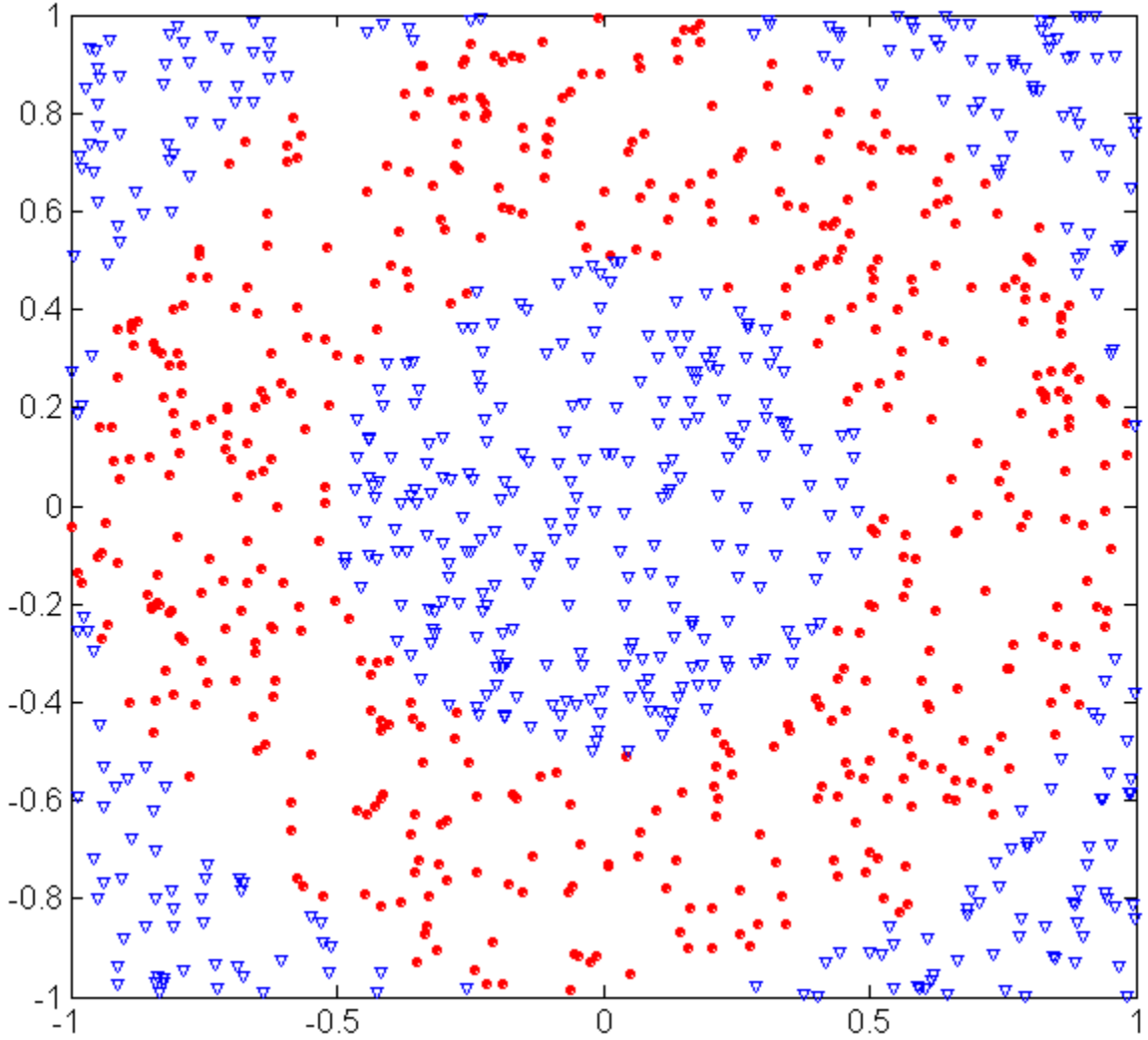
Data Science

04.01-04 의사결정나무 분류에서의 이슈

분류에서의 issue 들

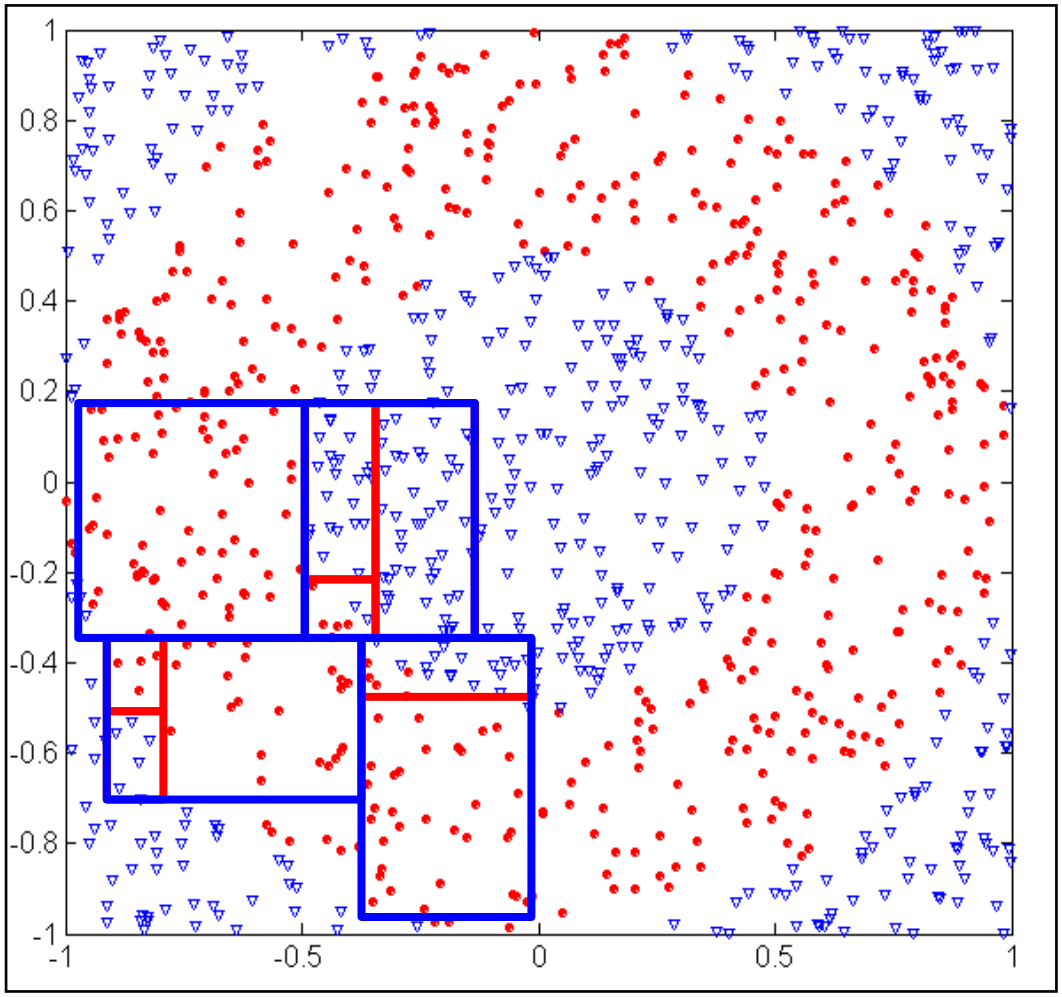
- Underfitting 과 Overfitting
- 분류의 비용(Costs of Classification)

Underfitting and Overfitting (Example)



- **Training data set.**
- **Red Class: 500**
  $0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$
- **Blue Class: 500**
  $\text{sqrt}(x_1^2+x_2^2) > 0.5$ or
  $\text{sqrt}(x_1^2+x_2^2) < 1$

Underfitting and Overfitting (Example)



• 데이터 셋을 학습 데이터로 삼아 나무모델 생성하였을 경우

• 각 잎 노드 는 직사각형의 영역을 차지함.

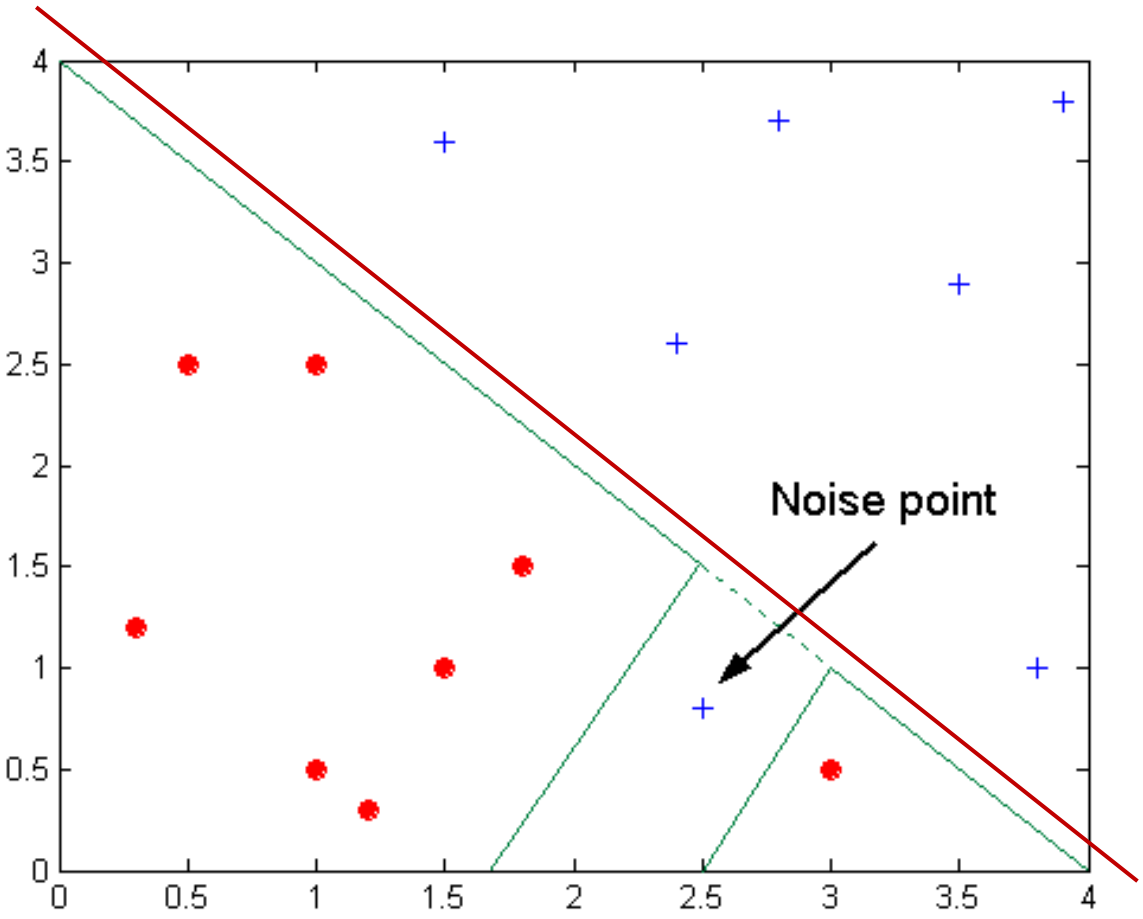
• Underfitting: 나무 크기가 적정수준보다 작은 경우 각 잎노드에 할당된 영역이 너무 커져 사례들을 정확한 클래스로 분리하기 힘들게 됨.

•

• Overfitting: 나무 크기가 적정수준보다 큰 경우 영역이 작아지게 되고 영역 안에 들어 있는 사례들의 수가 너무 적어 영역, 또는 잎노드에 대하여 정확한 클래스를 정하는 것이 힘들게 됨.

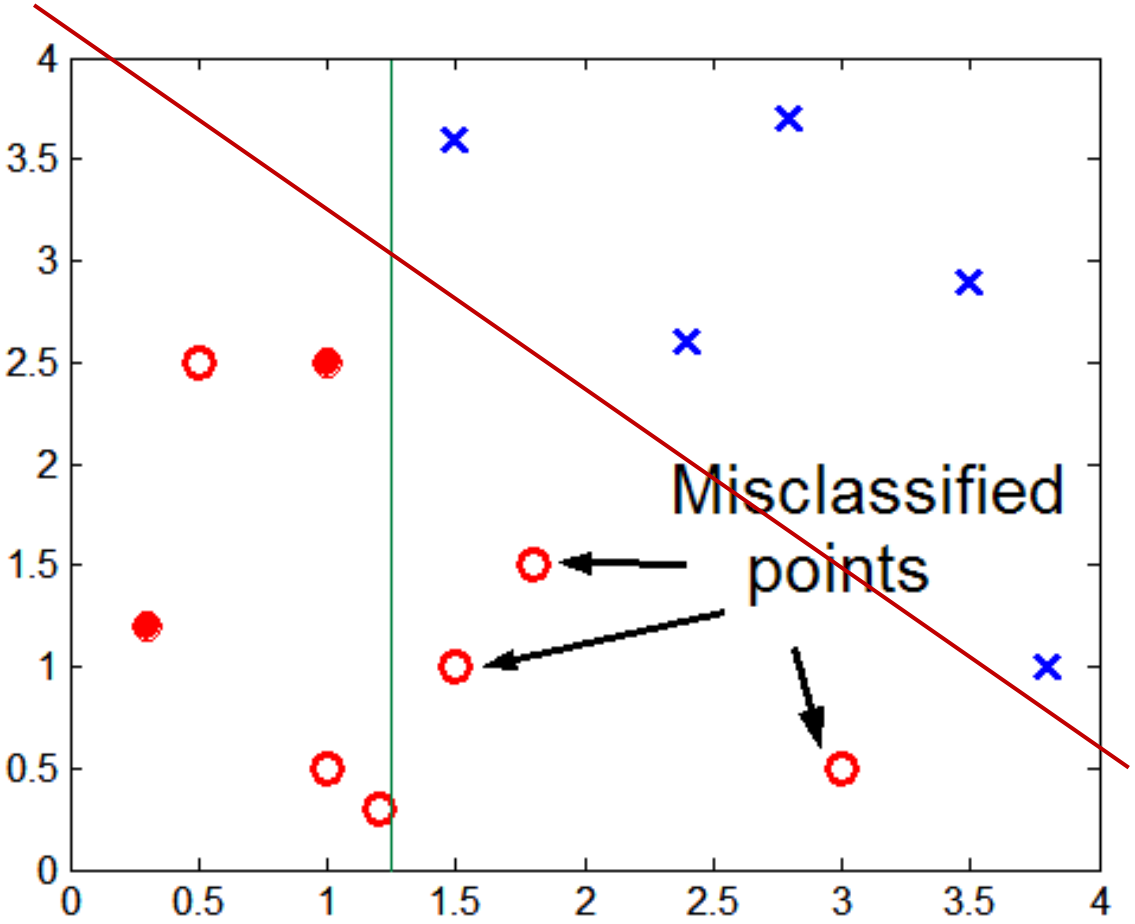
•최적의 나무 크기는?
→ 해상도의 문제

잡음에 의한 overfitting



결정경계(Decision boundary) 가 잡음 point 때문에 왜곡된 경우

부족한 사례에 의한 overfitting

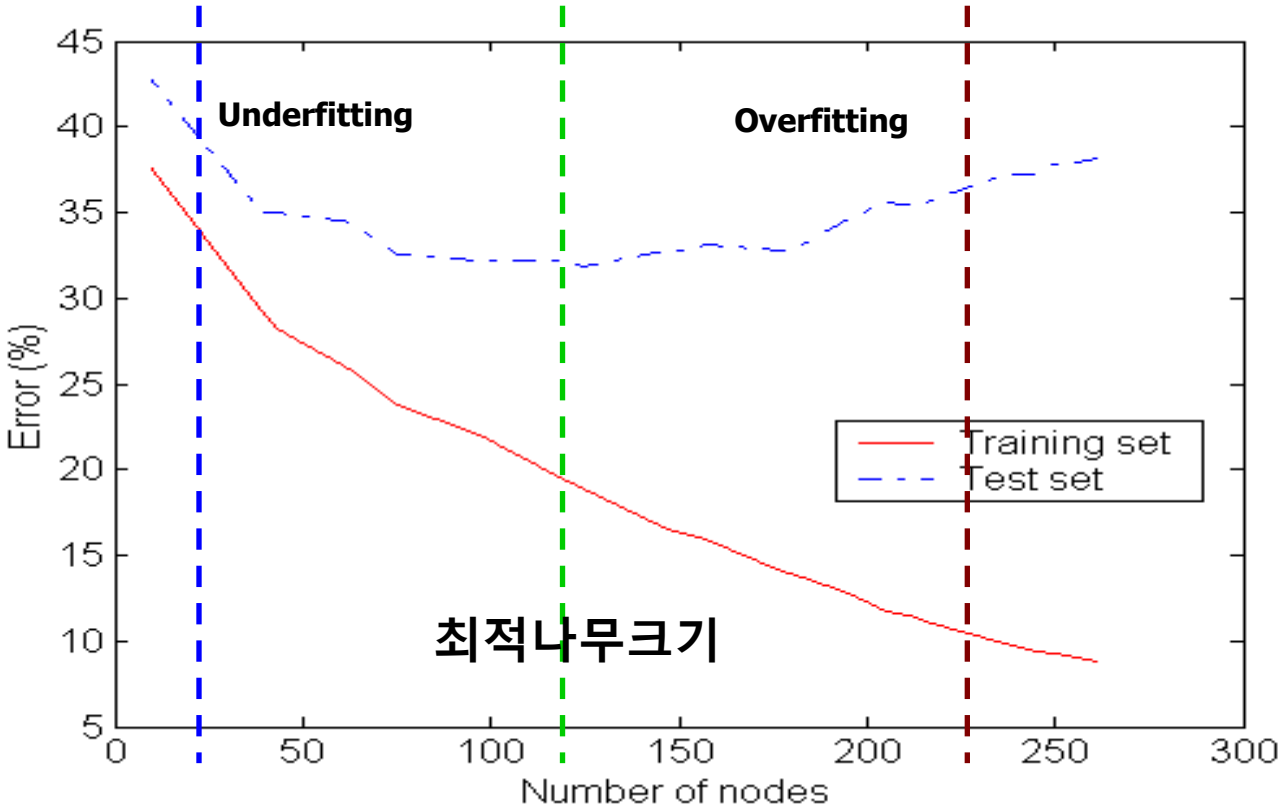


빨간 점 및 파랑 x 는 학습용 사례임.
속이 빈 점은 테스트 사례임.

학습 데이터셋에서의 데이터의 부족
→ 결정 경계 를 왜곡시킬 수 있다.

Underfitting and Overfitting

Error = 틀리게 분류한 사례 수 / 전체 평가용 사례 수



학습곡선(Learning Curve)

Underfitting: 모델이 너무 단순한 경우, training, test error 가 크다.
Overfitting: 모델이 너무 복잡한 경우, testing error 가 다시 커진다.

Notes on Overfitting

- Overfitting 은 필요한 만큼보다 더 복잡한 tree 를 만든다.

- 이 경우에는

  Training error 는 tree 의 성능을 제대로 나타내지 못한다.

  즉, 예측 성능을 제대로 나타내지 못함.

- Error 를 측정하는 새로운 방법이 필요함.

Estimating Generalization Errors

- Re-substitution (재대입) errors: error on training (Σ e(t))
- Generalization(일반화) errors: error on testing (Σ e'(t))
- generalization errors 를 예측하는 방법:
  - 낙관적 방법: e'(t) = e(t)
  - 비관적 방법:
    - 각 잎노드 t 에 대하여: e'(t) = (e(t)+0.5)
    - error 의 총합: e'(T) = e(T) + N × 0.5 (N: 잎노드의 개수)
    - 30개의 잎노드를 가진 트리 에서 1000개의 training 사례에 대하여 10개의 error 가 발생했다면

      재대입 error 율 = 10/1000 = 1%

      일반화 error 율 = (10 + 30×0.5)/1000 = 2.5%

Penalty on complexity

Occam's Razor

- 임의의 현상을 설명하는 이론이 여러 가지 있을 때 단순한 이론을 선택함. – Occam's Razor

- 예측 모델도 마찬가지
  - 복잡한 모델의 경우, 데이터에 들어 있는 error 또는 잡음에 대하여 overfitting 되어 있을 확률이 크다.
  - 예측 정확도가 비슷한 두 모델이 있을 경우 단순한 모델을 선호함.

- 따라서, model 을 평가할 때 정확도 뿐만 하니라 모델의 복잡도를 고려해야 된다.

모델 평가

- 성능 평가의 척도(Metrics for Performance Evaluation)
  - 모델의 성능을 어떻게 나타낼 것인가?
- 성능 평가의 방법(Methods for Performance Evaluation)
  - 신뢰성 있는 정확도의 예측치를 얻기 위한 방법

- 모델의 비교 방법
  - How to compare the relative performance among competing models?

성능 평가를 위한 척도

- Model 의 예측이 얼마나 정확한가에 중점.
  - Training 이나 classification 의 속도, 확장성 등 보다 정확도(Accuracy)에 중점.
- 혼동행렬(Confusion Matrix):

	예측된 클래스		
실제 클래스		Class=Yes	Class=No
	Class =Yes	a	b
	Class=No	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

관심을 두는 class 를 positive 다른 쪽을 negative 로 함

성능 평가를 위한 척도

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Type 1 error : false positive error
Type 2 error : false negative error

- 정확도의 정의:

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

정확도의 문제

- 2 class 문제에서 평가용 데이터 셋의 분포가
  - Class 0 사례수 = 9990
  - Class 1 사례수 = 10

- 만일 예측모델이 모든 입력 사례를 class 0로 예측,

  Accuracy = 9990/10000 = 99.9 %
  - 이 경우 Accuracy 는 오해의 소지가 있다. 사실 class 1 에 대한 예측은 전혀 하지 못한다.

정확도 척도

- 2 class 일 경우

  관심을 두는 class 의 example을 positive example이라 하고 다른 쪽을 negative example로 한다.

- Positive/negative class에 대한 Accuracy 의 구분
  - Sensitivity(민감도) = # of true positives / # of actual positives
  - Specificity(특이도) = # of true negatives/ # of actual negatives

정확도 척도

		PREDICTED CLASS	
		+	-
ACTUAL CLASS	+	a	b
	-	c	d

$$\text{Precision } p = \frac{a}{a+c}$$
정밀도

= Positive Predictive Value
(Medical Diagnosis, Search Engine)

$$\text{Recall } r = \frac{a}{a+b}$$
재현율

$$\text{F-measure } F = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

모델 평가

- 성능 평가의 척도(Metrics for Performance Evaluation)
  - 모델의 성능을 어떻게 나타낼 것인가?
- 성능 평가의 방법(Methods for Performance Evaluation)
  - 신뢰성 있는 정확도의 예측치를 얻기 위한 방법
- 모델의 비교 방법
  - How to compare the relative performance among competing models?

모델 평가의 방법

- 모델의 성능은 학습 알고리즘 외에 다음 요인에 의하여 영향을 받는다.
  - 클래스 분포 (Class distribution)
  - 오분류에 대한 비용 (Cost of misclassification)
  - 데이터 셋의 크기 (Size of training and test sets)

Methods of Sampling Records

- Holdout
  - Record set 중 2/3 는 training, 1/3 은 testing set 으로 무작위 추출 (Random sampling)
- Random subsampling
  - Repeated holdout
- k – fold Cross validation
  - Data 를 k 개의 partition 으로 분할
  - k - 1 개의 partition 은 training 에, 나머지 하나는 testing 에
  - Leave-one-out: k = n (= number of records) 인 경우
- Stratified sampling
  - Record 집합을 속성에 따라 복수 의 homogeneous group 으로 나누고 각 group 에서 random sampling 함.
  - oversampling vs undersampling
- Bootstrap
  - Replacement(선택된 record 를 대체)하면서 sampling 함.

* Note: training set 과 testing set 은 어떤 경우라도 중복되지 않도록 해야된다.

비용 행렬 Cost Matrix

		PREDICTED CLASS	
	C(*i*|*j*)	**Class=Yes**	**Class=No**
ACTUAL CLASS	**Class=Yes**	C(Yes|Yes) = w_1	C(No|Yes) = w_2
	Class=No	C(Yes|No) = w_3	C(No|No) = w_4

C(*i* | *j*): Class *j* 인 record 를 class *i* 로 분류 할 때의 비용

$$\text{총분류비용} = w_1 a + w_2 b + w_3 c + w_4 d$$

분류에 대한 비용(cost of classification)

500 records

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

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

Accuracy = 400/500 =80%

Cost = 3910

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

Accuracy = 450/500 = 90%

Cost = 4255

예시: 과자공장에서의 분류 비용

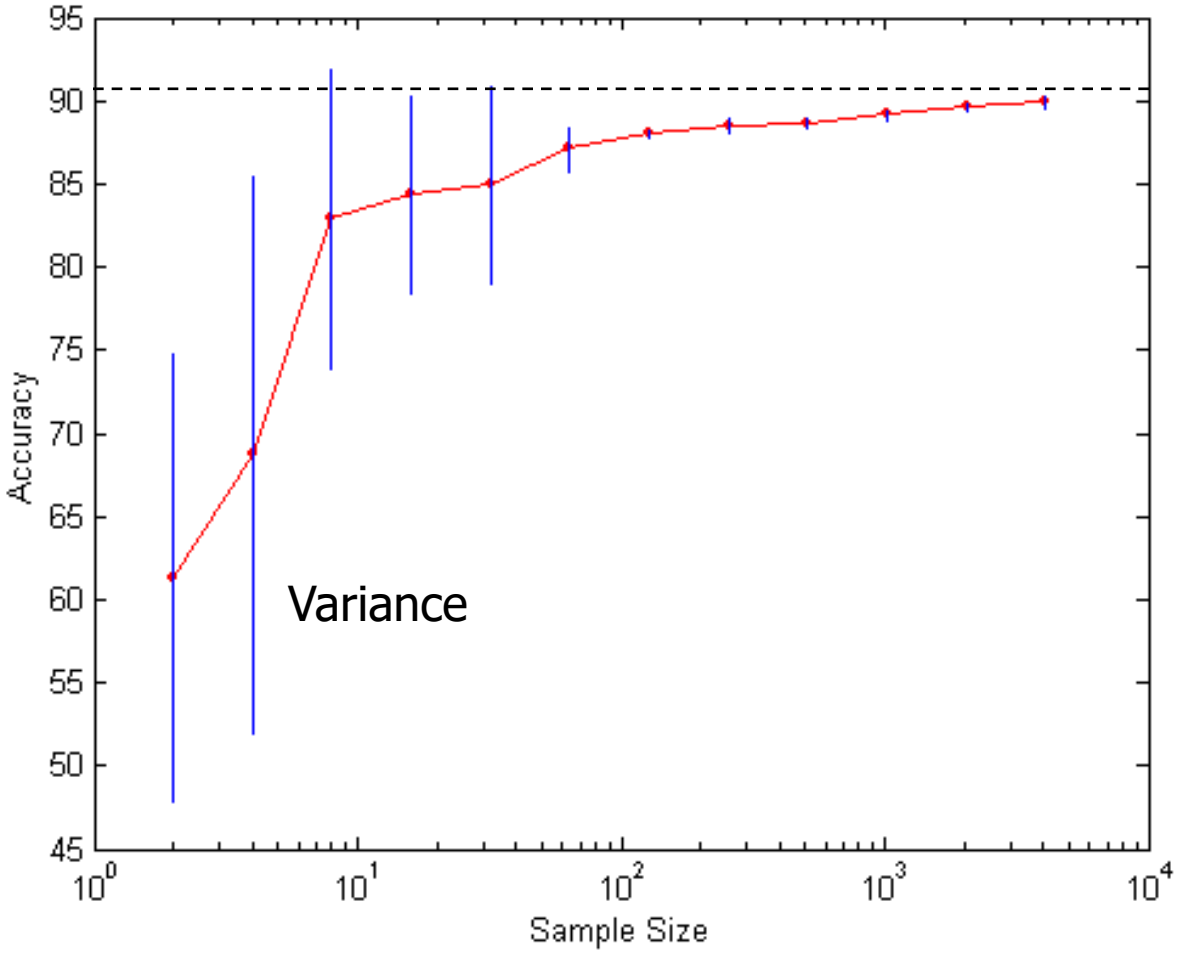
- 과자공장에서 완제품을 기계로 자동 분류하는 경우 오분류 비용은?
  - 정상품을 불량품으로 판별하는 경우
    - 정상품 가격만큼 손해
  - 불량품을 정상품으로 분류하는 경우
    - 소비자 신뢰에 문제
    - 매우 큰 오분류 비용 부여

분류에 대한 비용을 모델에 반영

- 분류에 대한 비용으로 모델을 최적화하는 방법
  - 모델 생성 알고리즘에서 처리하거나
  - 학습 데이터셋의 클래스 분포를 조정함.
- 사례의 분포(distribution)를 기반으로 하는 분류 모델에서
  - 오분류에 대한 비용을 학습데이터 셋의 클래스의 분포에 반영함.
  - 비용의 비율을 학습 데이터 셋의 클래스 비율에 적용함.
  - 예) 남/여 의 오분류 비용이 1 : 5 이고 전체 데이터에서 남/여의 클래스 비율이 90 : 10% 인 경우 학습 데이터셋에서 남 : 여 비율을 1 * 90 : 5 * 10으로 조정
- 사례의 분포를 기반으로 하는 분류 모델
  - 나무모델, 베이지안, k-nn 등
  - cf. SVM, 신경망 등은 결정경계(decision boundary)를 기반으로 함.

데이터 셋의 크기

Learning Curve(학습곡선)



- Sample size 에 따라 정확도가 변화
- a sampling schedule for creating learning curve:
  - Arithmetic sampling (Langley, et al)
  - Geometric sampling (Provost et al)

작은 sample size의 부작용:

- Bias in the estimate
- Variance of estimate

실습

- Tree model
  - Cross validation
  - 오분류 비용 적용
  - 데이터 셋 크기에 따른 정확도의 변화