



Performance Evaluation

강필성

고려대학교 산업경영공학부

Bflysoft & WIGO AI LAB

AGENDA

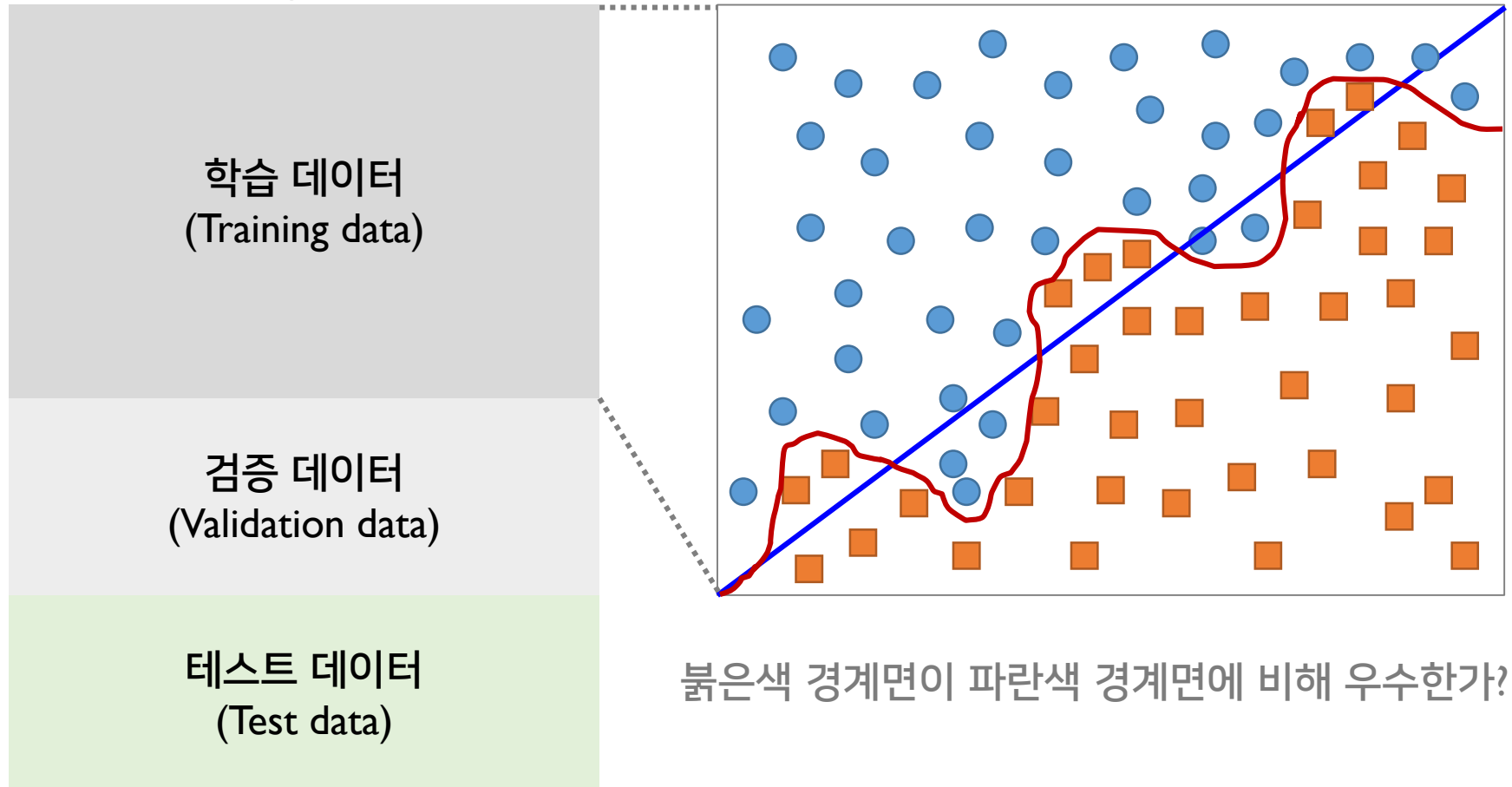
01 모델 평가의 필요성

02 회귀 모형의 성능 평가

03 분류 모형의 성능 평가

모델 평가의 필요성

- 과적합^{Overfitting}: 학습 데이터에 존재하는 불필요한 정보까지 학습하여 일반화 성능이 저하되는 현상

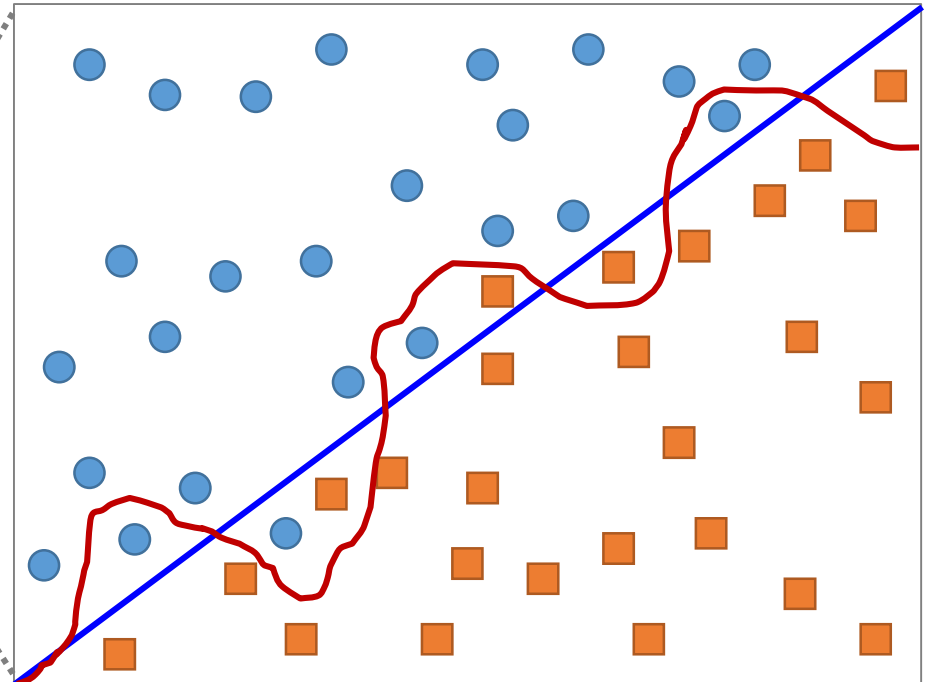


모델 평가의 필요성

- 과적합^{Overfitting}: 학습 데이터에 존재하는 불필요한 정보까지 학습하여 일반화 성능이 저하되는 현상



학습 데이터를 완벽히 외우는 것은 일반화 성능을 저하시키는 위험(과적합)이 존재!



모델 평가의 필요성

- 분류 문제나 회귀 문제를 풀 수 있는 다양한 알고리즘 존재
 - ✓ Classification:
 - Naïve bayes, linear discriminant, k-nearest neighbor, classification trees, etc.
 - ✓ Prediction:
 - Multiple linear regression, neural networks, regression trees, etc.
- 어떤 알고리즘은 최적의 파라미터 설정이 필요함
 - ✓ k-인접이웃기법: 이웃 개체의 수(k), 인공 신경망: 은닉 노드의 수 등
- 주어진 문제를 해결하기 위한 최적의 방법론을 선택하기 위해 개별 모델을 동등한 조건에서 평가할 필요가 있음
 - ✓ 검증 데이터: 다양한 파라미터 조합 중 최적의 파라미터를 찾는 데 주로 사용
 - ✓ 테스트 데이터: 여러 기계학습 알고리즘 중 최적의 알고리즘을 찾는 데 주로 사용

AGENDA

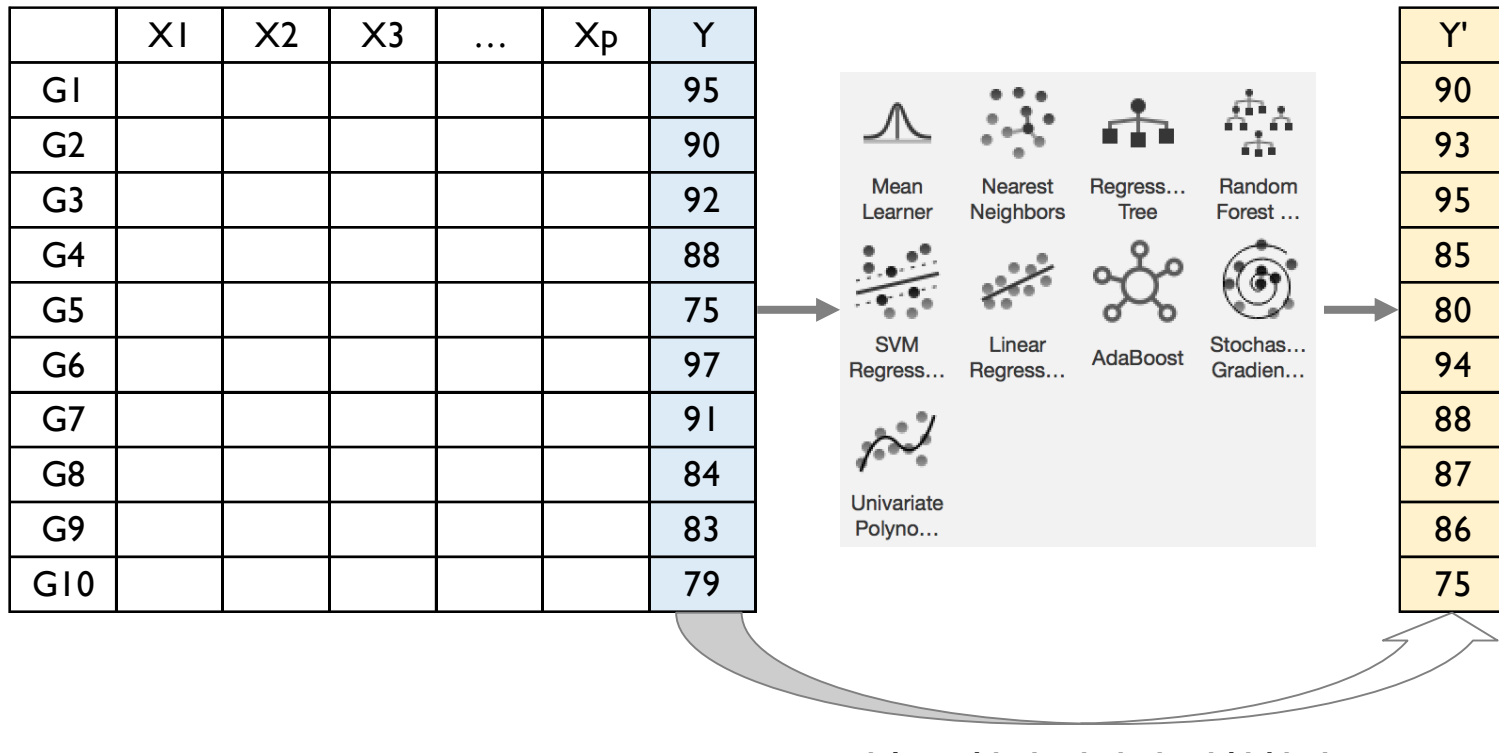
01 모델 평가의 필요성

02 회귀 모형의 성능 평가

03 분류 모형의 성능 평가

회귀모형 성능평가

- 예시: 설비 파라미터 X에 대한 제품의 수율(y) 예측



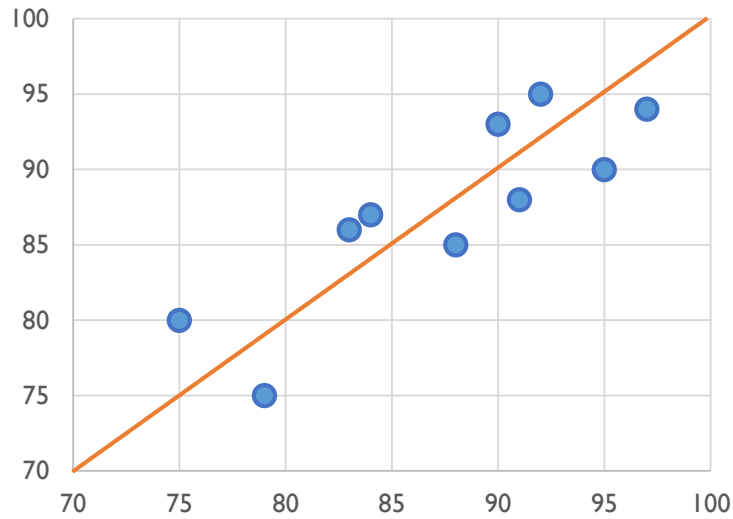
예측 모형이 얼마나 정확한가?

회귀모형 성능평가

- 성능지표 I: 평균오차 (Average Error)

- ✓ 실제 값에 비해 과대/과소 추정 여부를 판단
- ✓ 부호로 인해 잘못된 결론을 내릴 위험이 있음

$$\text{Average Error} = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)$$



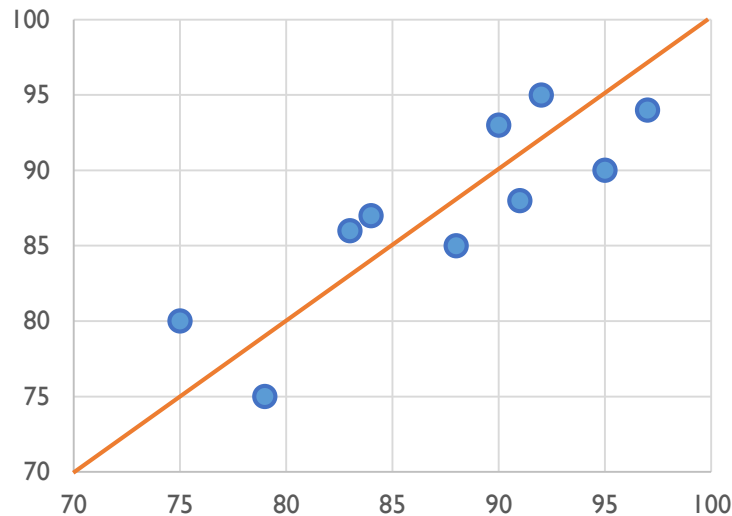
Y	Y'	Y-Y'
95	90	5
90	93	-3
92	95	-3
88	85	3
75	80	-5
97	94	3
91	88	3
84	87	-3
83	86	-3
79	75	4
Average Error		0.1

회귀모형 성능평가

- 성능지표 2: 평균 절대 오차(Mean absolute error; MAE)

✓ 실제 값과 예측 값 사이의 절대적인 오차의 평균을 이용

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - y'_i|$$



Y	Y'	Y-Y'
95	90	5
90	93	3
92	95	3
88	85	3
75	80	5
97	94	3
91	88	3
84	87	3
83	86	3
79	75	4
MAE		3.5

회귀모형 성능평가

- 성능지표 3: Mean absolute percentage error (MAPE)

- ✓ MAE의 단점: 실제 값과 절대적인 차이에 대한 정보만 제공하고, 상대적인 차이에 대한 정보를 제공하지 못함
- ✓ 아래 두 예시의 MAE는 모두 1임

Y	Y'	Y-Y'
1	0	1
1	2	1
1	0	1
1	2	1
1	0	1
1	2	1
1	0	1
1	2	1
1	0	1
1	2	1
MAE		1

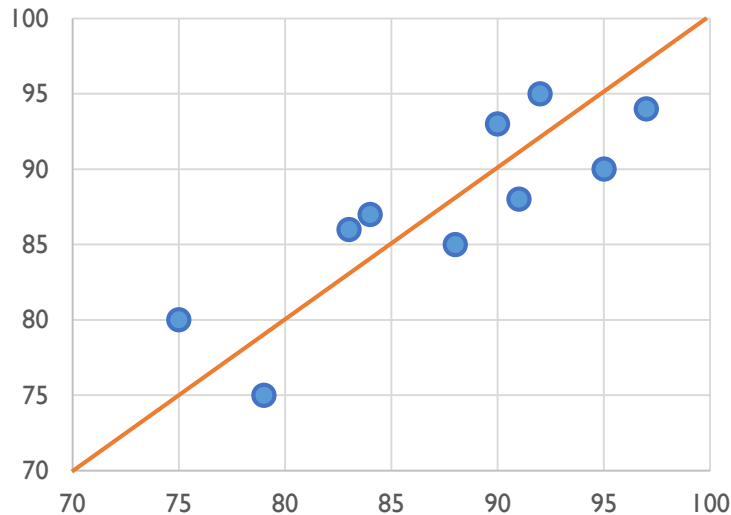
Y	Y'	Y-Y'
100	99	1
100	101	1
100	99	1
100	101	1
100	99	1
100	101	1
100	99	1
100	101	1
100	99	1
100	101	1
MAE		1

회귀모형 성능평가

- 성능지표 3: Mean absolute percentage error (MAPE)

- ✓ 실제값 대비 얼마나 예측 값이 차이가 있는지를 %로 표현
- ✓ 상대적인 오차를 추정하는데 주로 사용

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - y'_i}{y_i} \right|$$



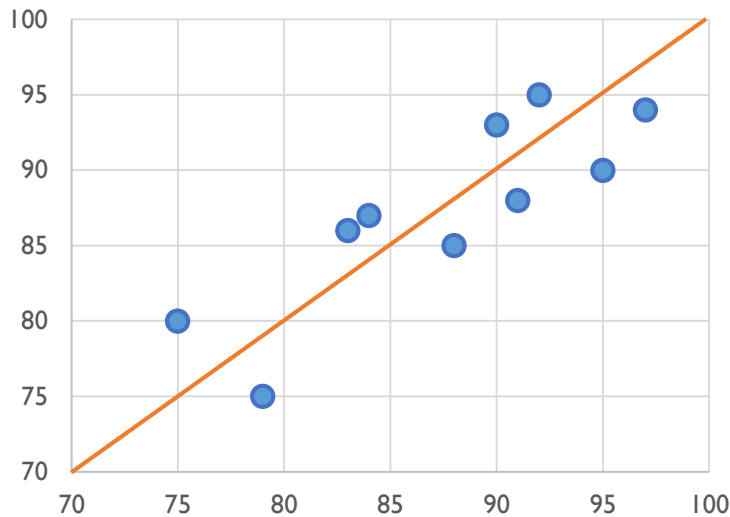
Y	Y'	Y-Y'	Y-Y' / Y
95	90	5	5.26%
90	93	3	3.33%
92	95	3	3.26%
88	85	3	3.41%
75	80	5	6.67%
97	94	3	3.09%
91	88	3	3.30%
84	87	3	3.57%
83	86	3	3.61%
79	75	4	5.06%
MAE		3.5	4.06%

회귀모형 성능평가

- 성능지표 4 & 5: (Root) Mean squared error ((R)MSE)

✓ 부호의 영향을 제거하기 위해 절대값이 아닌 제곱을 취한 지표

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2}$$



Y	Y'	(Y-Y') ²
95	90	25
90	93	9
92	95	9
88	85	9
75	80	25
97	94	9
91	88	9
84	87	9
83	86	9
79	75	16
MSE		12.9

$$\text{RMSE} = \sqrt{12.9} = 3.59$$

AGENDA

01 모델 평가의 필요성

02 회귀 모형의 성능 평가

03 분류 모형의 성능 평가





분류 모형 성능 평가

- 예시: 성별 분류

✓ 한 사람의 체지방률만을 이용하여 남성/여성 분류

									
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2

✓ 단순 분류기: 체지방률이 20보다 크면 여성으로, 작으면 남성으로 분류











									
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.8	23.2
M	F	M	M	F	F	M	F	F	F

✓ 위 분류기의 성능을 어떻게 평가할 것인가?

분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 실제 범주와 예측된 범주를 이용하여 생성한 2X2 행렬

									
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2
M	F	M	M	F	F	M	F	F	F

✓ 위 결과에 대한 정오 행렬은 다음과 같이 생성됨

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3

분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 정오행렬을 통해 다음과 같이 다양한 분류 성능 평가 지표를 계산할 수 있음

Confusion Matrix		Predicted	
		1(+)	0(-)
Actual	1(+)	n_{11}	n_{10}
	0(-)	n_{01}	n_{00}

- 민감도(Sensitivity), true positive, 재현율(recall) = $n_{11}/(n_{11}+n_{10})$
- 특이도(Specificity, true negative) = $n_{00}/(n_{01}+n_{00})$
- 정밀도(Precision) = $n_{11}/(n_{11}+n_{01})$
- 제1종 오류(Type I error, false negative) = $n_{10}/(n_{11}+n_{10})$
- 제2종 오류(Type II error, false positive) = $n_{01}/(n_{01}+n_{00})$

분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 정오행렬을 통해 다음과 같이 다양한 분류 성능 평가 지표를 계산할 수 있음

Confusion Matrix		Predicted	
		1(+)	0(-)
Actual	1(+)	n_{11}	n_{10}
	0(-)	n_{01}	n_{00}

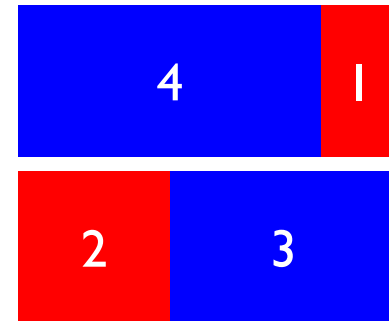
- 오분류율(Misclassification error) = $(n_{01} + n_{10}) / (n_{11} + n_{10} + n_{01} + n_{00})$
- 정분류율(Accuracy = 1 - misclassification error) = $(n_{11} + n_{00}) / (n_{11} + n_{10} + n_{01} + n_{00})$
- 균형 정확도 (Balanced correction rate) =
$$\sqrt{\frac{n_{11}}{n_{11} + n_{10}} \cdot \frac{n_{00}}{n_{01} + n_{00}}}$$
- F1 measure (정밀도와 재현율의 조화평균) =
$$F1 \text{ measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

분류 모형 성능 평가

- 정오 행렬 Confusion Matrix

✓ 이전 예시에서 여성(F)을 1(+) 범주로 정의할 경우

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	2	3



- Sensitivity: $4/5 = 0.8$, Specificity: $3/5 = 0.6$
- Recall: $4/5 = 0.8$, Precision: $4/6 = 0.67$
- Type I error: $1/5 = 0.2$, Type II error: $2/5 = 0.4$
- Misclassification error: $(1+2)/(4+1+2+3) = 0.3$, accuracy = 0.7
- Balanced correction rate: $\sqrt{0.8 \times 0.6} = 0.69$
- F1 measure: $(2 \times 0.8 \times 0.67) / (0.8 + 0.67) = 0.85$

분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 새로운 분류기: 체지방률이 θ 보다 크면 여성으로 분류



✓ 레코드들을 체지방률의 내림차순으로 정렬



✓ 분류를 위한 최적의 cut-off를 어떻게 설정할 것인가?

분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

▪ If $\theta = 24$,

Confusion Matrix		Predicted	
		F	M
Actual	F	2	3
	M	1	4

- Misclassification error: 0.4
- Accuracy: 0.6
- Balanced correction rate: 0.57
- F1 measure = 0.5

분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

▪ If $\theta = 22$,

Confusion Matrix		Predicted	
		F	M
Actual	F	4	1
	M	1	4

- Misclassification error: 0.2
- Accuracy: 0.8
- Balanced correction rate: 0.8
- F1 measure = 0.8

분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

✓ 다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	F
2	25.4	M
3	24.2	F
4	23.6	F
5	22.7	F
6	21.5	M
7	19.9	F
8	15.7	M
9	10.0	M
10	8.9	M

▪ If $\theta = 18$,

Confusion Matrix		Predicted	
		F	M
Actual	F	5	0
	M	2	3

- Misclassification error: 0.2
- Accuracy: 0.8
- Balanced correction rate: 0.77
- F1 measure = 0.83

분류 모형 성능 평가

- 분류 알고리즘의 Cut-off 설정

- ✓ 일반적으로 분류 알고리즘은 특정 범주에 속할 확률(probability)이나 우도(likelihood) 값을 생성함
- ✓ 동일한 확률값 하에서도 Cut-off가 어떻게 설정되느냐에 따라서 분류 성능이 크게 좌우되는 상황이 발생할 수 있음
- ✓ 분류 알고리즘간의 정확한 비교를 위해서는 Cut-off에 독립적인 측정 지표가 필요함
- ✓ 리프트 도표(Lift charts), receiver operating characteristic (ROC) curve 등이 사용

분류 모형 성능 평가

- ROC Curve 예시

- ✓ Glass 불량 진단 문제:

- Glass의 불량(NG) 여부를 판별
 - 총 100장의 Glass 중 20장의 Glass가 불량
 - 불량 확률: 0.2
 - Label: 1(NG), 0(G)

분류 모형 성능 평가

- 특정 분류 알고리즘에 의해 산출된 NG 범주에 속할 확률과 실제 Label 정보

Glass	P(NG)	Label	Glass	P(NG)	Label	Glass	P(NG)	Label	Glass	P(NG)	Label
1	0.976	1	26	0.716	1	51	0.41	0	76	0.186	0
2	0.973	1	27	0.676	0	52	0.406	1	77	0.183	0
3	0.971	0	28	0.672	0	53	0.378	0	78	0.178	0
4	0.967	1	29	0.662	0	54	0.376	0	79	0.176	0
5	0.937	0	30	0.647	0	55	0.362	0	80	0.173	0
6	0.936	1	31	0.64	1	56	0.355	0	81	0.17	0
7	0.929	1	32	0.625	0	57	0.343	0	82	0.133	0
8	0.927	0	33	0.624	0	58	0.338	0	83	0.12	0
9	0.923	1	34	0.613	1	59	0.335	0	84	0.119	0
10	0.898	0	35	0.606	0	60	0.334	0	85	0.112	0
11	0.863	1	36	0.604	0	61	0.328	0	86	0.093	0
12	0.862	1	37	0.601	0	62	0.313	0	87	0.086	0
13	0.859	0	38	0.594	0	63	0.285	1	88	0.079	0
14	0.855	0	39	0.578	0	64	0.274	0	89	0.071	0
15	0.847	1	40	0.548	0	65	0.273	0	90	0.069	0
16	0.845	1	41	0.539	1	66	0.272	0	91	0.047	0
17	0.837	0	42	0.525	1	67	0.267	0	92	0.029	0
18	0.833	0	43	0.524	0	68	0.265	0	93	0.028	0
19	0.814	0	44	0.514	0	69	0.237	0	94	0.027	0
20	0.813	0	45	0.51	0	70	0.217	0	95	0.022	0
21	0.793	1	46	0.509	0	71	0.213	0	96	0.019	0
22	0.787	0	47	0.455	0	72	0.204	1	97	0.015	0
23	0.757	1	48	0.449	0	73	0.201	0	98	0.01	0
24	0.741	0	49	0.434	0	74	0.2	0	99	0.005	0
25	0.737	0	50	0.414	0	75	0.193	0	100	0.002	0

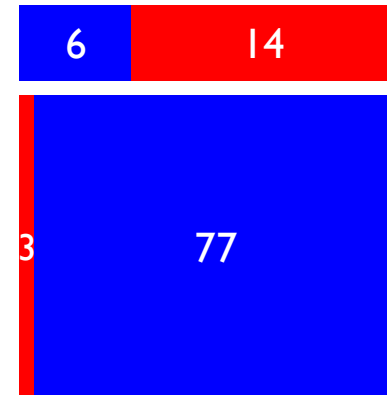
분류 모형 성능 평가

- 정오행렬

- ✓ Cut-off를 0.9로 설정할 경우

- NG if $P(NG) > 0.9$, else G

Confusion Matrix		Predicted	
		M	B
Actual	M	6	14
	B	3	77



- Misclassification error = 0.17
- Accuracy = 0.83

- ✓ 이 모델은 우수한 분류 모델인가?

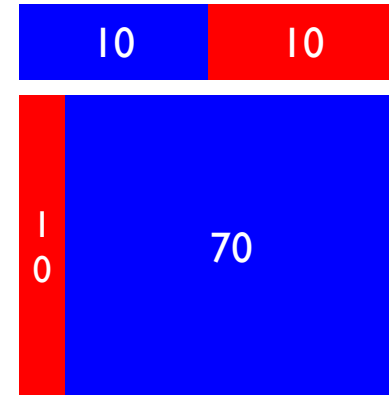
분류 모형 성능 평가

- 정오행렬

- ✓ Cut-off를 0.9로 설정할 경우

- NG if $P(NG) > 0.8$, else G

Confusion Matrix		Predicted	
		M	B
Actual	M	10	10
	B	10	70



- Misclassification error = 0.20

- Accuracy = 0.80

- ✓ 이 모델은 이전 모델보다 열등한 모델인가?

분류 모형 성능 평가

- ROC 생성 절차

- ✓ 모든 개체를 $P(\text{interesting class})$ 를 기준으로 내림차순 정렬
- ✓ 가능한 모든 Cut-off 경우에 대해 True Positive Rate와 False Positive Rate를 계산
 - $P(NG)$ 에 동물이 없을 경우 이론적으로 101개의 cut-off 설정이 가능
- ✓ X축이 False Positive Rate, Y축이 True Positive Rate가 되는 2차원 그래프 도시

분류 모형 성능 평가

- ROC 생성 절차

✓ 첫 번째 Cut-off 설정

Glass	P(NG)	Label
1	0.976	1
2	0.973	1
3	0.971	0
4	0.967	1
5	0.937	0
⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	0	20
	G	0	80

$$\text{TPR} = \frac{0}{20} = 0$$

$$\text{FPR} = \frac{0}{80} = 0$$

분류 모형 성능 평가

- ROC 생성 절차

✓ 두 번째 Cut-off 설정

Glass	P(NG)	Label	TPR	FPR
			0	0
1	0.976	1		
2	0.973	1		
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	1	19
	G	0	80

$$\text{TPR} = \frac{1}{20} = 0.05$$

$$\text{FPR} = \frac{0}{80} = 0$$

분류 모형 성능 평가

- ROC 생성 절차

✓ 세 번째 Cut-off 설정

Glass	P(NG)	Label	TPR	FPR
			0	0
1	0.976	1	0.05	0
2	0.973	1		
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	2	18
	G	0	80

$$\text{TPR} = \frac{2}{20} = 0.10$$

$$\text{FPR} = \frac{0}{80} = 0$$

분류 모형 성능 평가

- ROC 생성 절차

✓ 네 번째 Cut-off 설정

Glass	P(NG)	Label	TPR	FPR
			0.00	0.00
1	0.976	1	0.05	0.00
2	0.973	1	0.10	0.00
3	0.971	0		
4	0.967	1		
5	0.937	0		
⋮	⋮	⋮	⋮	⋮

Confusion Matrix		예측	
		NG	G
실제	NG	2	18
	G	1	79

$$\text{TPR} = \frac{2}{20} = 0.10$$

$$\text{FPR} = \frac{1}{80} = 0.0125$$

분류 모형 성능 평가

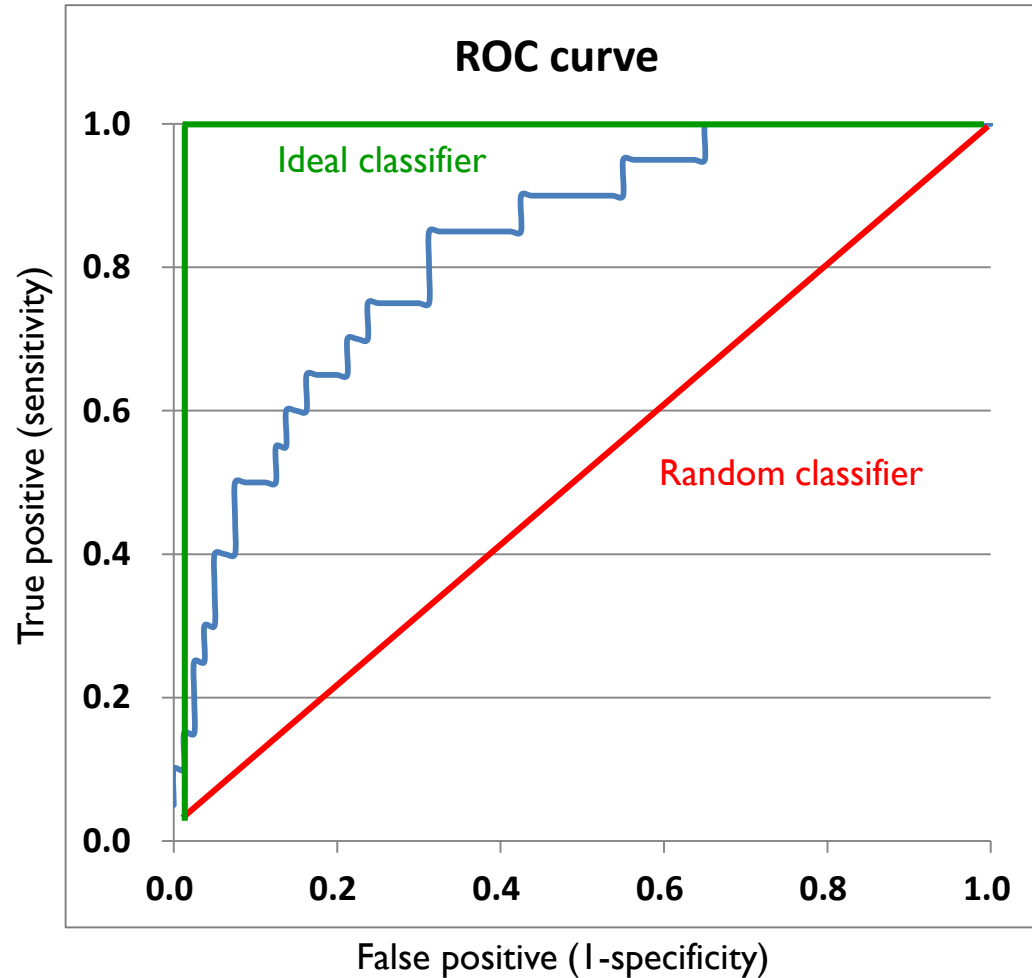
- ROC 생성 절차

- ✓ 모든 가능한 Cut-off 값에 대한 TPR/FPR 계산 완료
- ✓ FPR을 x축으로 하고, TPR을 y축으로 하는 그래프 생성

Glass	P(NG)	Label	TPR	FPR
			0.000	0.000
1	0.976	1	0.050	0.000
2	0.973	1	0.100	0.000
3	0.971	0	0.100	0.013
4	0.967	1	0.150	0.013
5	0.937	0	0.150	0.025
6	0.936	1	0.200	0.025
7	0.929	1	0.250	0.025
8	0.927	0	0.250	0.038
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
96	0.019	0	1.000	0.950
97	0.015	0	1.000	0.963
98	0.01	0	1.000	0.975
99	0.005	0	1.000	0.988
100	0.002	0	1.000	1.000

분류 모형 성능 평가

- ROC Curve 범위



분류 모형 성능 평가

- Area Under ROC Curve (AUROC)

- ✓ ROC curve 아래의 면적
- ✓ 이상적인 분류기는 1의 값을 갖고, 무작위 분류기는 0.5의 값을 가짐
- ✓ Cut-off에 독립적인 알고리즘 성능 평가 지표로 사용될 수 있음

