

Grid Search 격자 찾기

Evaluation metrics

평가 지표

python (**scikitlearn**)

MACHINE LEARNING

accurate 미국·영국 ['ækjəreɪt]  영국식  

1. 형용사 정확한
2. 형용사 정밀한
3. 형용사 (목표물에) 명중하는, 정확한 (↔inaccurate)

평가지표의 중요성

“ 어떤 분류가 정확하다 accurate ”라는
말은 사실 정확하지 않다 (정보 생략이 많다)

1. 형용사 정확한
2. 형용사 정밀한
3. 형용사 (목표물에) 명중하는, 정확한 (↔inaccurate)

평가지표의 중요성

Accuracy

= 정확성?

정밀성?

1. 형용사 정확한

2. 형용사 정밀한

3. 형용사 (목표물에) 명중하는, 정확한 (↔inaccurate)

평가지표의 중요성

“ 어떤 분류가 정확하다 accurate ”라는
말은 사실 정확하지 않다 (정보 생략이 많다)

“ 어떤 분류의 평가지표의 수치가 ~~만큼이다 ”
라고 표현해야 함

지난 시간에

pipeline module. Pipeline class

```
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
```

```
lr_tfidf = Pipeline([('vect', tfidf),
                      ('clf', LogisticRegression(solver='liblinear', random_state=0))])

gs_lr_tfidf = GridSearchCV(lr_tfidf, param_grid,
                           scoring='accuracy',
                           cv=5,
                           verbose=1,
                           n_jobs=1)
```

pipeline module. Pipeline class

```
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
```

`sklearn.pipeline.Pipeline`

`class sklearn.pipeline.Pipeline(steps, *, memory=None, verbose=False)` Verb 동사 말 많은,
= 지금 상황 다 설명해줘

데이터 처리, 분류 등의 steps단계를 묶어서 라인으로 만든 class의
인스턴스를 만들 수 있다.

```
lr_tfidf = Pipeline([('vect', tfidf),  
                    ('clf', LogisticRegression(solver='liblinear', random_state=0))])
```




feature_extraction module

```
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import TfidfVectorizer
```

TfidfVectorizer

: 문서를 tf-idf의 **feature matrix**로 벡터변환하는 클래스

BOW (Bag of Words)

고정된 bag (multiset) 가방 자리 를 만들고

D_i 라는 개별 문서의 가방 자리에 해당하는 단어들이 포함되어 있는지 표시

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Vectorizer 벡터화 class

CountVectorizer:

문서 집합에서 단어 토큰을 생성하고
각 단어의 수를 세어 BOW 인코딩한 벡터를 만든다.

TfidfVectorizer:

TF-IDF 방식으로 단어의 중요도를 조정한 BOW 벡터를 만든다.

```
sklearn.feature_extraction.text.TfidfVectorizer
```

```
class sklearn.feature_extraction.text.TfidfVectorizer(*, input='content', encoding='utf-8', decode_error='strict',
```

`sklearn.feature_extraction.text` submodule gathers utilities to build feature vectors from text documents.

`sklearn.feature_extraction.text.TfidfVectorizer`

```
class sklearn.feature_extraction.text.TfidfVectorizer(*, input='content', encoding='utf-8', decode_error='strict',
```

TfidfVectorizer:

TF-IDF 방식으로 단어의 가중치를 조정한 BOW 벡터를 만든다.

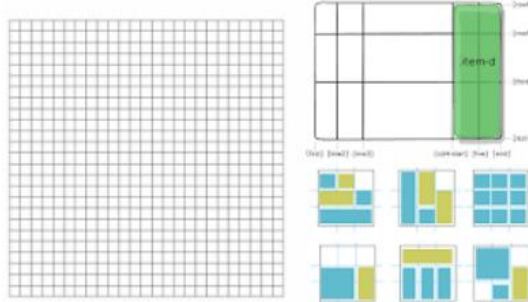
단순 단어 빈도로 접근하는 게 아니라,

어떤 단어가 한 문서에서 많이 나타난 동시에

다른 문서에서는 잘 나타나지 않는 것까지 고려하기 위한 개념

TF-IDF (Term Frequency-Inverse Document Frequency)

GridSearch



Grid

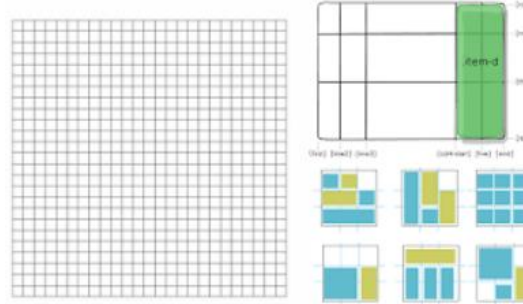
Pipeline [vectorizer, classifier]

GridSearch (Pipeline, parameter, 점수 지표, 몇 분할, 많았은, n_P.U. 개수)

```
lr_tfidf = Pipeline([('vect', tfidf),
                      ('clf', LogisticRegression(solver='liblinear', random_state=0))])

gs_lr_tfidf = GridSearchCV(lr_tfidf, param_grid,
                             scoring='accuracy',
                             cv=5,
                             verbose=1,
                             n_jobs=1)
```


GridSearch



Grid

Pipeline [vectorizer, classifier]

GridSearch (Pipeline, parameter, 점수 지표, 몇 분할, 많았으면, n_P.U. 개수)



Install User Guide API Examples More ▾

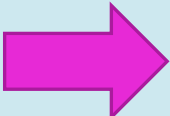
UpNext

scikit-learn 0.23.1
Other versions

Cite us if you use the software.

model_selection.GridSearchCV

using
model_selection.GridSearchCV



sklearn.model_selection.GridSearchCV

```
class sklearn.model_selection.GridSearchCV(estimator, param_grid, *, scoring=None, n_jobs=None, iid='deprecated', refit=True, cv=None, verbose=0, pre_dispatch='2*n_jobs', error_score=nan, return_train_score=False)
```

[source]

Exhaustive search over specified parameter values for an estimator.

Important members are fit, predict.

GridSearchCV implements a “fit” and a “score” method. It also implements “predict”, “predict_proba”, “decision_function”, “transform” and “inverse_transform” if they are implemented in the estimator used.

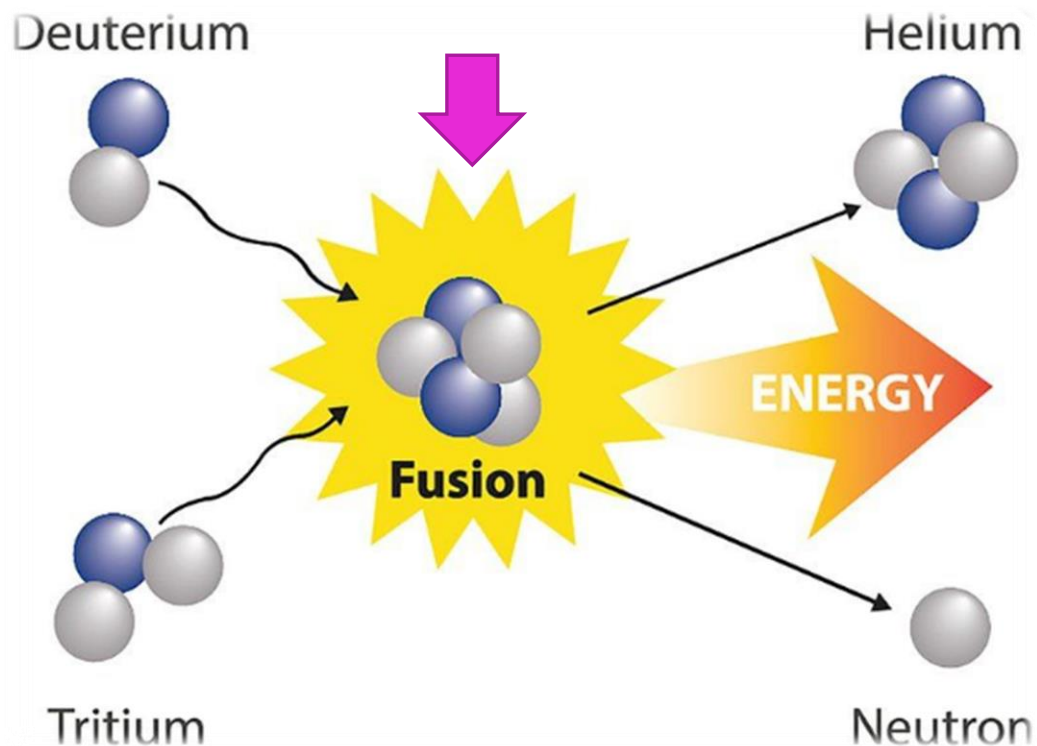
분류기 성능 평가

실제 actual class 대비

얼마나 잘 맞았는가?

Confusion matrix (혼합 행렬)

실제 라벨과 예측 라벨의 일치 개수를 matrix 형태로 표현



Actual
Class

Estimated Class

Prediction

	Estimated Class	
	Prediction	
Actual Class	1	0
	1	0
1	True Positive	False Negative
0	False Positive	True Negative

Confusion matrix (혼합 행렬)

True Positive (TP)

- 참 + 양성으로 예측

True: 참

Positive: 양성으로 예측

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

Confusion matrix (혼합 행렬)

True Negative (TN)

- 참 + 음성으로 예측

True: 참

Negative: 음성으로 예측

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

Confusion matrix (혼합 행렬)

False Positive (FP)

- 거짓 + 양성으로 예측

False: 거짓

Positive: 양성으로 예측

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

Confusion matrix (혼합 행렬)

False Negative (FN)

- 거짓 + 음성으로 예측

False: 거짓

Negative: 음성으로 예측

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

Confusion Matrix

True Positive (TP)

True Negative (TN)

False Positive (FP)

False Negative (FN)

Metrics for classification performance

Accuracy (정확도)

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

Error Rate (오차율)

$$Errorrate = \frac{FP + FN}{TP + TN + FP + FN} = (1 - Accuracy)$$

Precision (정밀도)

$$Precision = \frac{TP}{TP + FP} \text{ (PPV: Positive Predict Value)}$$

Specificity (특이도)

$$Specificity = \frac{TN}{TN + FP} \text{ (TNR: True Negative Rate)}$$

Sensitivity (민감도)

$$Sensitivity = \frac{TP}{TP + FP} \text{ (TPR: True Positive Rate)}$$

Metrics for classification performance

Accuracy (정확도)

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

Error Rate (오차율)

$$Errorrate = \frac{FP + FN}{TP + TN + FP + FN} = (1 - Accuracy)$$

Precision (정밀도)

$$Precision = \frac{TP}{TP + FP} \text{ (PPV: Positive Predict Value)}$$

Specificity (특이도)

$$Specificity = \frac{TN}{TN + FP} \text{ (TNR: True Negative Rate)}$$

Sensitivity (민감도)

$$Sensitivity = \frac{TP}{TP + FP} \text{ (TPR: True Positive Rate)}$$

일상적으로 많이 쓰는 말 (균형잡힌 클래스)

정확도 (Accuracy, ACC, Agreement)

전체 데이터 대비 정확하게 예측한 개수의 비율

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$ACC = 1 - ERR$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

오차율 (Error Rate, ERR)

전체 데이터 대비 부정확하게 예측 개수의 비율

$$ERR = \frac{FP + FN}{TP + TN + FP + FN}$$

$$ERR = 1 - ACC$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

불균형적 Imbalanced Dataset

대학의 학사경고자 평균비율 3%

하버드 입학 지원자의 합격률 2%

이메일 수신자 중 2%만 물건을 구매

Metrics for Imbalanced Dataset



이탈리아 코판사가 원천기술을 가지고 있는 다양한 형태의 의료용 면봉 제품. <출처=코판그

Re call 다시 부름?

우리가 본래 원했던 **실제 양성** retrieved 목표 중에
참으로 연관성 relevant 있는 것이 **불러와진** 비율

Fraction of relevant instances
that are retrieved

민감도 (Recall, True Positive Rate, Sensitivity)

양성이라고 분류된 중에, 참 (연관된) 비율.

실제 양성 중에, 참을 민감하게 다시 불러왔는가?

$$RECALL(TPR) = \frac{TP}{TP + FN} = \frac{TP}{P} \quad \text{Actual Class}$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

정밀도 (**Precision**, **P**ositive **P**redictive **V**alue)

양성이라고 분류한 결과 중에서 진짜 참인 비율.

양성이라고 **분류**한 결과가 **얼마나 정밀**한지 나타냄

$$PRECISION(PPV) = \frac{TP}{TP + FP}$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

F₁ Score (F-measure)

Precision 정밀 & Recall 민감도의 통합 지표

$$F_1 = 2 \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

F-measures do not take the true negatives

F_β Score (F-measure)

Recall이 Precision보다 **베타** 제공 만큼 **중요**할 때

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

$$\begin{aligned} F_{\beta=2} & \text{ (if prec = 0.6, recall = 0.4)} \\ &= 5 \cdot \text{prec} \cdot \text{recall} / (4\text{prec} + \text{recall}) \\ &= 5 \cdot 0.6 \cdot 0.4 / (2.4 + 0.4) = 1.2 / 2.8 \doteq 0.43 \end{aligned}$$

$$\begin{aligned} F_{\beta=1} & \text{ (if prec = 0.6, recall = 0.4)} \\ &= 2 \cdot 0.6 \cdot 0.4 / (0.6 + 0.4) = 0.48 / 1 = 0.48 \end{aligned}$$

$$\begin{aligned} F_{\beta=2} & \text{ (if prec = 0.4, recall = 0.6)} \\ &= 5 \cdot \text{prec} \cdot \text{recall} / (4\text{prec} + \text{recall}) \\ &= 5 \cdot 0.4 \cdot 0.6 / (1.6 + 0.6) = 1.2 / 2.2 \doteq 0.55 \end{aligned}$$

F-measures do not take the true negatives

F_β Score (F-measure)

Recall이 Precision보다 베타 제곱 만큼 더 중요할 때

weighs recall higher than precision (by placing more emphasis on false negatives)

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

$$F_1 = 2 \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

F-measures do not take the true negatives

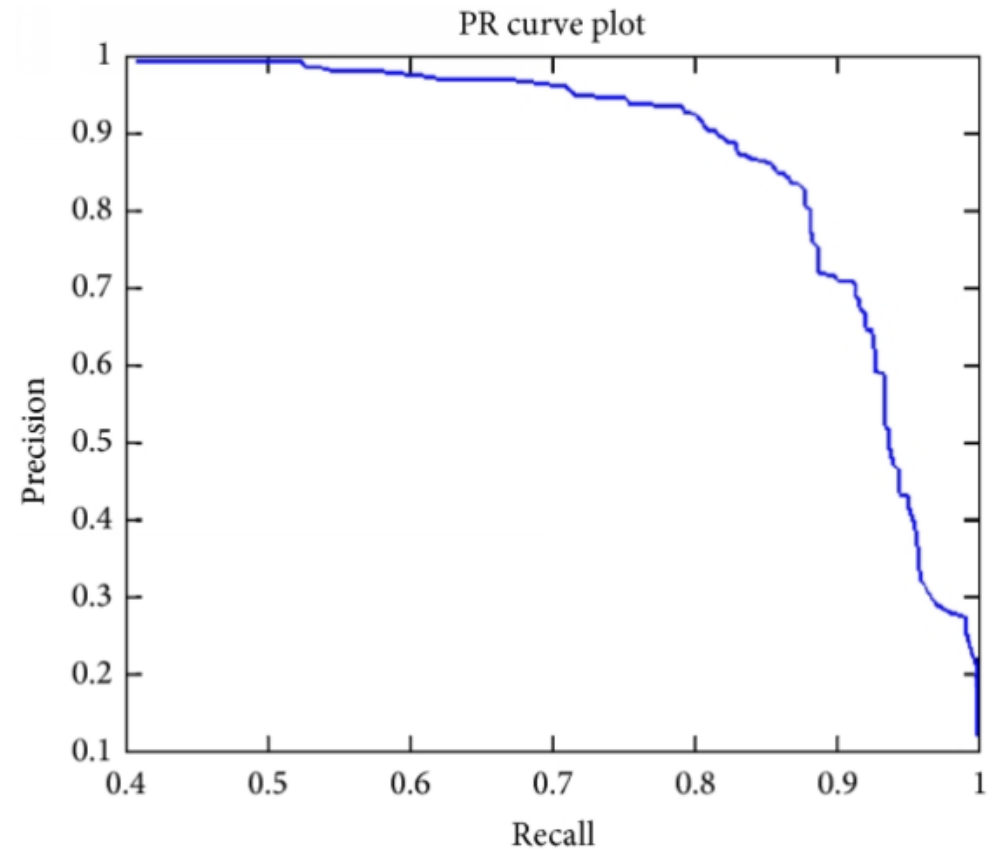
$$\begin{aligned} F_{\beta=2} & \text{ if prec} = 0.6, \text{ recall} = 0.4 \\ &= 5 \cdot \text{prec} \cdot \text{recall} / (4\text{prec} + \text{recall}) \\ &= 5 \cdot 0.6 \cdot 0.4 / (2.4 + 0.4) = 1.3 / 2.8 = 0.46 \end{aligned}$$

$$\begin{aligned} F_{\beta=1} & \text{ (if prec} = 0.6, \text{ recall} = 0.4) \\ &= 0.48 / 1 = 0.48 \end{aligned}$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

Precision-Recall Curve (PR-curve)

Test데이터의 Class가 불균등할 때 사용
Positive가
Negative 보다 더 의미 있을 때 사용

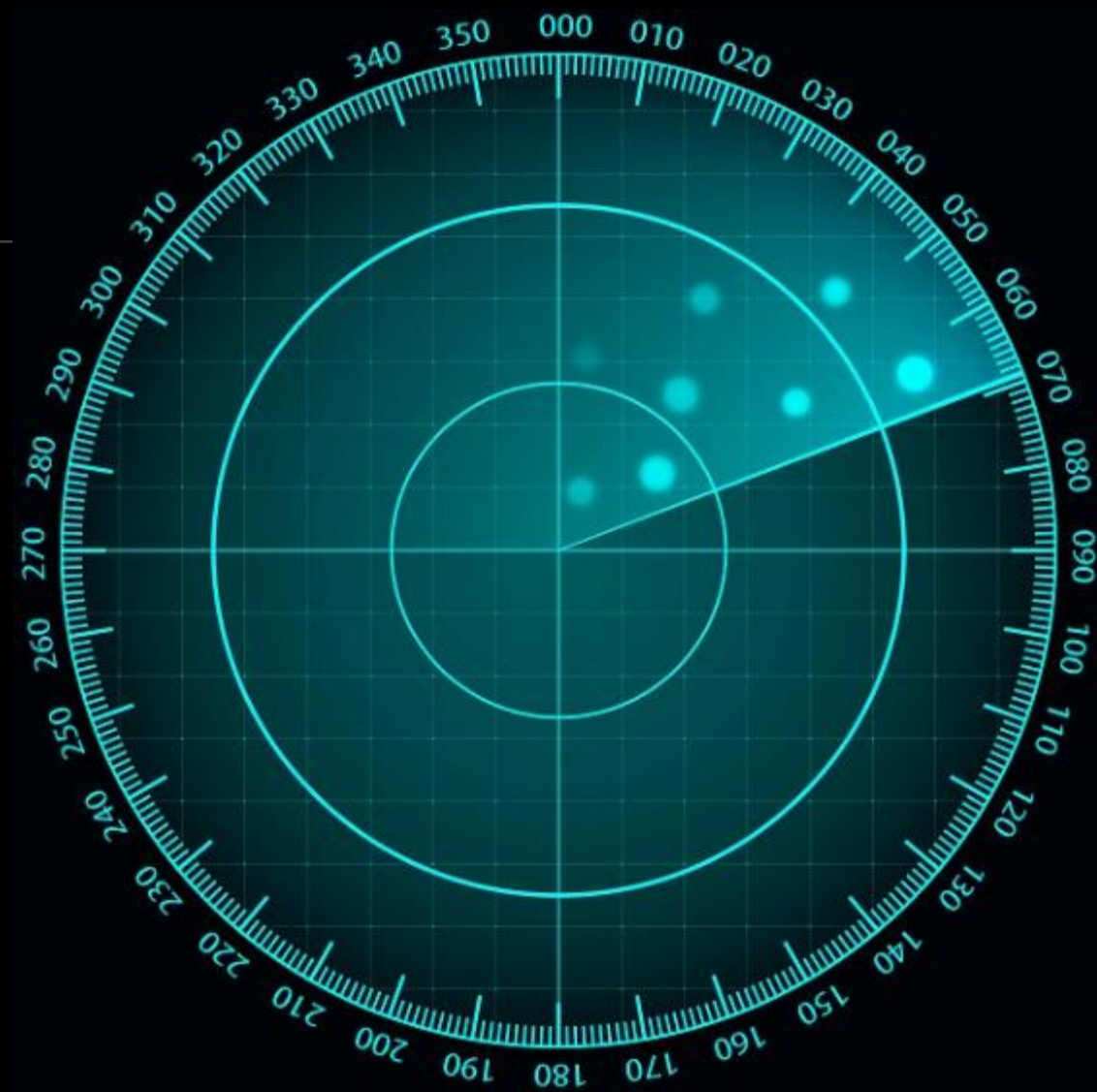
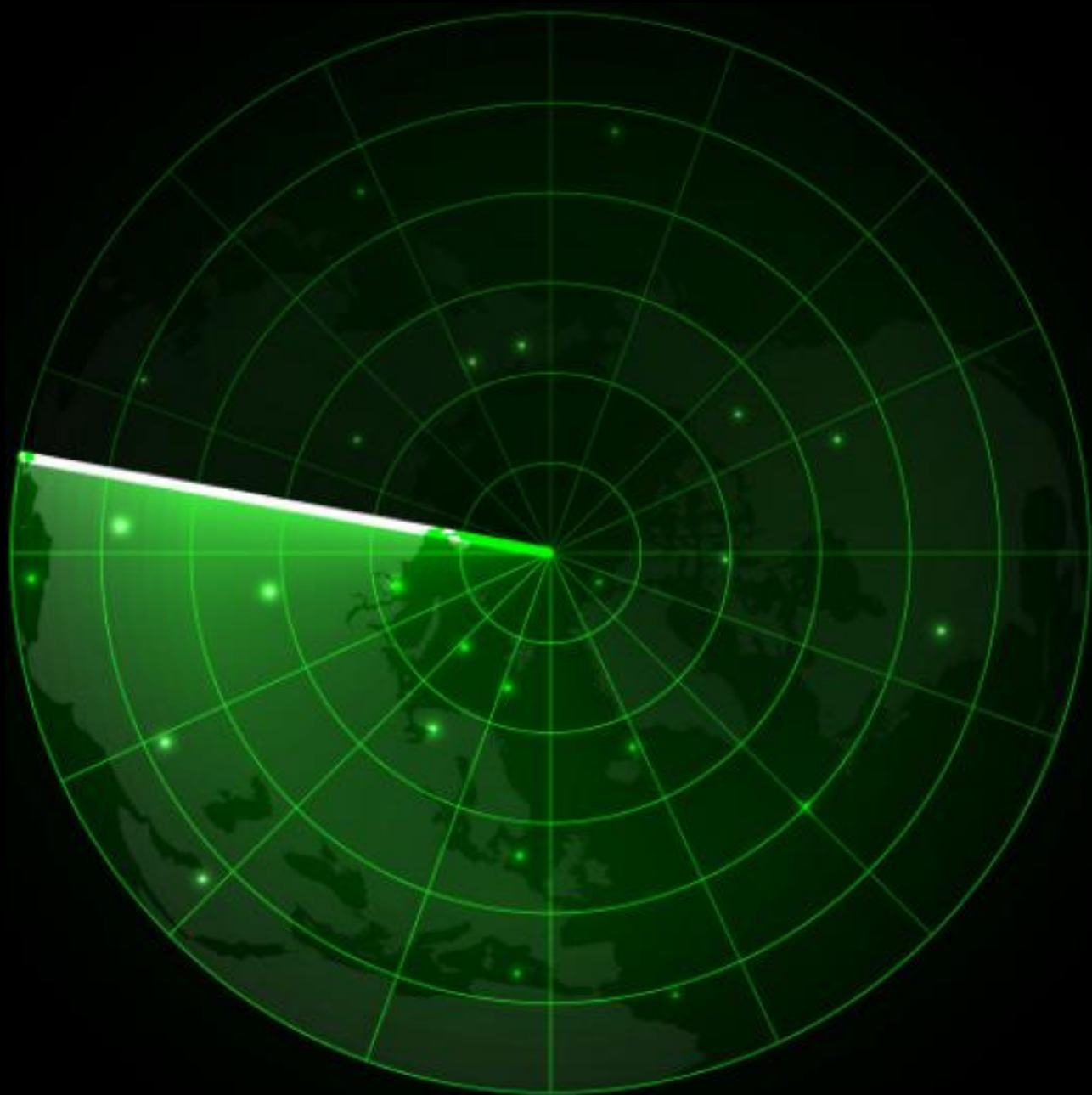


특이도 (Specificity, True Negative Rate)

모든 실제 음성 중에 참 음성이라고 밝혀진 비율
음성(병없음)중에, 참 음성을 잘 진단하여 **스펙**으로,
병 걸릴 환경에서, 참 음성이라면 **특이** (면역)하다.

$$SPC = \frac{TN}{TN + FP} = \frac{TN}{N}$$

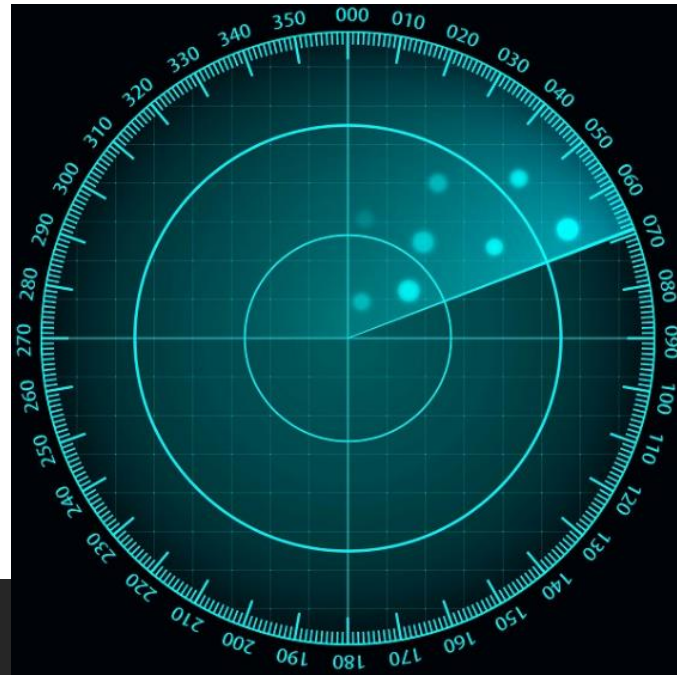
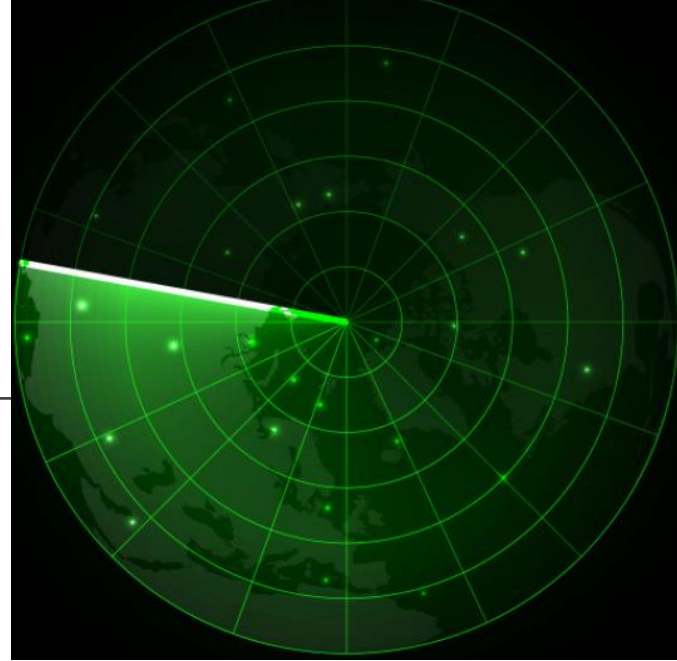
		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative



Radar 레이더 1
VS

Radar 레이더 2

**수신자 성능을
어떻게 비교하지?**



ROC Curve

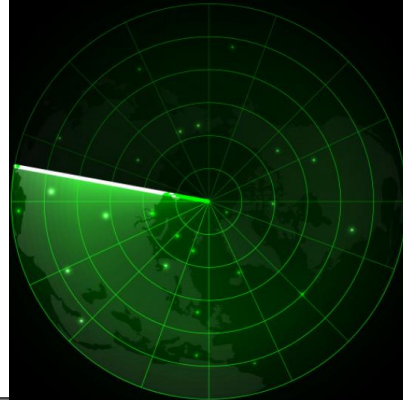
Reciever **O**perating **C**haracteristics

수신자

조작

특성

ROC curve



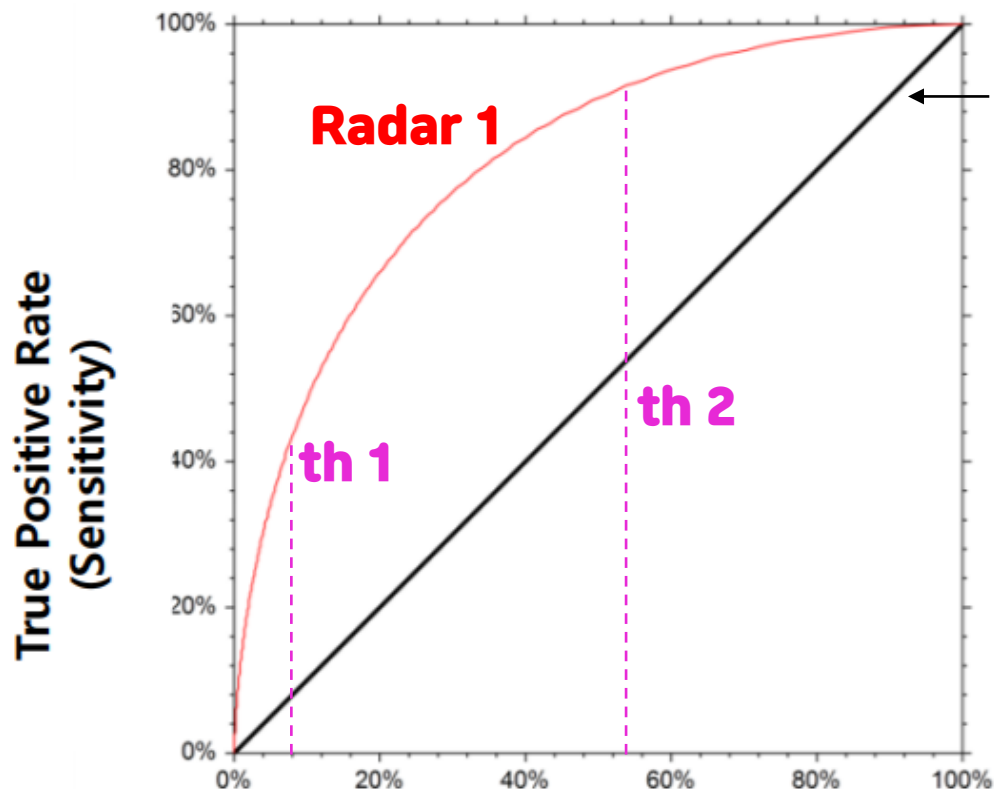
2차 세계 대전 중 레이더 신호 감지 (수신자 조작) 이론에서 시작

Basic Principles of ROC Analysis (Charles Metz, 1978)

분류기 패러미터의 경계치(Threshold)를 조작 operating,

민감도-특이도간 비율을 도식화

ROC curve

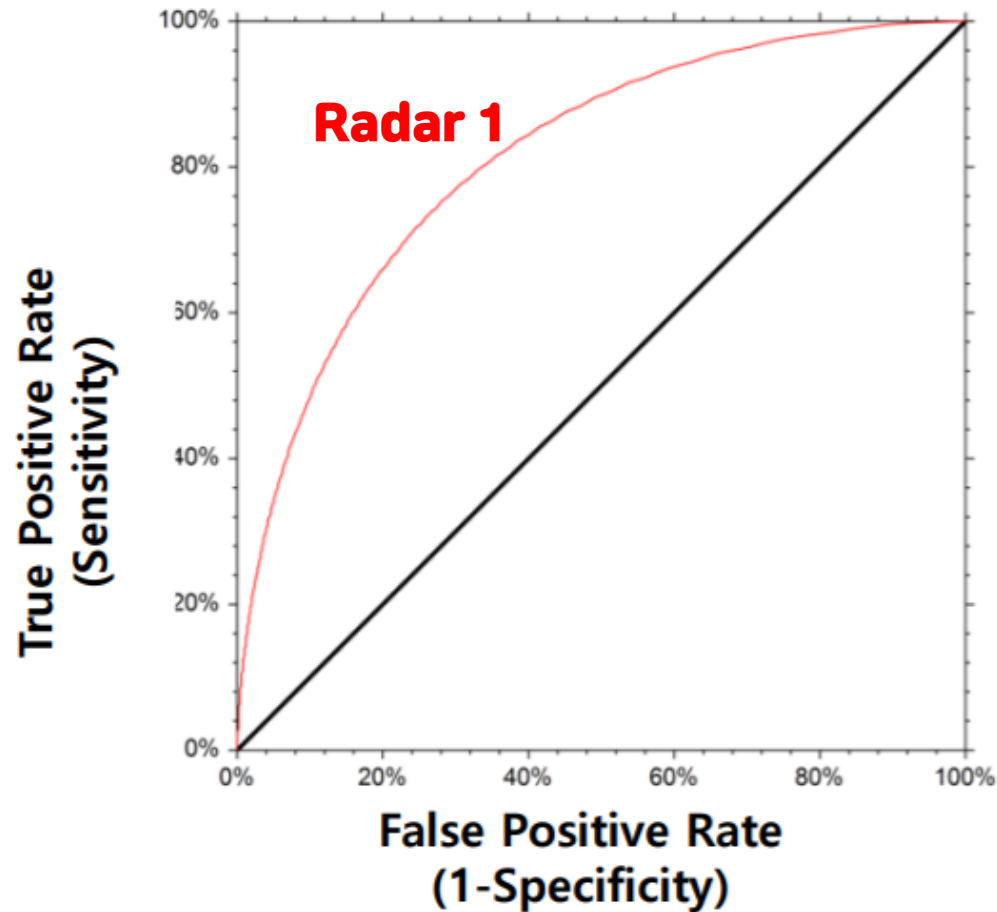


← TPR = FPR의 가상선

False Positive Rate
(1-Specificity)

= False Alarm 잘못된 경고 (적군 비행기 발견!, 근데 실수)

ROC curve



$$\text{Sensitivity}(TPR) = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$\begin{aligned} FPR &= 1 - \text{Specificity}(TNR) \\ &= 1 - \frac{TN}{TN + FP} = 1 - \frac{TN}{N} \end{aligned}$$

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

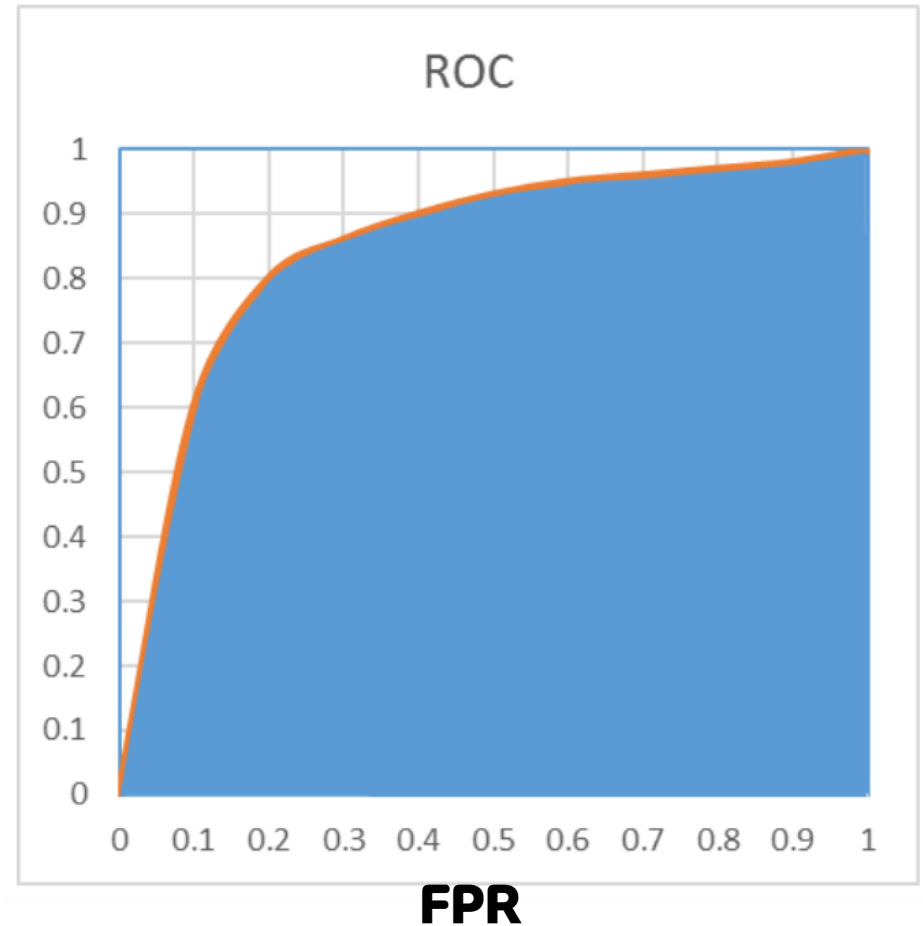
AUC, Area Under Curve

ROC curve의 하단의 넓이를 의미

ROC curve를 단순한 single metrics 로 표현 할 수 있음

대각선을 중심으로 상단에 붙어 있을 수록 높은 성능을 의미

TPR



정리

Classifier를 평가하는 방법

- Recall (TPR, Sensitivity)
- Precision (PPV)
- F-score
- PR-curve

- Specificity (TNR)
- ROC-curve, AUC