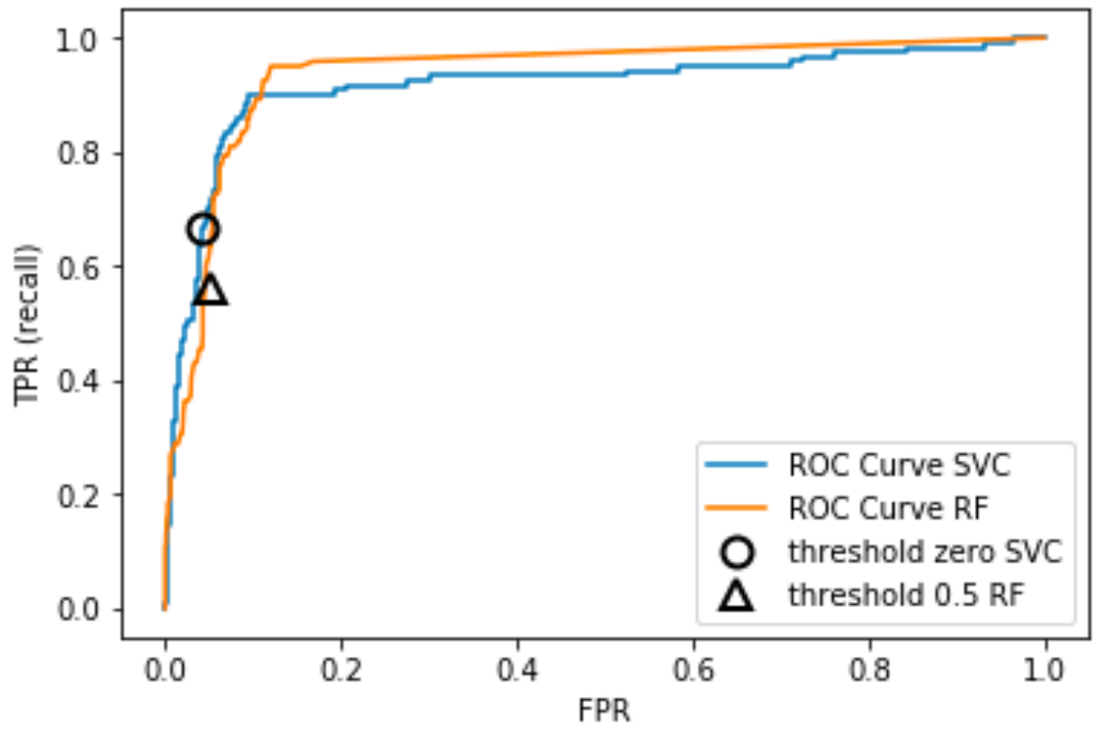
- ROC(Receiver Operating Characteristics Curve)
- AUC(Area Under the Curve)

- FPR(False Positive Rate)
- TPR(True Positive Rate) = recall

$$FPR = \frac{FP}{FP+TN}$$

ROC & AUC

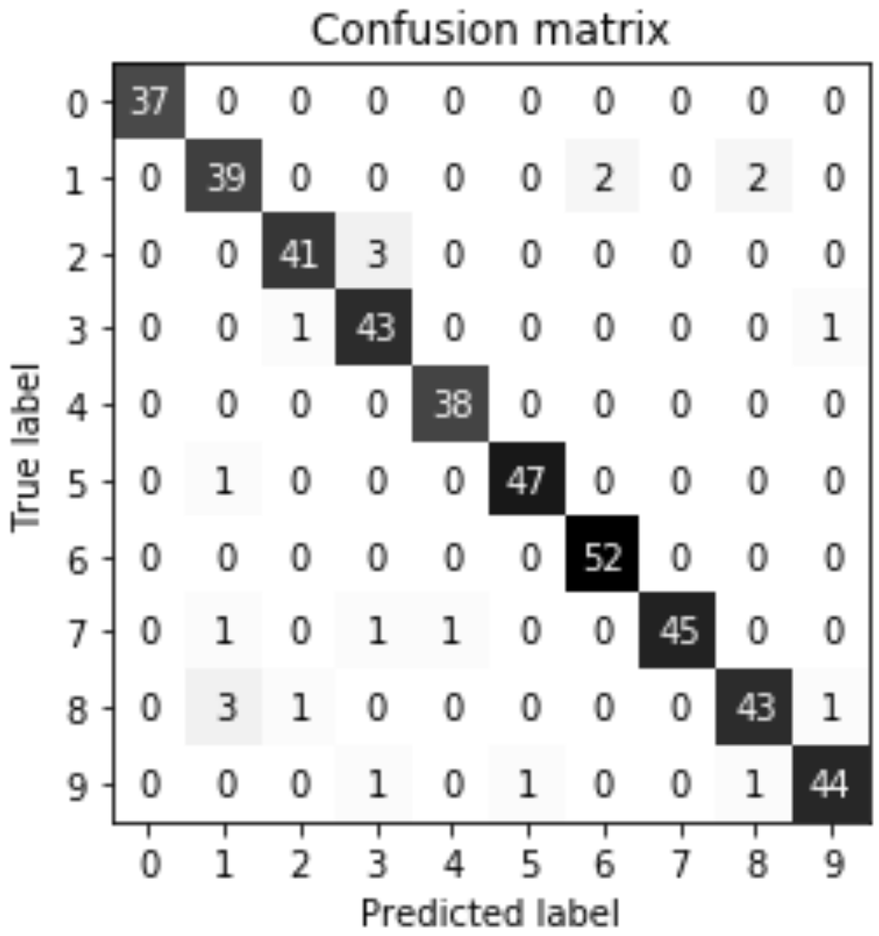




HYUNDAI MOBIS

Matrix for Multiclass

```
Accuracy: 0.953
Confusion matrix:
[[37  0  0  0  0  0  0  0  0  0]
 [ 0 39  0  0  0  0  2  0  2  0]
 [ 0  0 41  3  0  0  0  0  0  0]
 [ 0  0  1 43  0  0  0  0  0  1]
 [ 0  0  0  0 38  0  0  0  0  0]
 [ 0  1  0  0  0 47  0  0  0  0]
 [ 0  0  0  0  0  0 52  0  0  0]
 [ 0  1  0  1  1  0  0 45  0  0]
 [ 0  3  1  0  0  0  0  0 43  1]
 [ 0  0  0  1  0  1  0  0  1 44]]
```



HYUNDAI MOBIS

Matrix for Multiclass

	precision	recall	f1-score	support
0	1.00	1.00	1.00	37
1	0.89	0.91	0.90	43
2	0.95	0.93	0.94	44
3	0.90	0.96	0.92	45
4	0.97	1.00	0.99	38
5	0.98	0.98	0.98	48
6	0.96	1.00	0.98	52
7	1.00	0.94	0.97	48
8	0.93	0.90	0.91	48
9	0.96	0.94	0.95	47
micro avg	0.95	0.95	0.95	450
macro avg	0.95	0.95	0.95	450
weighted avg	0.95	0.95	0.95	450

정규 표현식
(Regular Expression)

- 복잡한 문자열을 처리
- 메타 문자: 원래 문자의 뜻이 아닌 특별한 용도로 사용되는 문자

  . ^ $ * + ? { } [ ] \ | ( )

- 문자 클래스 [ ]: [와] 사이의 문자들과 매치
  - [abc]: a, b, c 중 한 개의 문자와 매치
  - [a-zA-Z]: 알파벳 모두
  - [0-9]: 숫자
  - [^0-9]: 숫자가 아닌 문자

정규식	문자열	매치 여부
[abc]	a	Yes
	before	Yes
	dude	No

HYUNDAI MOBIS

정규 표현식
(Regular Expression)

정규 표현식	설명
\d	= [0-9], 숫자와 매치
\D	= [^0-9], 숫자가 아닌 것과 매치
\s	whitespace 문자와 매치
\S	whitespace 문자가 아닌 것과 매치
\w	=[a-zA-Z0-9], 문자+숫자와 매치
\W	=[^a-zA-Z0-9], 문자+숫자가 아닌 문자와 매치

정규 표현식
(Regular Expression)

- Dot(.): 줄바꿈 문자인 \n을 제외한 모든 문자와 매치됨을 의미

정규식	문자열	매치 여부
a.b	aab	Yes
	a0b	Yes
	abc	No

- a[.]b?

HYUNDAI MOBIS

정규 표현식
(Regular Expression)

- 반복(*): 바로 앞에 있는 문자가 0번 이상 반복

정규식	문자열	매치 여부
ca*t	ct	Yes
	cat	Yes
	caaat	Yes

- 반복(+): 바로 앞에 있는 문자가 1번 이상 반복

정규식	문자열	매치 여부
ca*t	ct	No
	cat	Yes
	caaat	Yes

HYUNDAI MOBIS

정규 표현식
(Regular Expression)

- 반복({m,n}, ?): 반복 횟수 고정시키기
  - ca{2}t
  - ca{2,5}t
  - ca{,3}t
  - ab?c

HYUNDAI MOBIS

정규 표현식
(Regular Expression)

- 정규식을 이용한 문자열 검색

메서드	목적
match()	문자열의 처음부터 정규식과 매치되는지 조사
search()	문자열 전체를 검색하여 정규식과 매치되는지 조사
findall()	정규식과 매치되는 모든 문자열을 리스트로 리턴
finditer()	정규식과 미치되는 모든 문자열을 반복 가능한 객체로 리턴

HYUNDAI MOBIS

정규 표현식
(Regular Expression)

- match 객체의 메서드

메서드	목적
group()	매치된 문자열을 리턴
start()	매치된 문자열의 시작 위치를 리턴
end()	매치된 문자열의 끝 위치를 리턴
span()	매치된 문자열의 (시작, 끝)에 해당되는 튜플을 리턴

HYUNDAI MOBIS

정규 표현식
(Regular Expression)

- 컴파일 옵션

옵션명	약어	설명
DOTALL	S	줄바꿈 문자를 포함하여 모든 문자와 매치할 수 있도록
IGNORECASE	I	대.소문자에 관계 없이 매치할 수 있도록
MULTILINE	M	여러 줄과 매치할 수 있도록
VERBOSE	X	verbose 모드를 사용할 수 있도록