Data Mining

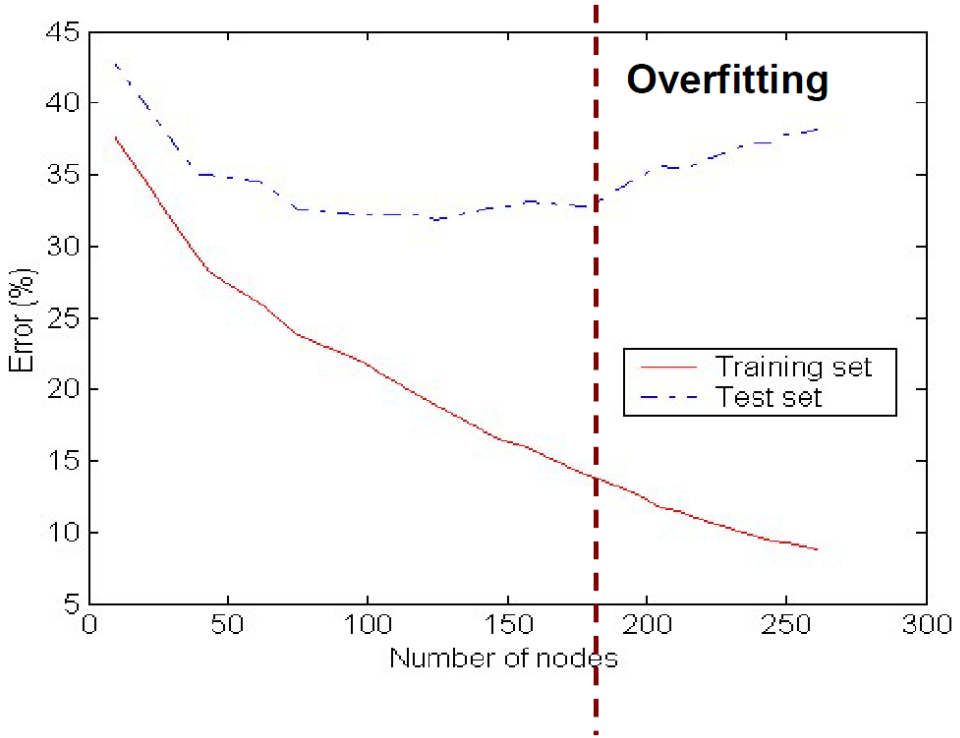
2017. 3. 24.

김영철

모델 평가 방법과 수치

Notes on Overfitting

- Model overfitting
  - 트레이닝 데이터에 딱 맞는 분류 방법을 택하여 일반 데이터에 대한 에러율이 증가



Estimating Generalization Errors

- Optimistic approach

  Generalization error = training error rate

- Pessimistic approach

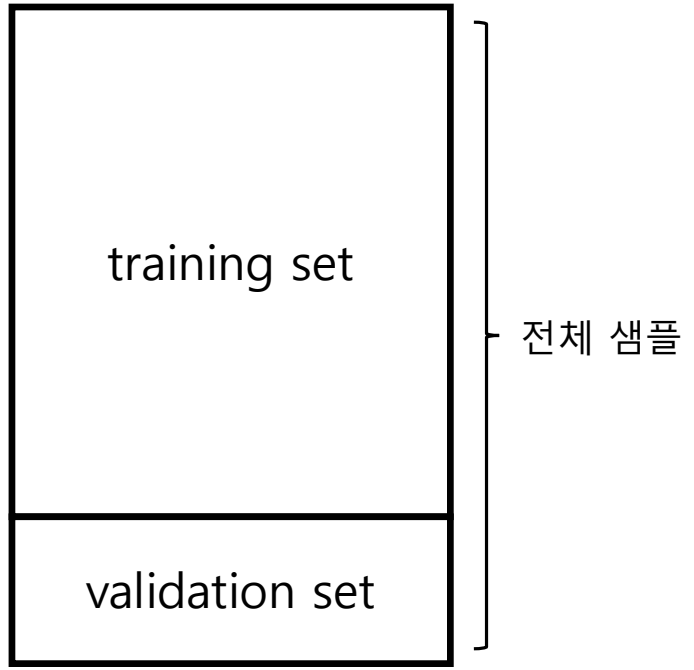
  Generalization error = $\frac{e(T)+N*penalty\ terms}{N}$

  ex) 30 leaf nodes, 10 error on 1000 instances

  generalization error = (10+30*0.5)/1000 = 2.5%

- Reduced error pruning (REP)

  validation data set을 사용하여 측정

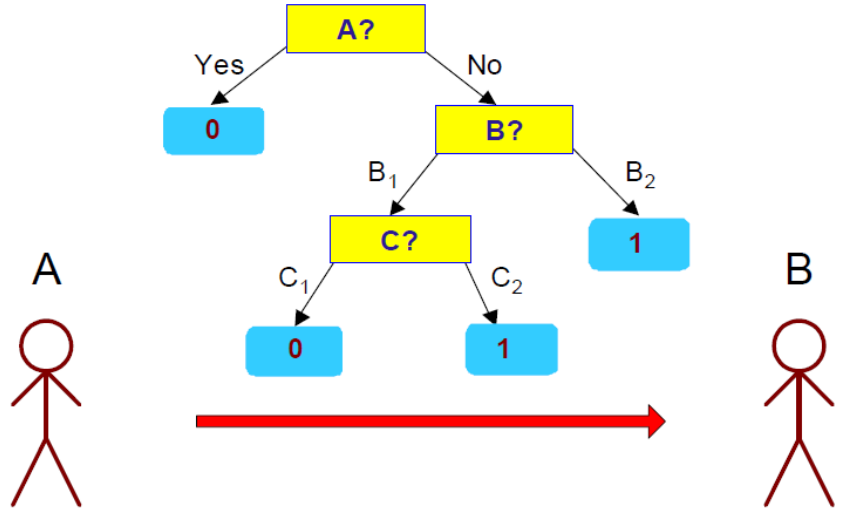


Occam's Razor

- 모델의 복잡도가 클수록 Overfitting의 확률이 증가한다.
- generalization error가 비슷한 두 모델이 주어졌을 때, 복잡도가 작은 분류 모델을 선택하는 것이 더 좋다.
- 모델을 평가할 때 모델의 복잡도를 포함해야 한다.

Minimum Description Length (MDL)

X	y
X_1	1
X_2	0
X_3	0
X_4	1
…	…
X_n	1

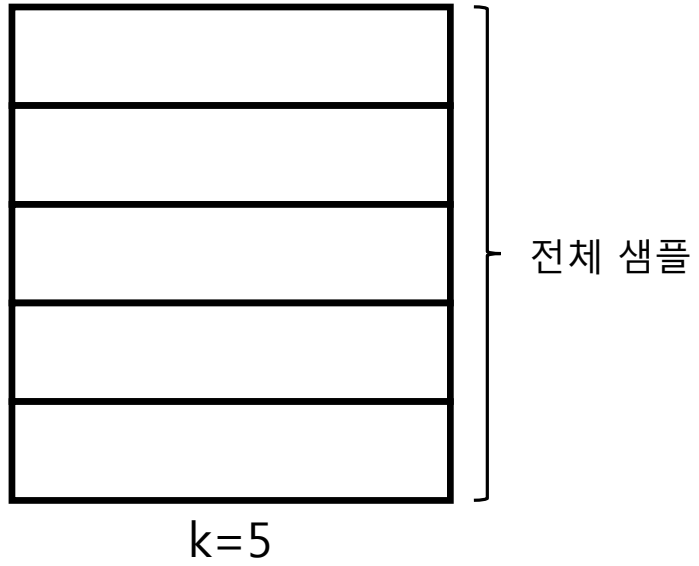


X	y
X_1	?
X_2	?
X_3	?
X_4	?
…	…
X_n	?

- $Cost(Model, Data) = Cost(Model) + Cost(Data|Model)$
  - Cost(Model): 모델을 인코딩하는 비용
  - Cost(Data|Model): 레이블이 잘못된 레코드들을 인코딩하는 비용

Methods of Estimation

- Holdout
  샘플 데이터에서 일부는 테스트
  나머지는 트레이닝
- Random subsampling
  Holdout을 반복
- Cross validation
  샘플 데이터를 k개의 파티션으로 분할
  k-fold : k-1개 training, 1개 test
- .632 bootstrap



Metrics for Performance Evaluation

- 분류 모델의 새로운 데이터에 대한 예측 능력에 집중

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Error\ rate = \frac{b+c}{a+b+c+d}$$

Limitation of Accuracy

- OpenSSL 두 개의 버전 간에 대한 함수 유사도 분석
  - 이름이 같은 함수 쌍의 수(Class 0) = 약 2,000개
  - 이름이 다른 함수 쌍의 수(Class 1) = 약 3,000,000개

- 분류 모델이 모든 함수 쌍에 대하여 Class 1로 판정한다고 하면, accuracy = 3000000/3002000 = 99.93%
  - 상황에 따라 Accuracy의 수치가 쓸모가 없다.
  - Class 0에 대해서는 탐지할 수가 없다.

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

C(i|j) : class i를 class j로 잘못 분류하는 비용

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

M_2 모델이 Accuracy가 높지만 Cost는 M_1 모델이 더 좋음

Cost-Sensitive Measures

$$Precision(p) = \frac{TP}{TP + FP} \qquad Recall(r) = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2rp}{r + p} = \frac{2TP}{2 * TP + FP + FN}$$

- Precision : positive 판별 중에 올바른 판별의 비율
- Recall : 실제의 positive 중에서 positive로 판별한 비율
- F-measure : Precision과 Recall의 조화 평균

ROC (Receiver Operating Characteristic)

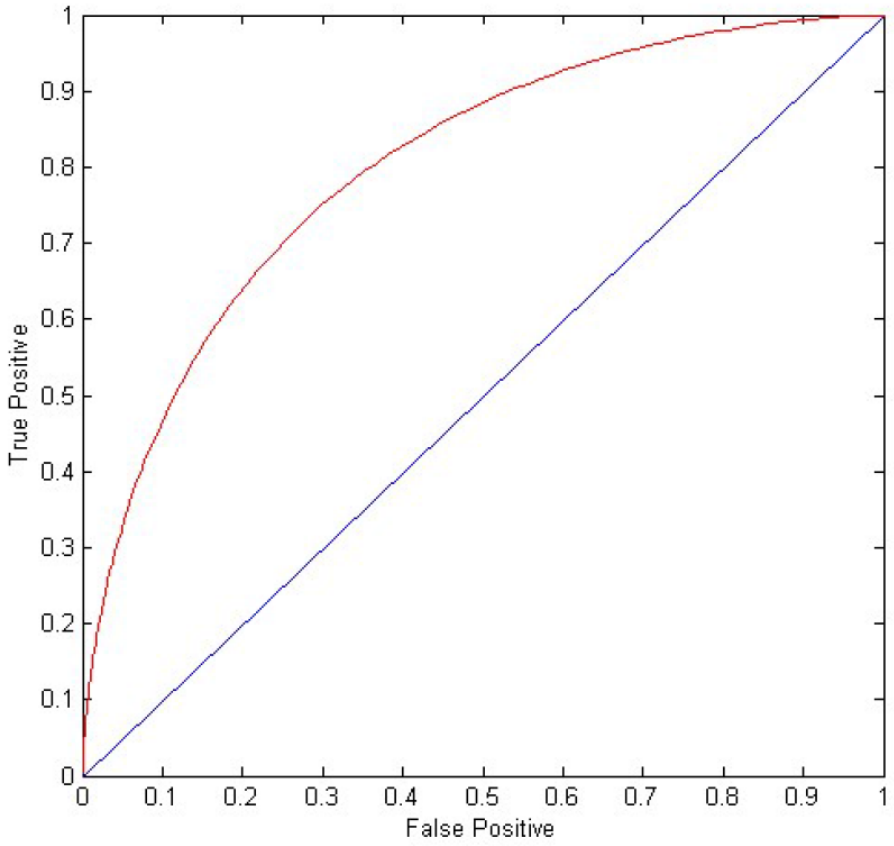
- 신호 탐지 분야에서 개발됨
- ROC curve : TPR과 FPR의 관계에 대한 커브
  - $TPR(True\ Positive\ Rate) = \frac{TP}{TP+FN}$: 민감도
  - $FPR(Flase\ Positive\ Rate) = \frac{FP}{TN+FP}$: 특이도

- 분류기의 성능이 ROC curve의 점으로 표현됨
  - 임계값이 바뀌면 점의 위치가 바뀜.

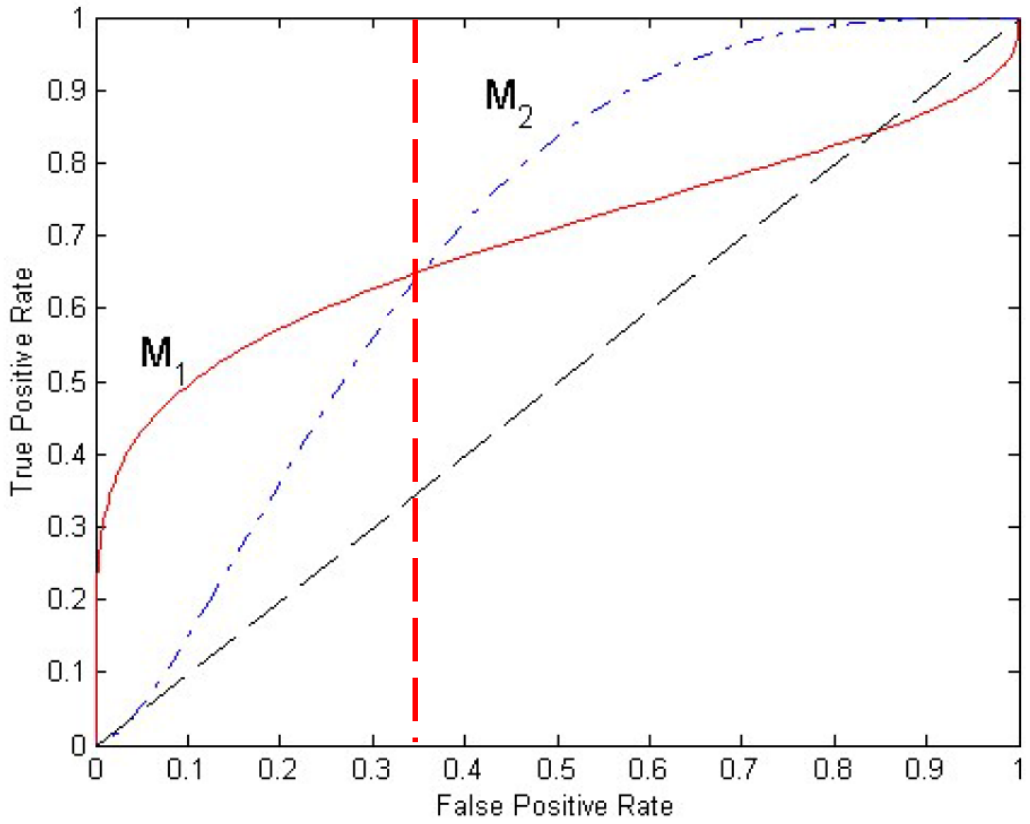
ROC Curve

- Diagonal line:
  Random Guessing

- Area Under the ROC curve (AUC)
  - Ideal : Area = 1
  - Random : Area = 0.5



Using ROC for Model Comparison



- 모델 M_1은 낮은 FPR에 대해서 좋은 성능을 나타냄
- 모델 M_2는 높은 FPR에 대해서 좋은 성능을 나타냄

Class Imbalance Problem

- precision, recall 등을 사용
- Cost-sensitive learning
- Sampling-based approaches
  - Undersampling
  - Oversampling
  - Generating artificial positive examples

Rule-Based Classifier

Rule-Based Classifier

- Rule: A -> B : A이면 B이다.
  - A : 조건
  - B : 결과

ex) (Status=Single) -> No

- Rule Coverage

  전체 10개 인스턴스 중 4개가 해당 : 40%

- Rule Accuracy

  Single로 분류되는 것 중 No가 50%

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Characteristics of Rule-Based Classifier

- Mutually exclusive rules
  - 하나의 샘플은 하나의 rule로만 커버됨.

- Exhaustive rules
  - 샘플은 적어도 하나의 rule에 매칭 되어야 함.

Characteristics of Rule-Based Classifier

- Mutually exclusive rules을 위반하는 경우
  - Ordered rule set
  - Unordered rule set – 더 많이 분류하는 rule에 맞춤.

- Exhaustive rules을 위반하는 경우
  - default class 사용

Ordered Rule Set

- Rule 우선순위에 의해 먼저 매칭된 룰 선택
- Rule ordering 방법
  - rule-ordering
  - class-ordering

Rule-based Ordering

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

Class-based Ordering

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Married}) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

Building Classification Rules

- Direct Method

  data로부터 직접 rule을 뽑아냄.

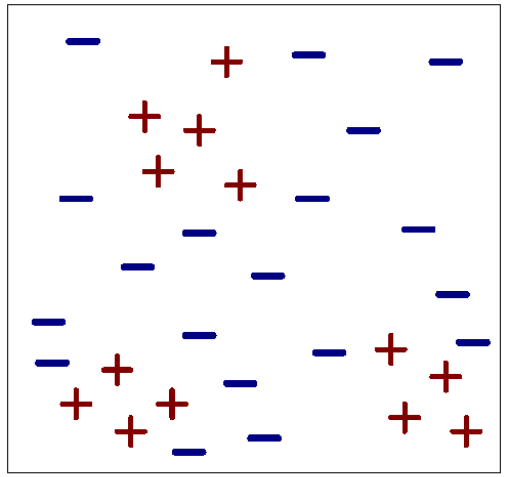
  RIPPER, CN2, Holte's 1R : Direct Method 방법을 사용

- Indirect Method

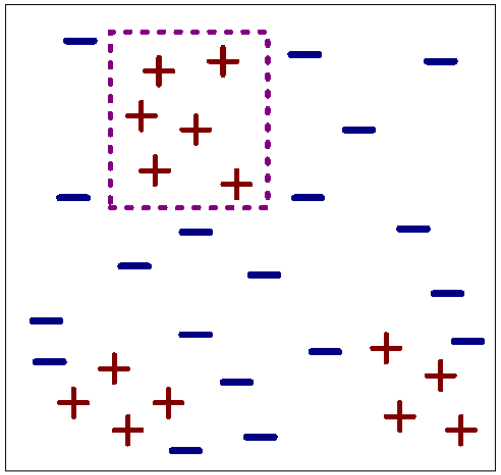
  Decision Tree 같은 분류 모델로부터 rule을 뽑아냄

  C4.5 rules가 indirect method 방법을 사용

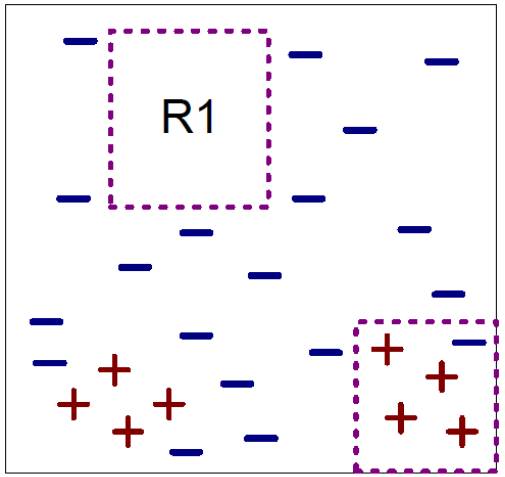
Direct Method: Sequential Covering



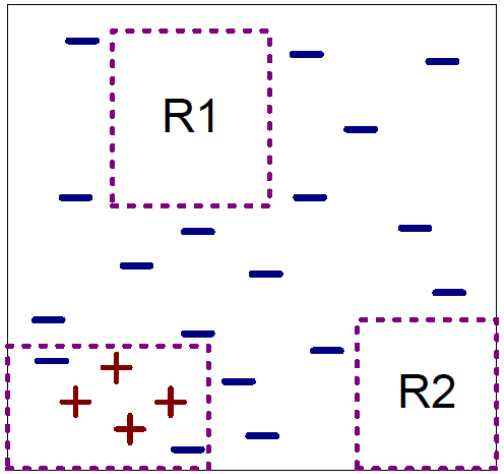
(i) Original Data



(ii) Step 1



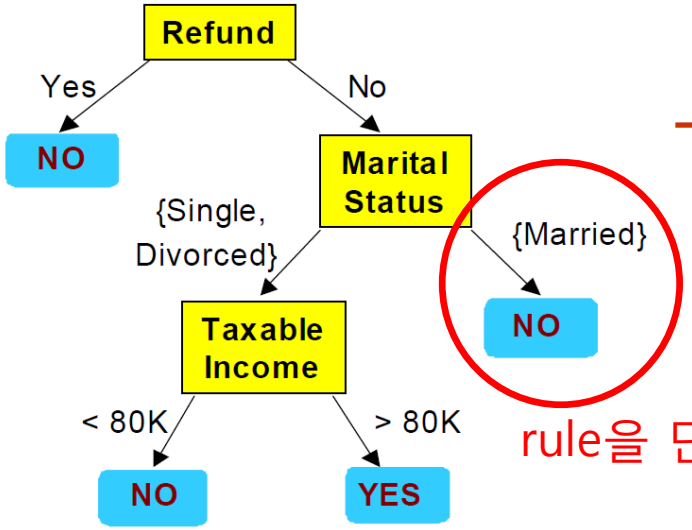
(iii) Step 2



(iv) Step 3

- empty rule로 시작
- 룰을 만들고 커버되는 샘플 삭제하는 과정 반복

Indirect Methods: From Decision Trees



Classification Rules

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(~~Refund=No~~, Marital Status={Married}) ==> No

rule을 단순화 시키는 것이 가능

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Rules are mutually exclusive and exhaustive

Advantages of Rule-Based Classifiers

- 표현력이 좋다
- 해석하기 쉽다
- 만들기 쉽다
- 새로운 인스턴스를 빠르게 분류할 수 있다
- ~~Decision tree와 성능을 비교할만 하다~~