기계 학습

기계 학습

Part I

1. 기계학습

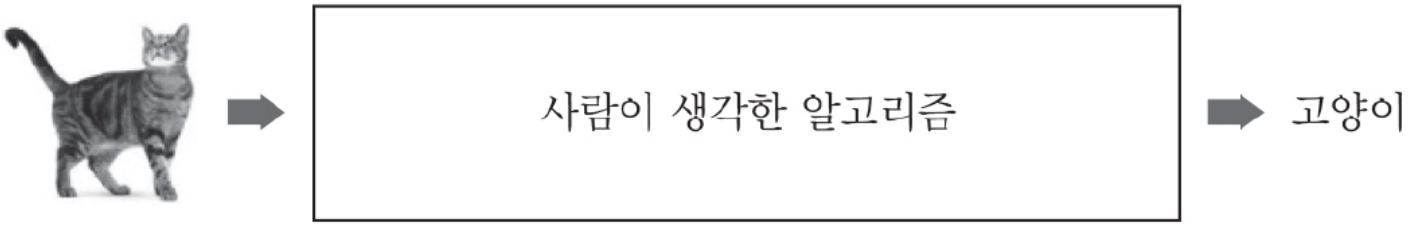
❖ **기계학습**(機械學習, machine learning)

- **경험**을 통해서 나중에 유사하거나 같은 **일**(task)를 더 **효율적**으로 처리할 수 있도록 **시스템**의 **구조**나 **파라미터**를 **바꾸는 것**
- 컴퓨터가 데이터로부터 특정 문제해결을 위한 지식을 자동으로 추출해서 사용할 수 있게 하는 기술

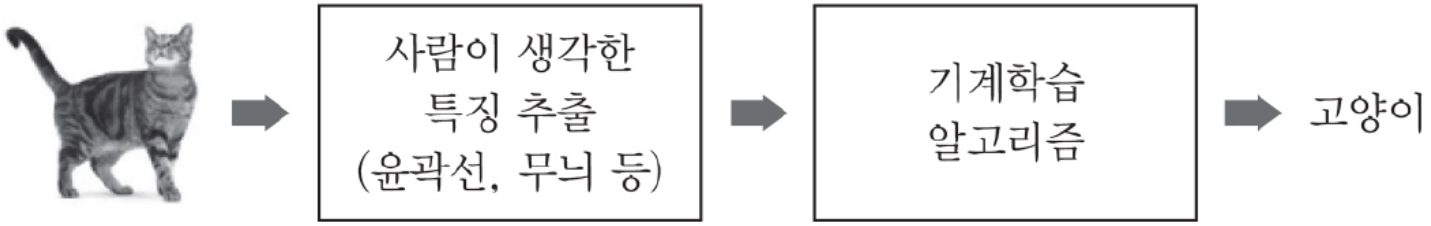
경험	일	효율(성능)
필기문자 이미지, 글자	문자 판독(인식)	정확도
사진, 얼굴영역	사진에서 얼굴영역 식별	정확도
이메일, 스팸여부	스팸 이메일 판단	정확도
풍경 사진	유사한 풍경 사진 식별	유사도
바둑 대국	바둑두는 방법	승률

기계학습

❖ 일반 프로그래밍 방식



❖ 기계 학습



기계학습

❖ PlayTennis 문제

- 어떤 사람이 테니스를 치는 날의 기상 상황을 조사한 데이터
  - **학습데이터 (training data)**

표 4.1 PlayTennis 데이터

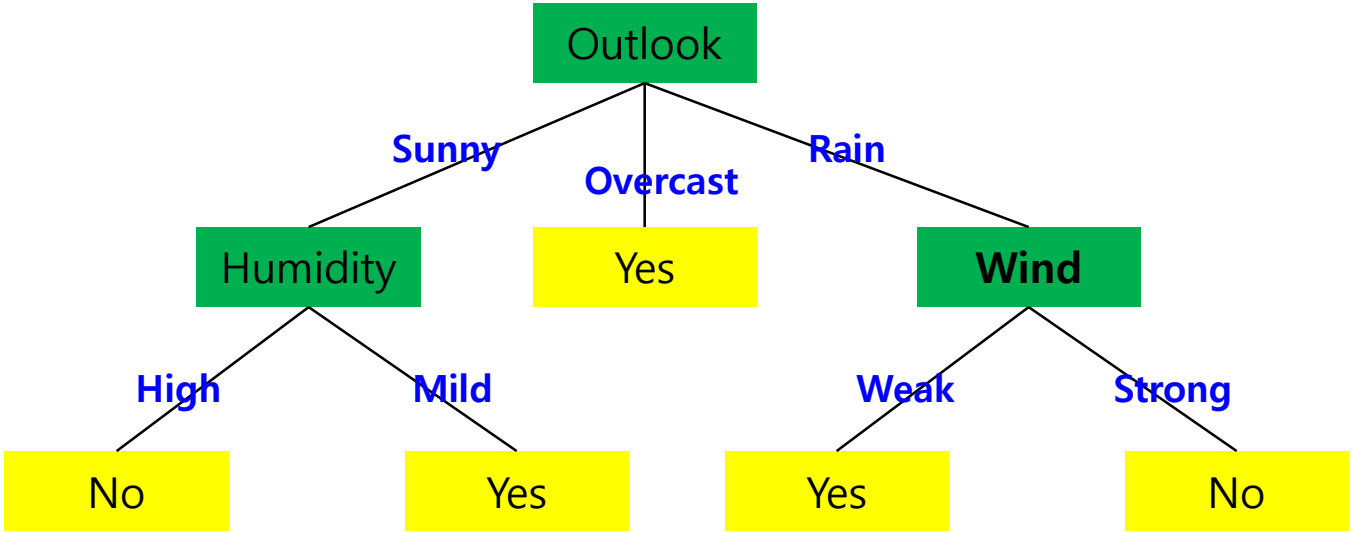
Day 날짜	Outlook 조망	Temperature 기온	Humidity 습도	Wind 바람	PlayTennis 테니스 여부
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

(출처: Machine Learning, Tom Mitchell, 1995)

- **테니스를 치는 날은?**
- **'흐리고 적당한 온도에 습도는 높고 바람이 센 날' 테니스를 칠까?**

기계학습

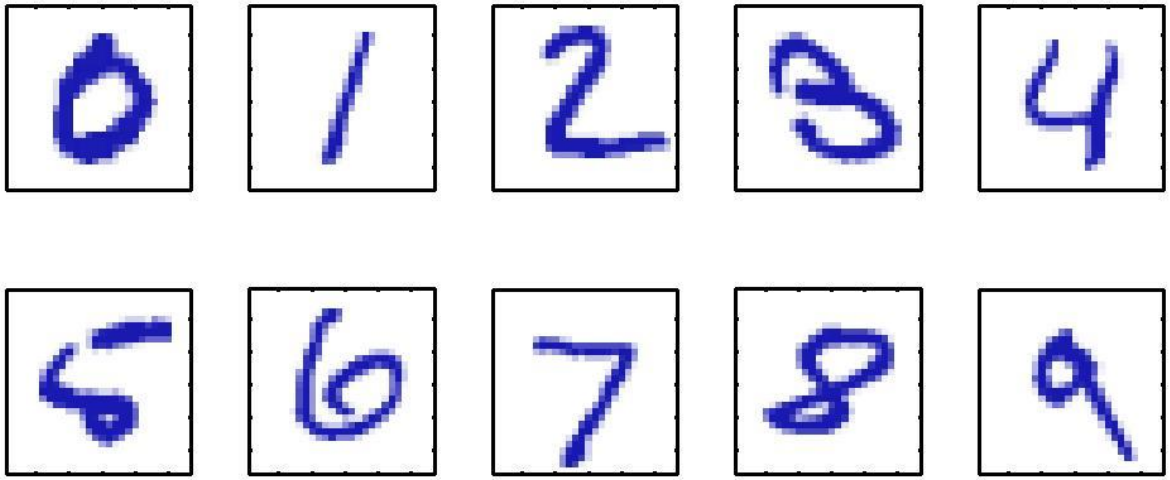
❖ **PlayTennis 문제 – cont.**



Outlook 조망	**Temperature** 기온	**Humidity** 습도	**Wind** 바람	**PlayTennis** 테니스 여부
Sunny	Hot	Mild	Weak	**?**
Rain	Hot	High	Weak	**?**

기계학습

❖ 필기문자 인식



- **직접 만든 규칙이나 휴리스틱**(heuristics)
  - 복잡
  - 불충분한 성능

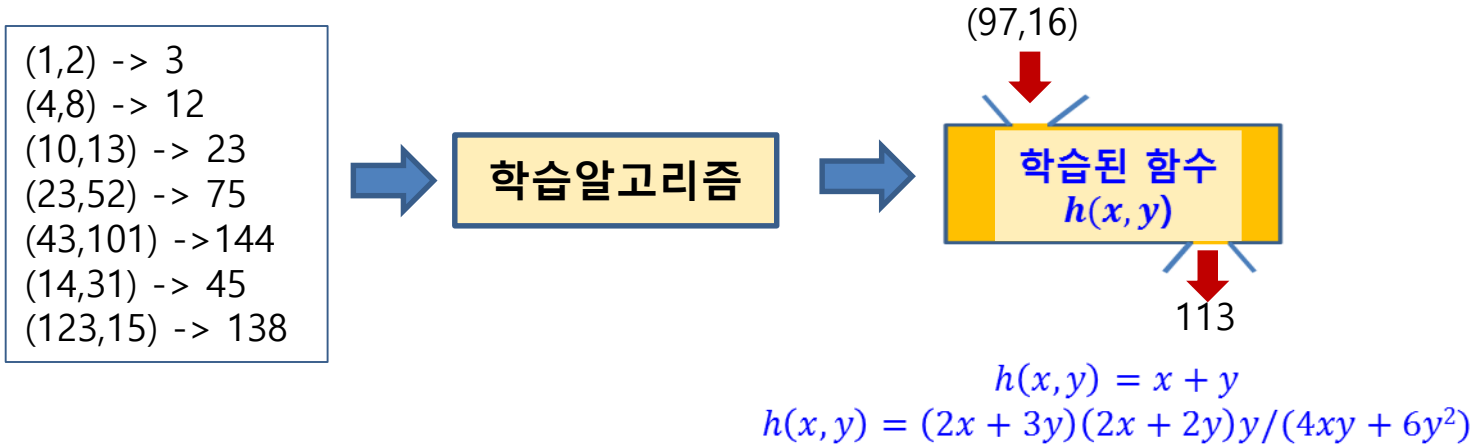
- **기계학습 방법**
  - **자동**으로 **분류 규칙**이나 **프로그램 생성**
  - 괄목할 만한 성능

Image : MNIST data

기계학습

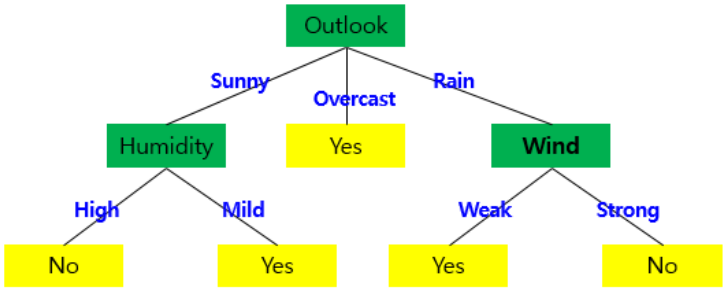
❖ **연역적 학습** (deductive learning)
- 연역적 추론(deductive inference)을 통한 학습

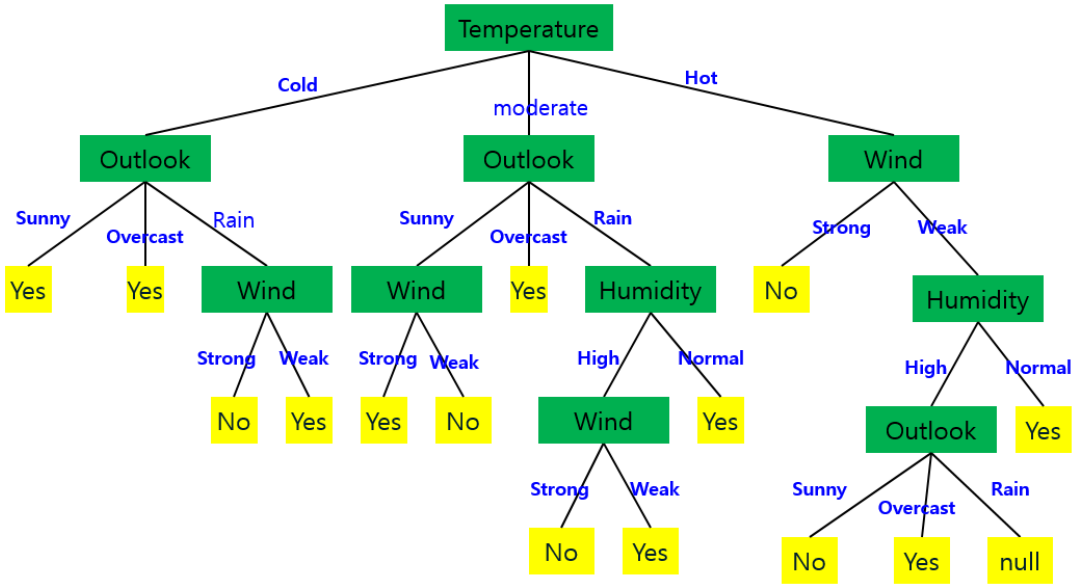
❖ **귀납적 학습** (inductive learning)
- **사례들**(examples)을 **일반화(generalization)**하여 **패턴**(pattern) 또는 **모델**(model)을 추출하는 것
- 일반적인 기계학습의 대상
- 학습 데이터를 **잘 설명**할 수 있는 **패턴**을 찾는 것
  - **오컴의 면도날**(Occam's razor)
    - 가능하면 학습 결과를 간단한 형태로 표현하는 것이 좋다



기계학습

- **오컴의 면도날**(Occam's razor) **원리**에 따른 선택





2. 기계학습의 종류

- **지도학습**(supervised learning)
  - 입력(**문제**)-출력(**답**)의 데이터들로 부터 새로운 입력에 대한 출력을 결정할 수 있는 패턴 추출

- **비지도학습**(unsupervised learning, **자율학습**)
  - **출력**에 대한 정보가 **없는 데이터**로 부터 필요한 패턴 추출

- **반지도학습**(semisupervised learning)
  - **일부 학습 데이터**만 **출력값**이 주어진 상태에서 일반화한 패턴 추출

- **강화학습**(reinforcement learning)
  - 출력에 대한 정확한 정보를 제공하지는 않지만, **평가정보**(reward)는 주어지는 문제에 대해 **각 상태**에서의 **행동**(action)을 결정

3. 기계학습 대상 문제

3.1 분류

- 과적합 학습의 문제
- 학습 데이터가 적은 경우의 성능평가
- 불균형 데이터 문제
- 이진 분류기의 성능 평가

3.2 회귀

3.3 군집화

3.4 밀도 추정

3.5 차원축소

3.6 이상치 탐지

3.7 반지도 학습

3.1 지도학습 – 분류

❖ **지도학습**(supervised learning)

- 주어진 (입력, 출력) 에 대한 데이터 이용 : 학습(training) 데이터
  - $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
- 새로운 입력이 있을 때 결과를 결정할 수 있도록 하는 방법 찾아내는 것
  - $y = \boldsymbol{f}(\boldsymbol{x})$

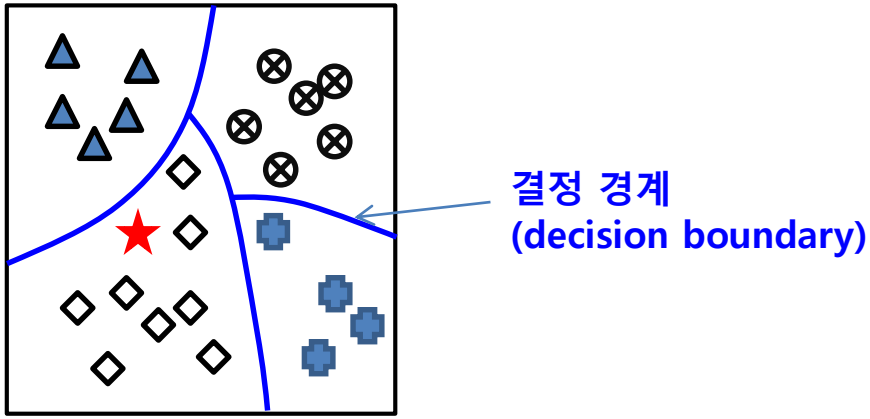
- **분류 (classification)**
  - **출력**이 정해진 **부류**(class, category) 중의 하나로 결정

- **회귀 분석(regression)**
  - **출력**이 연속인 **영역**(continuous domain)의 값 결정

분류

❖ **분류(classification)**

- 데이터들을 정해진 몇 개의 부류(class)로 대응시키는 문제



- **분류 문제의 학습**
  - **학습 데이터**를 잘 **분류**할 수 있는 **함수**를 찾는 것
  - 함수의 형태는 **수학적 함수**일 수도 있고, **규칙**일 수도 있음

- **분류기(classifier)**
  - 학습된 함수를 이용하여 데이터를 분류하는 프로그램

분류

❖ **분류기 학습 알고리즘**

- **결정트리(decision tree) 알고리즘**
- **K-근접이웃 (K-nearest neighbor, KNN) 알고리즘**
- **다층 퍼셉트론 신경망**
- **딥러닝(deep learning) 알고리즘**
- **서포트 벡터 머신(Support Vector Machine, SVM)**
- **에이다부스트(AdaBoost)**
- **랜덤 포리스트(random forest)**
- **확률 그래프 모델** (probabilistic graphical model)

분류

❖ **이상적인 분류기**

- **학습에 사용되지 않은 데이터**에 대해서 분류를 잘 하는 것
- **일반화**(generalization) **능력**이 좋은 것

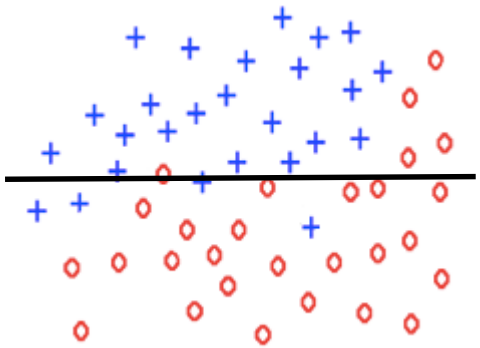
❖ **데이터의 구분**

- **학습 데이터**(training data)
  - 분류기(classifier)를 학습하는데 사용하는 데이터 집합
  - 학습 데이터가 많을 수록 유리
- **테스트 데이터**(test data)
  - 학습된 모델의 성능을 평가하는데 사용하는 데이터 집합
  - 학습에 사용되지 않은 데이터이어야 함
- **검증 데이터**(validation data)
  - 학습 과정에서 학습을 중단할 시점을 결정하기 위해 사용하는 데이터 집합

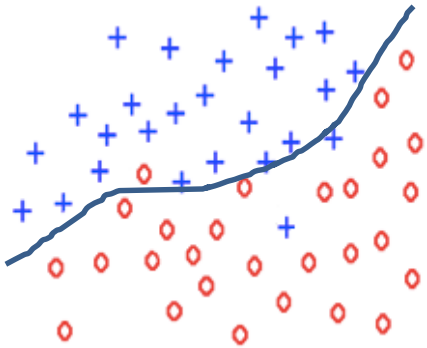
분류

❖ **과적합**(overfitting)과 **부적합**(underfitting)

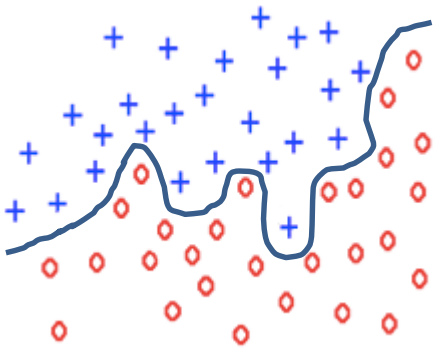
- **과적합**
  - **학습 데이터**에 대해서 **지나치게 잘 학습**된 상태
  - 데이터는 **오류**나 **잡음**을 포함할 개연성이 크기 때문에, 학습 데이터에 대해 매우 높은 성능을 보이더라도 **학습되지 않은 데이터**에 대해 좋지 않은 **성능**을 보일 수 있음
- **부적합**
  - 학습 데이터를 충분히 학습하지 않은 상태



부적합(underfitting)



정적합(good fitting)

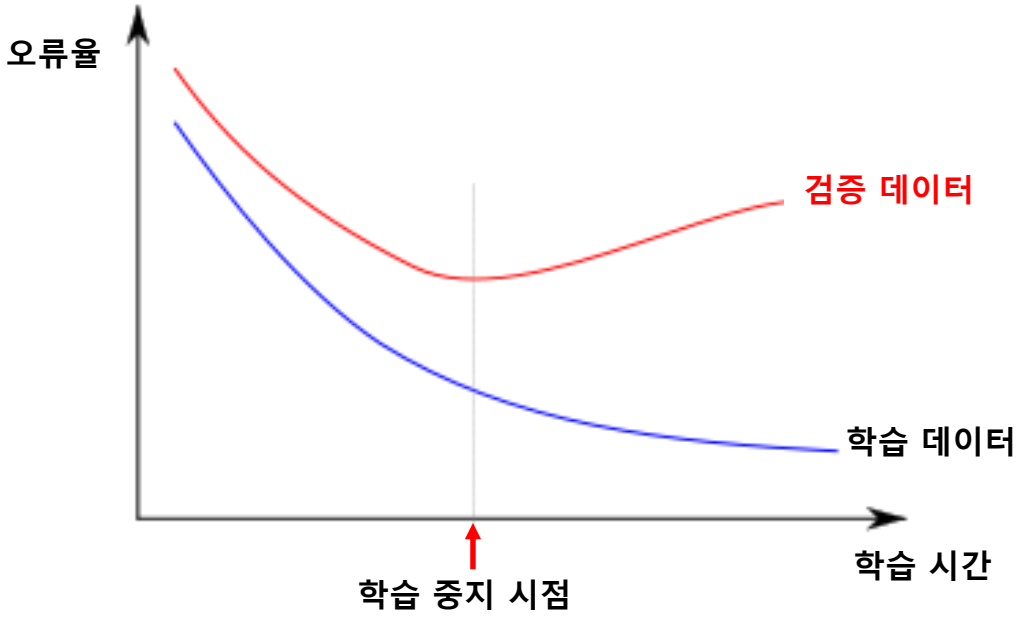


과적합(overfitting)

분류

❖ 과적합 회피 방법

- **학습데이터**에 대한 **성능**
  - **학습**을 **진행**할 수록 오류 **개선** 경향
  - 지나치게 학습이 진행되면 과적합 발생
- 학습과정에서 별도의 **검증 데이터**(validation data)에 대한 성능 평가
  - 검증 데이터에 대한 오류가 감소하다가 증가하는 시점에 학습 중단



분류

❖ 분류기의 성능 평가

- **정확도** (accuracy)
  - 얼마나 정확하게 분류하는가
  - **정확도** = (**옳게 분류한 데이터 개수**)/(**전체 데이터 개수**)
  - **테스트 데이터**에 대한 정확도를 분류기의 정확도로 사용
- 정확도가 높은 분류기를 학습하기 위해서는 **많은 학습데이터**를 사용하는 것이 유리
- **학습데이터**와 **테스트 데이터**는 **겹치게 않도록** 해야 함

분류

❖ 데이터 부족한 경우 성능평가

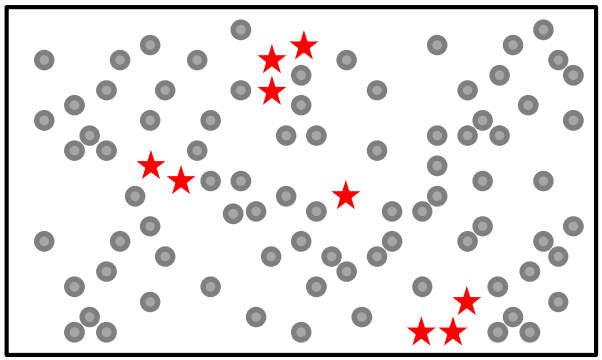
- 별도로 테스트 데이터를 확보하면 비효율적
- 가능하면 많은 데이터를 학습에 사용하면서, 성능 평가하는 방법 필요

- **K-겹 교차검증(k-fold cross-validation)** 사용
  - 전체 데이터를 **k 등분**
  - **각 등분**을 한번씩 **테스트 데이터**로 사용하여, 성능 평가를 하고 **평균값** 선택

분류

❖ 불균형 데이터(imbalanced data) 문제

- **특정 부류**에 속하는 **학습 데이터**의 **개수**가 다른 부류에 비하여 **지나치게 많은 경우**
- **정확도**에 의한 성능 평가는 **무의미**할 수 있음
  - 예. A 부류의 데이터가 전체의 99%인 경우, 분류기의 출력을 항상 A 부류로 하더라도 정확도는 99%가 됨.



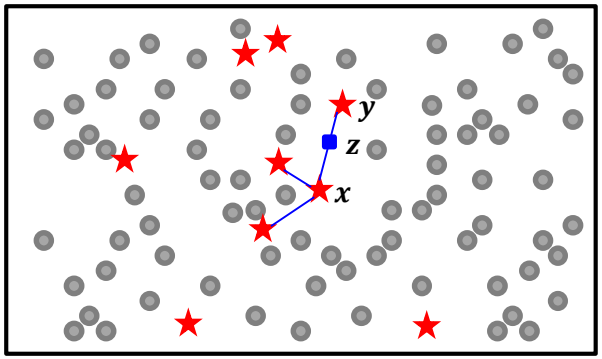
- 대응방안
  - 가중치를 고려한 정확도 척도 사용
  - 많은 학습데이터를 갖는 부류에서 재표본추출(re-sampling)
  - 적은 학습데이터를 갖는 부류에 대해서 인공적인 데이터 생성

분류

❖ 불균형 데이터(imbalanced data) 문제 – cont.

- **SMOTE**(Synthetic Minority Over-sampling Technique) 알고리즘
  - 빈도가 낮은 부류의 학습 데이터를 인공적으로 만들어 내는 방법

1. 임의로 닞은 빈도 부류의 학습 데이터 $\boldsymbol{x}$ 선택
2. $\boldsymbol{x}$의 k-근접이웃(k-nearest neighbor, KNN)인 같은 부류의 데이터 선택
3. k-근접이웃 중에 무작위로 하나 y를 선택
4. x와 y를 연결하는 직선 상의 무작위 위치에 새로운 데이터 생성



분류

❖ 이진 분류기의 성능 평가

- **이진 분류기**(binary classifier)
  - 두 개의 부류만을 갖는 데이터에 대한 분류기

표 4.2 이진 분류기의 혼동행렬

		예 측	
		양성	음성
실제	양성	진양성(True Positive) TP	위음성(False Negative) FN
	음성	위양성(False Positive) FP	진음성(True Negative) TN

  - 민감도(sensitivity)/재현율(recall)/진양성율(true positive rate)

$$\text{민감도} = \frac{TP}{TP+FN}$$

  - 특이도(specificity)/진음성율(true negative rate)

$$\text{특이도} = \frac{TN}{FP+TN}$$

분류

❖ 이진 분류기의 성능 평가 – cont.

표 4.2 이진 분류기의 혼동행렬

		예 측	
		양성	음성
실제	양성	진양성(True Positive) TP	위음성(False Negative) FN
	음성	위양성(False Positive) FP	진음성(True Negative) TN

- 정밀도(precision)

$$\text{정밀도} = \frac{TP}{TP+FP}$$

- 음성 예측도

$$\text{음성 예측도} = \frac{TN}{TN+FN}$$

- 위양성율

$$\text{위양성율} = \frac{FP}{FP+TN} = 1 - \text{특이도}$$

- 위발견율

$$\text{위발견율} = \frac{FP}{TP+FP} = 1 - \text{정밀도}$$

- 정확도

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

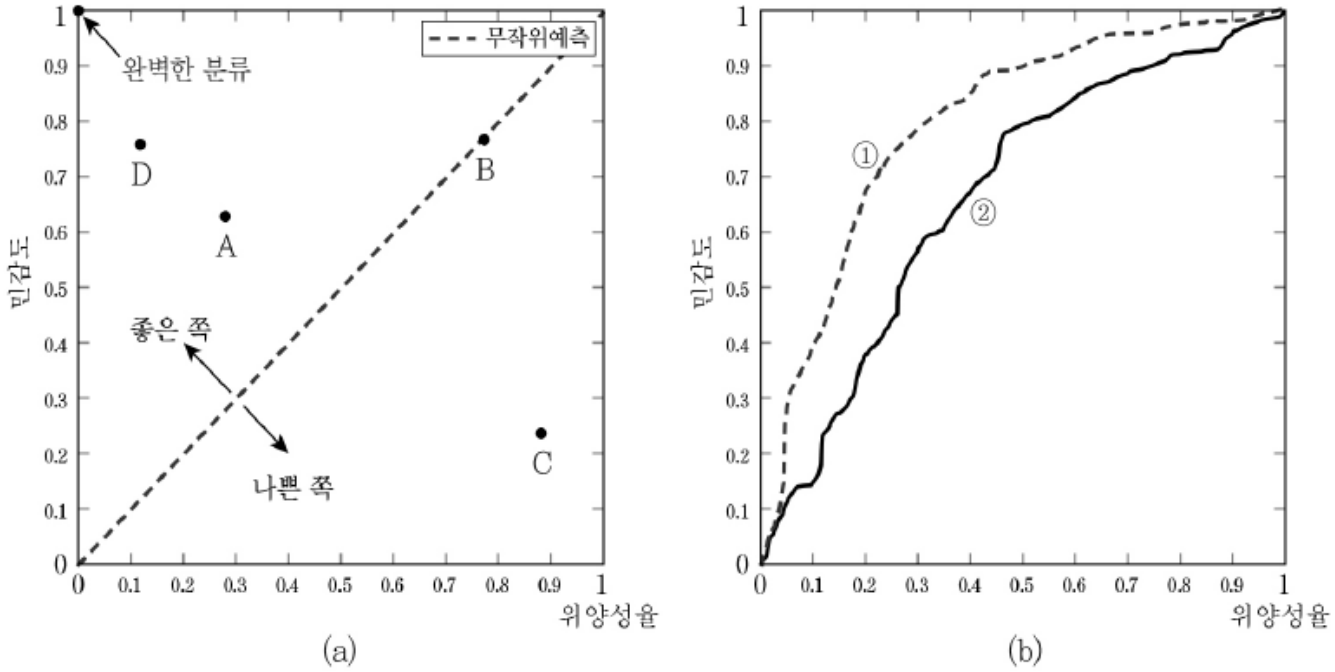
- F1 측도

$$F1 = 2\frac{(\text{정밀도}) \cdot (\text{재현율})}{(\text{정밀도}) + (\text{재현율})}$$

분류

❖ **이진 분류기의 성능 평가 – cont.**

- **ROC 곡선**
  - 부류 판정 임계값에 따른 (위양성율, 민감도) 그래프
- **AUC**(Area Under the Curve)
  - ROC 곡선에서 곡선 아래 부분의 면적
  - 클 수록 바람직

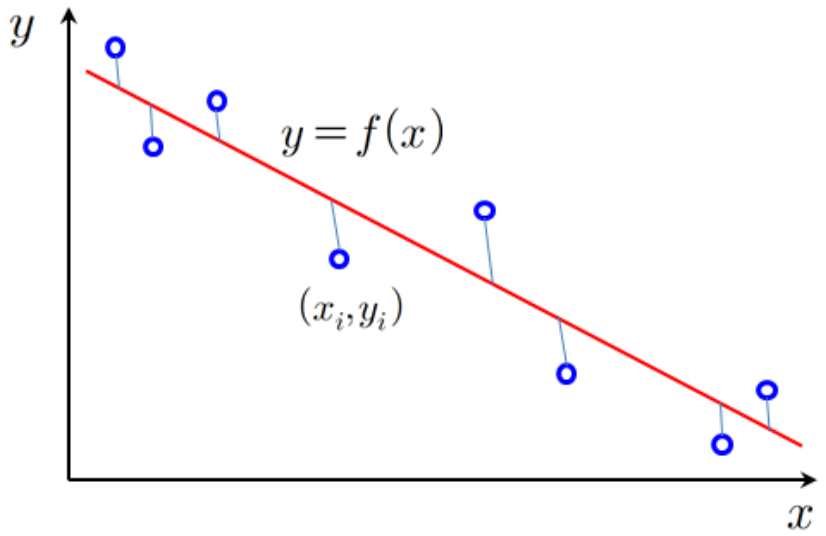


3.2 지도학습 - 회귀

❖ **회귀** (regression)

- **학습 데이터에 부합**되는 **출력**값이 **실수**인 함수를 찾는 문제

$$f^*(x) = \arg\min_f \sum_{i=1}^{n} (\boldsymbol{y_i - f(x_i)})^2$$



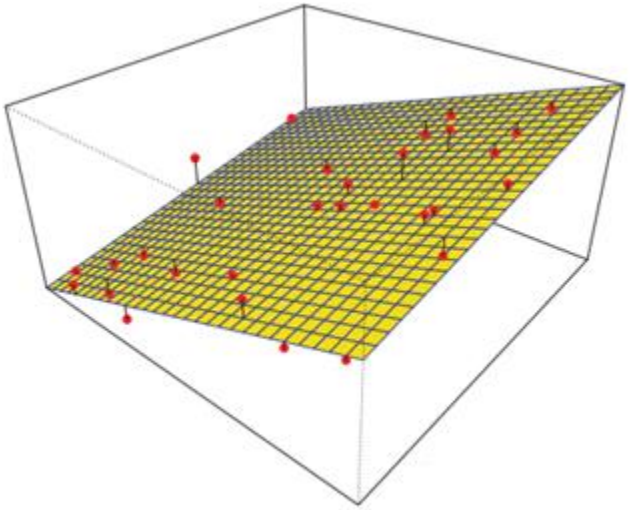
회귀

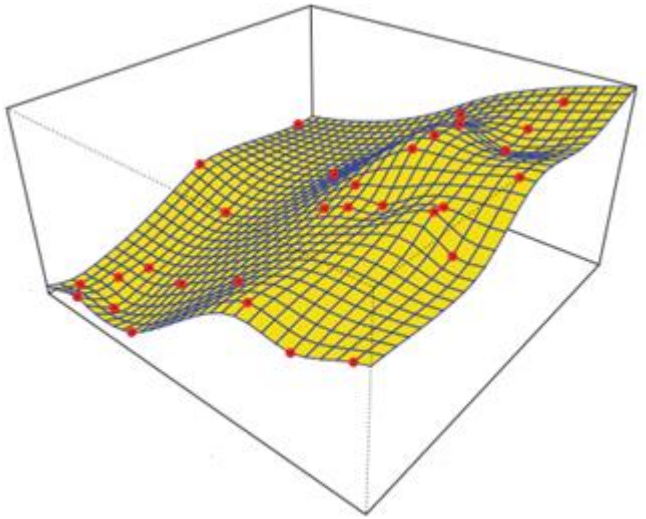
❖ **회귀** (regression) – cont.

- **성능**
  - **오차** : 예측값과 실제값의 차이
    - 테스트 데이터들에 대한 (예측값 – 실제값)^{2}의 평균 또는 평균의 제곱근

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

  - **모델**의 종류(**함수의 종류**)에 영향을 받음

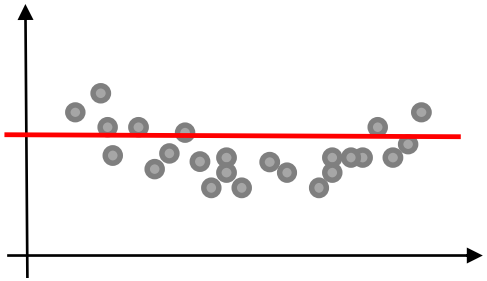




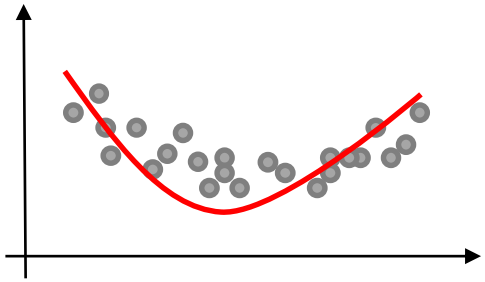
회귀

❖ **회귀**의 **과적합**(overfitting)과 **부적합**(underfitting)

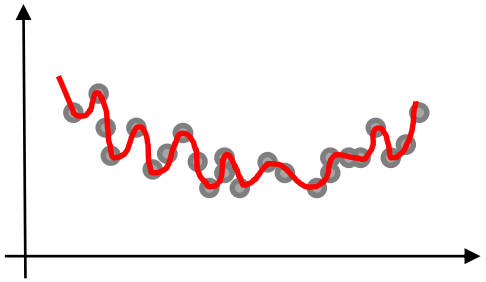
- **과적합**
  - 지나치게 복잡한 모델(함수) 사용
- **부적합**
  - 지나치게 단순한 모델(함수) 사용



부적합(underfitting)



정적합(good fitting)



과적합(overfitting)

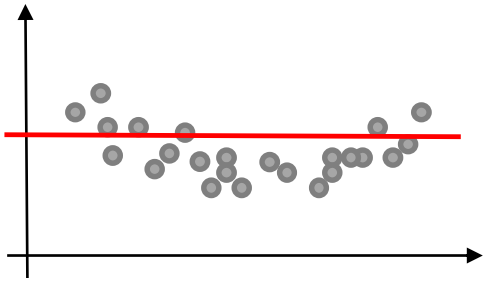
회귀

❖ **회귀**의 **과적합**(overfitting) 대응 방법

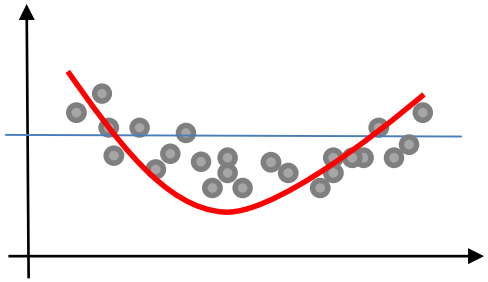
- **모델의 복잡도(model complexity)를 성능 평가에 반영**

목적함수 = 오차의 합 + (가중치)*(모델 복잡도)

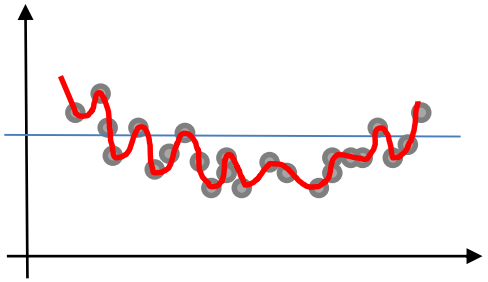
벌점(penalty) 항



부적합(underfitting)



정적합(good fitting)



과적합(overfitting)

회귀

❖ 로지스틱 회귀 (logistic regression)

- 학습 데이터 : $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, $y_1 \in \{0, 1\}$
- 로지스틱 함수를 이용하여 함수 근사

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^\top \boldsymbol{x}}}$$

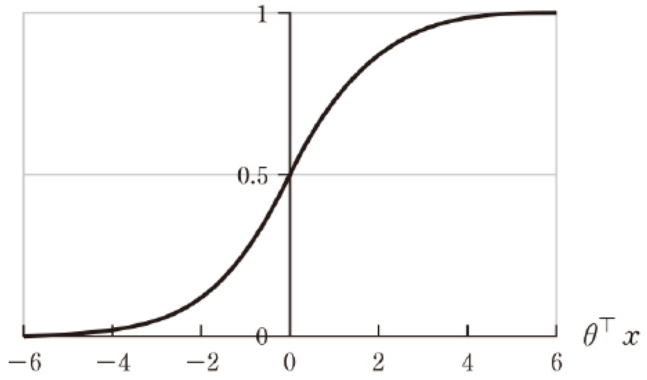


그림 4.13 로지스틱 함수.

- **학습시 목적 함수**

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}(y_i \log f(\boldsymbol{x}_i) + (1 - y_i)\log(1 - f(\boldsymbol{x}_i)))$$

  - **경사 하강법 사용 학습**

3.3 비지도학습

❖ **비지도학습**(unsupervised learning)

- **결과정보**가 **없는 데이터**들에 대해서 특정 **패턴**을 찾는 것
  - 데이터에 **잠재한 구조**(structure), **계층구조**(hierarchy) 를 찾아내는 것
  - **숨겨진 사용자 집단**(hidden user group)을 찾는 것
  - **문서들**을 **주제**에 따라 구조화하는 것
  - 로그(log) 정보를 사용하여 **사용패턴**(usage pattern)을 찾아내는 것

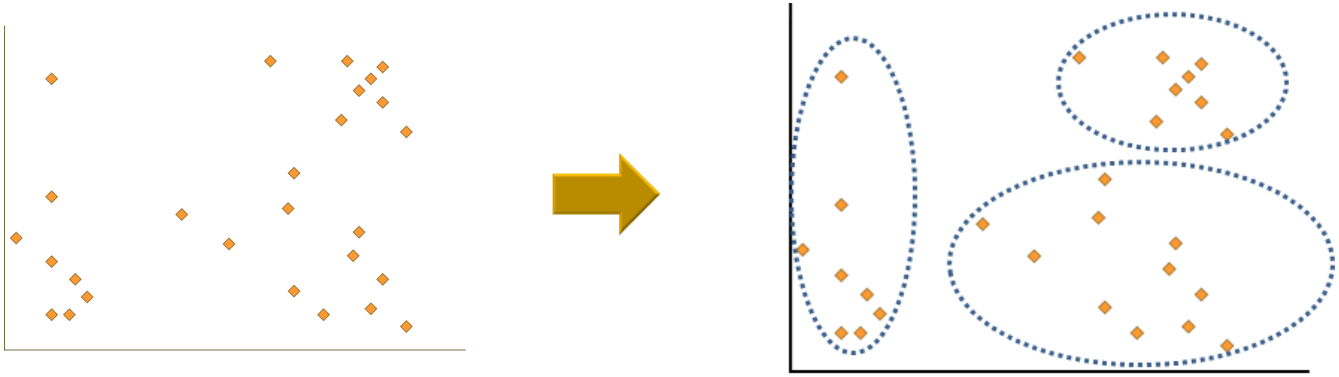
- **비지도 학습의 대상**
  - 군집화(clustering)
  - 밀도추정(density estimation)
  - 차원축소(dimensionality reduction)

http://www.youtube.com/watch?v=rhallml-juk

군집화

❖ **군집화(clustering)**

- **유사성**에 따라 데이터를 분할하는 것



영상 분할(segmentation)

image : Pedro Felzenswalb

군집화

❖ 군집화 – cont.

- **일반 군집화**(hard clustering)
  - 데이터는 하나의 군집에만 소속
    - 예. k-means 알고리즘
- **퍼지 군집화**(fuzzy clustering)
  - 데이터가 여러 군집에 부분적으로 소속
  - 소속정도의 합은 1이 됨
    - 예. 퍼지 k-means 알고리즘

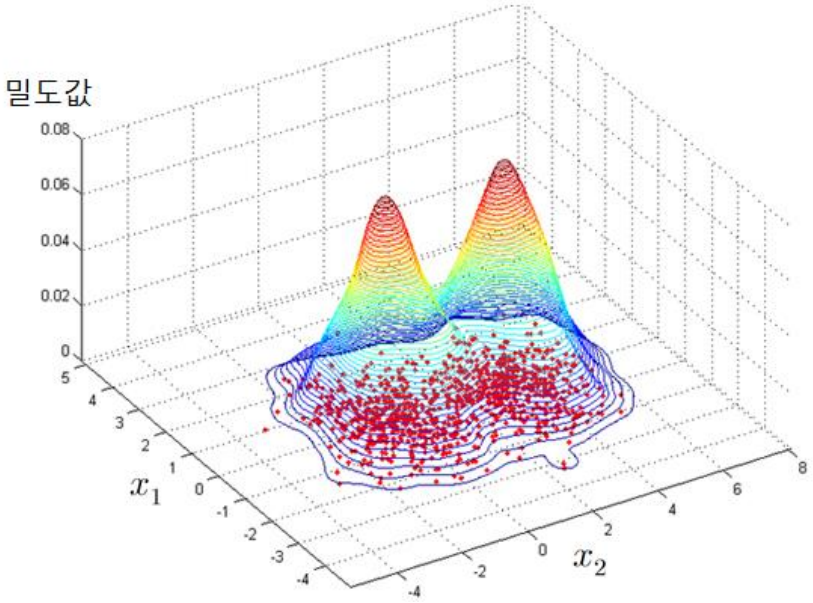
- **용도**
  - 데이터에 내재된 구조(underlying structure) 추정
  - 데이터의 전반적 구조 통찰
  - 가설 설정, 이상치(anomaly, outlier) 감지
  - 데이터 압축 : 동일 군집의 데이터를 같은 값으로 표현
  - 데이터 전처리(preprocessing) 작업

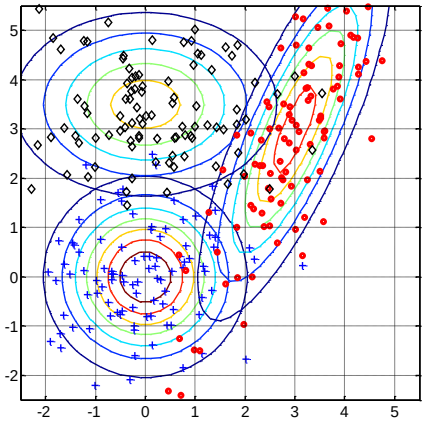
- **성능**
  - **군집내**의 **분산**과 **군집간**의 **거리**

3.4 비지도 학습 – 밀도 추정

❖ **밀도 추정**(density estimation)

- 부류(class)별 데이터를 만들어 냈을 것으로 추정되는 확률분포을 찾는 것





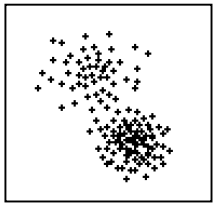
- **용도**
  - 각 부류 별로 주어진 **데이터**를 **발생**시키는 **확률** 계산
  - 가장 확률이 높은 부류로 **분류**

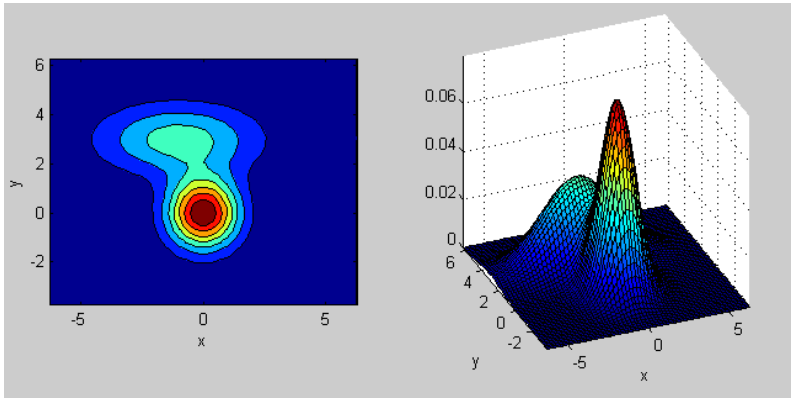
밀도 추정

❖ **밀도 추정** – cont.

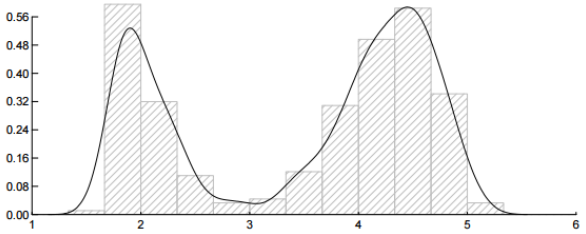
- **모수적(parametric) 밀도 추정**
  - 분포가 특정 수학적 함수의 형태를 가지고 있다고 가정
  - 주어진 데이터를 가장 잘 반영하도록 함수의 파라미터 결정
  - 전형적인 형태 : **가우시안(Gaussian) 함수** 또는 여러 개의 **가우시안 함수의 혼합**(Mixture of Gaussian)

- **비모수적(nonparametric) 밀도 추정**
  - 분포에 대한 특정 함수를 가정하지 않고, 주어진 데이터를 사용하여 밀도함수의 형태 표현
  - 전형적인 형태 : **히스토그램**(histogram)





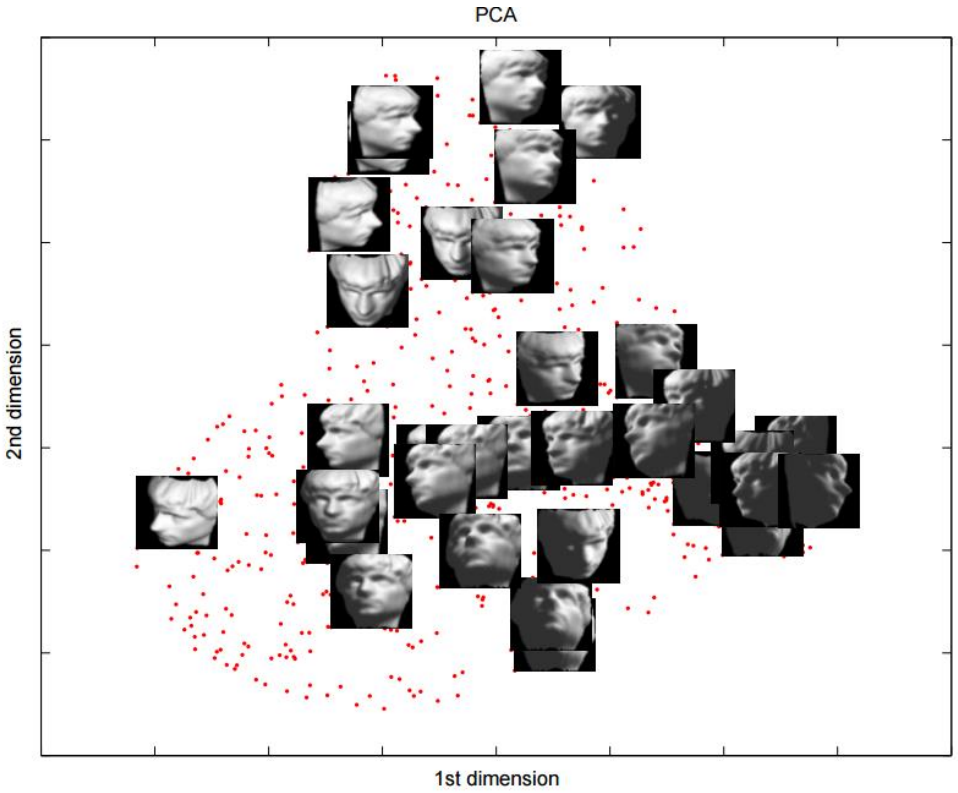
http://i.stack.imgur.com/pE0Xu.gif



3.5 비지도 학습 – 차원축소

❖ **차원축소**(dimension reduction)

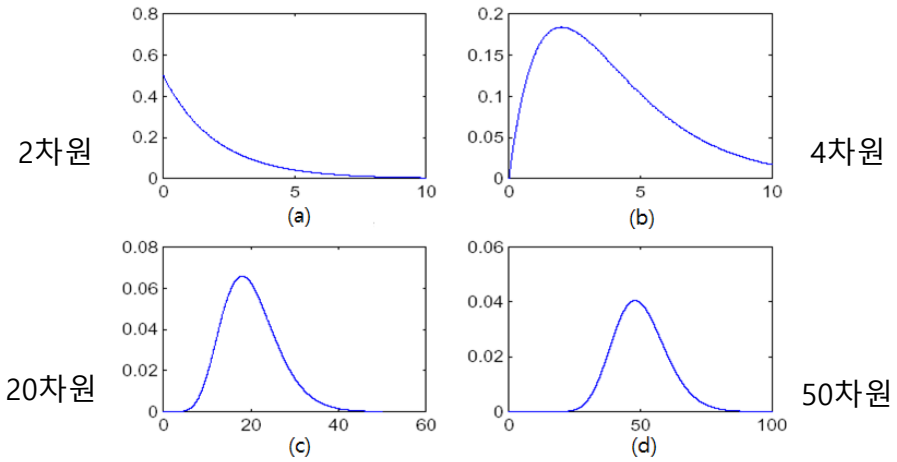
- **고차원의 데이터**를 정보의 손실을 최소화하면서 **저차원**으로 **변환**하는 것
- **목적**
  - 2, 3차원으로 변환해 시각화하면 직관적 데이터 분석 가능
  - **차원의 저주**(curse of dimensionality) 문제 완화



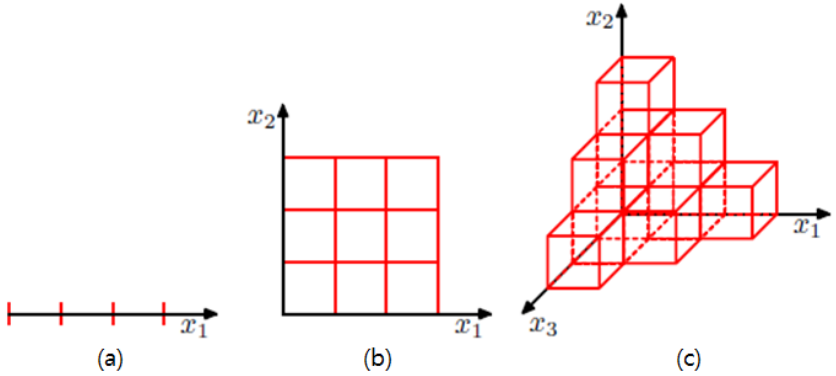
차원축소

❖ **차원축소** – cont.

- **차원의 저주**(curse of dimensionality)
  - 차원이 커질수록 **거리분포**가 **일정**해지는 경향



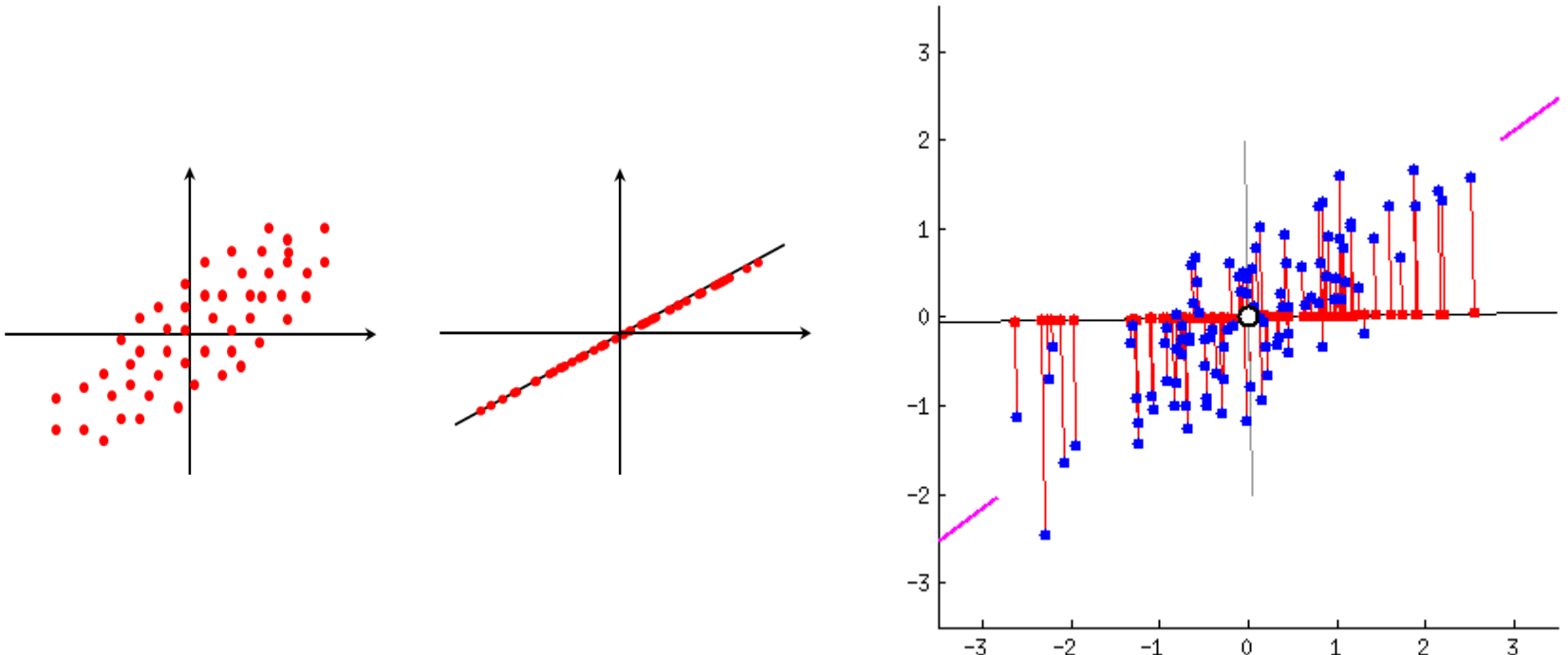
  - 원이 증가함에 따라 **부분공간의 개수**가 **기하급수적**으로 **증가**



차원축소

❖ **차원축소** – cont.

- **주성분 분석** (Principle Component Analysis, **PCA**)
  - 분산이 큰 소수의 축들을 기준으로 데이터를 사상(projection)하여 저차원으로 변환
  - 데이터의 **공분산행렬**(covariance matrix)에 대한 **고유값**(eigenvalue)가 큰 소수의 **고유벡터**(eigenvector)를 사상 축으로 선택



3.6 이상치 탐지

- **이상치**(outlier) 탐지
  - **이상치**
    - 다른 데이터와 크게 달라서 다른 메커니즘에 의해 생성된 것이 아닌지 의심스러운 데이터
    - 관심 대상
  - **잡음(noise)**
    - 관측 오류, 시스템에서 발생하는 무작위적인 오차
    - **관심이 없는 제거할 대상**
  - **신규성 탐지**(novelty detection)와 관련

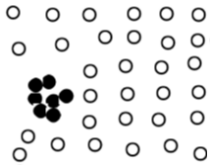
이상치 탐지

❖ **이상치**(outlier) 탐지 – cont.

- **점 이상치**(point outlier)
  - 다른 데이터와 비교하여 차이가 큰 데이터

- **상황적 이상치**(contextual outlier)
  - 상황에 맞지 않는 데이터
    예) 여름철에 25도인 데이터는 정산, 겨울철에 25도는 이상치

- **집단적 이상치**(collective outlier)
  - 여러 데이터를 모아서 보면 비정상으로 보이는 데이터들의 집단



... http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail, http-web, **ssh, buffer-overflow, ftp,** http-web, ftp, smtp-mail. http-web ...

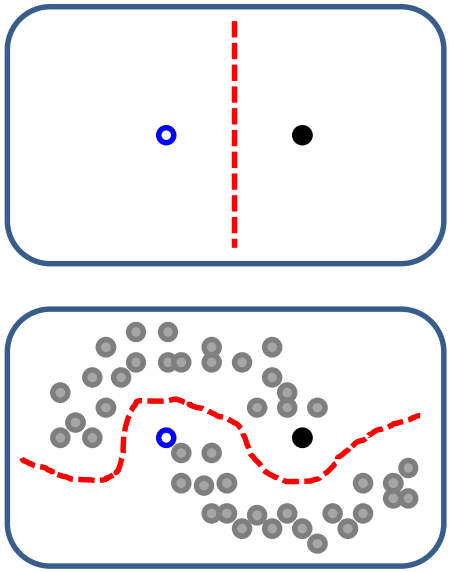
이상치 탐지

- **이상치**(outlier) 탐지 – cont.
  - **부정사용감지 시스템**(fraud detection system, FDS)
    - 이상한 거래 승인 요청 시에 카드 소유자에게 자동으로 경고 메시지 전송
  - **침입탐지 시스템**(intrusion detection system, IDS)
    - 네트워크 트래픽을 관찰하여 이상 접근 식별
  - 시스템의 고장 진단
  - 임상에서 질환 진단 및 모니터링
  - 공공보건에서 유행병의 탐지
  - 스포츠 통계학에서 특이 사건 감지
  - 관측 오류의 감지

3.7 반지도 학습

❖ **반지도 학습(semi-supervised learning)**

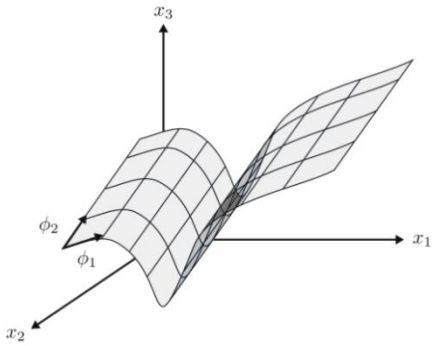
- 입력에 대한 결과값이 없는 **미분류 데이터**(unlabeled data)를 **지도학습에 사용**하는 방법
  - 분류된 데이터(labeled data)는 높은 획득 비용, 미분류 데이터는 낮은 획득 비용
  - 분류 경계가 인접한 미분류 데이터들이 동일한 집단에 소속하도록 학습
  - 같은 군집에 속하는 것은 가능한 동일한 부류에 소속하도록 학습



반지도 학습

❖ **반지도 학습의 가정**

- **평활성(smoothness, 平滑性) 가정**
  - 가까이 있는 점들은 서로 같은 부류에 속할 가능성이 높음
- **군집 (cluster) 가정**
  - 같은 군집에 속하는 데이터는 동일한 부류에 속할 가능성이 높음
- **매니폴드(manifold) 가정**
  - 원래 차원보다 낮은 차원의 매니폴드에 데이터에 분포할 가능성이 높음



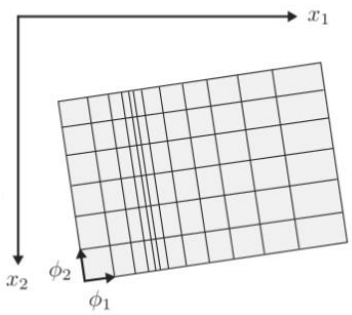


Image: ludovicarnold.altervista.org/