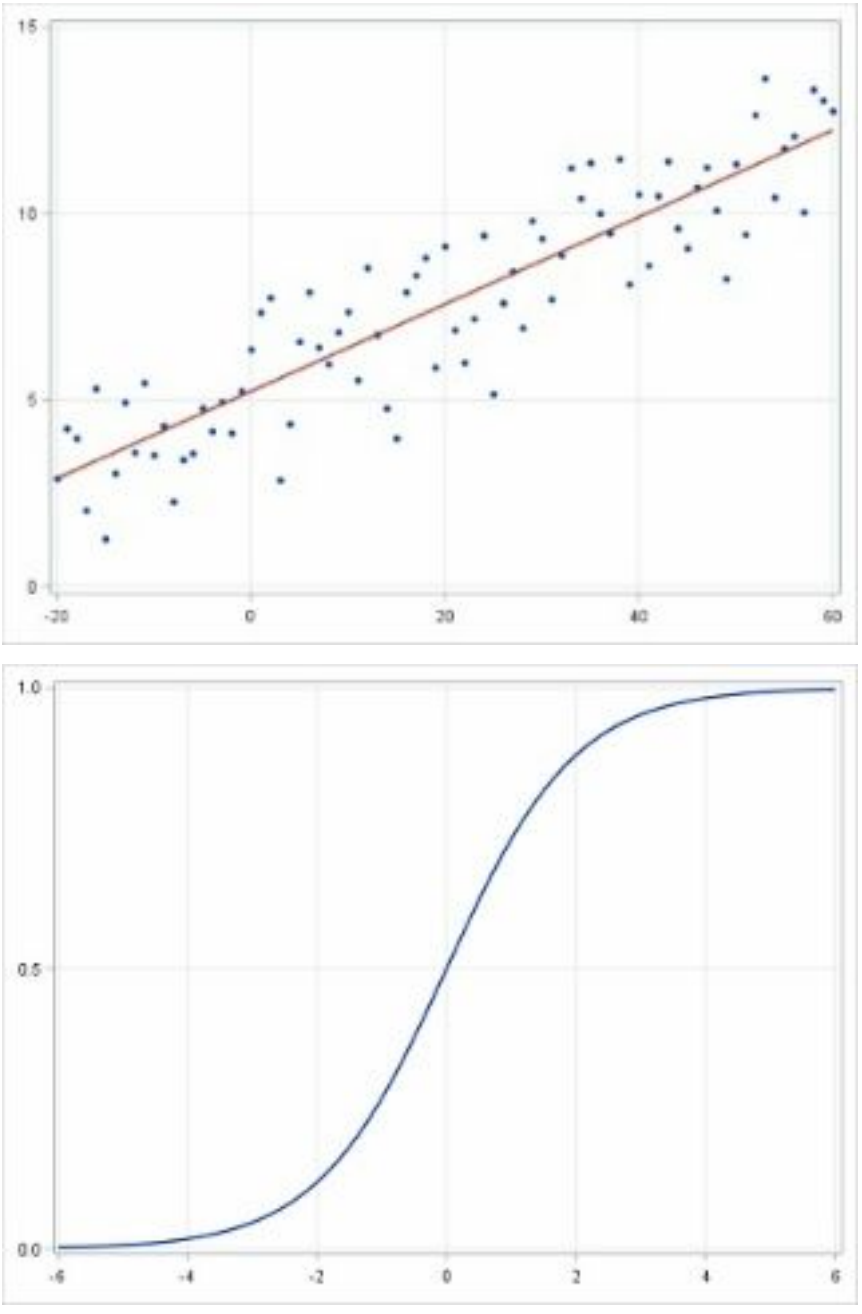
개별 알고리즘을 보다 자세히 들여다보면 알고리즘이 무엇을 제공하고, 어떻게 사용되는지 이해하는 데 도움이 된다.

## 1. 선형 회귀(Linear regression)와 로지스틱 회귀(Logistic regression)




▲선형 회귀(위쪽)와 로지스틱 회귀(아래쪽)

**선형 회귀는 연속적인 종속 변수 $y$ 와 한 개 이상의 예측 변수인 $X$ 사이의 관계를 모델링하는 접근법**입니다. $y$ 와 $X$ 사이의 관계는 $y=\beta^{T}X+\epsilon$ 와 같이 선형적으로 모델링할 수 있는데요. 매개 변수 벡터 $\beta$ 를 학습하는 학습 사례 $\{xi,yi\}_{i=1}^{N}$ 를 살펴본다

종속 변수가 연속형이 아니라 범주형이라면 선형 회귀는 로짓 연결(logit link) 함수를 이용해 로지스틱 회귀로 변환될 수 있다. 로지스틱 회귀는 단순하고 빠르지만 강력한 분류 알고리즘인데. 여기에서는 종속 변수 $y$ 가 오직 이진 값 $\{yi\in(-1,1)\}_{i=1}^{N}$ 만을 취하는 이진 사례에 대해 논의해보겠다. 이 경우 다중 클래스 분류 문제로 쉽게 확장될 수 있다.

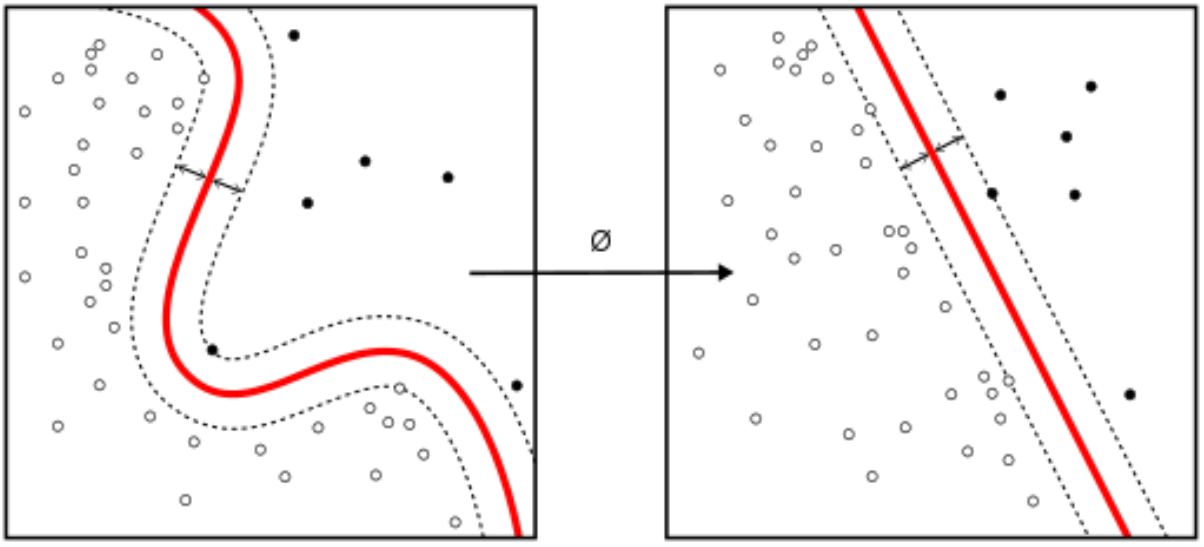
로지스틱 회귀에서는 주어진 자료가 $1$ 클래스에 속할 때와 $-1$ 클래스에 속할 때의 개연성을 비교 예측하기 위해 각기 다른 가설 클래스(hypothesis class)를 사용한다. 구체적으로 $p(y_i=1|x_i)=\sigma(\beta^{T}x_i)$ 와 $p(y_i=-1|x_i)=1-\sigma(\beta^{T}x_i)$ 형식의 함수를 학습하고자 한다. 여기에서 $\sigma(x)=\frac{1}{1+exp(-x)}$ 은 시그모이드(sigmoid) 함수인데. 학습 사례

$\{xi,yi\}_{i=1}^{N}$ 의 경우, 매개 변수 벡터 $\beta$ 는 데이터 세트를 기반으로 $\beta$ 의 로그 우도값(log-likelihood)을 극대화함으로써 학습할 수 있다.

## 2. 선형(Linear) SVM 및 커널(Kernel) SVM

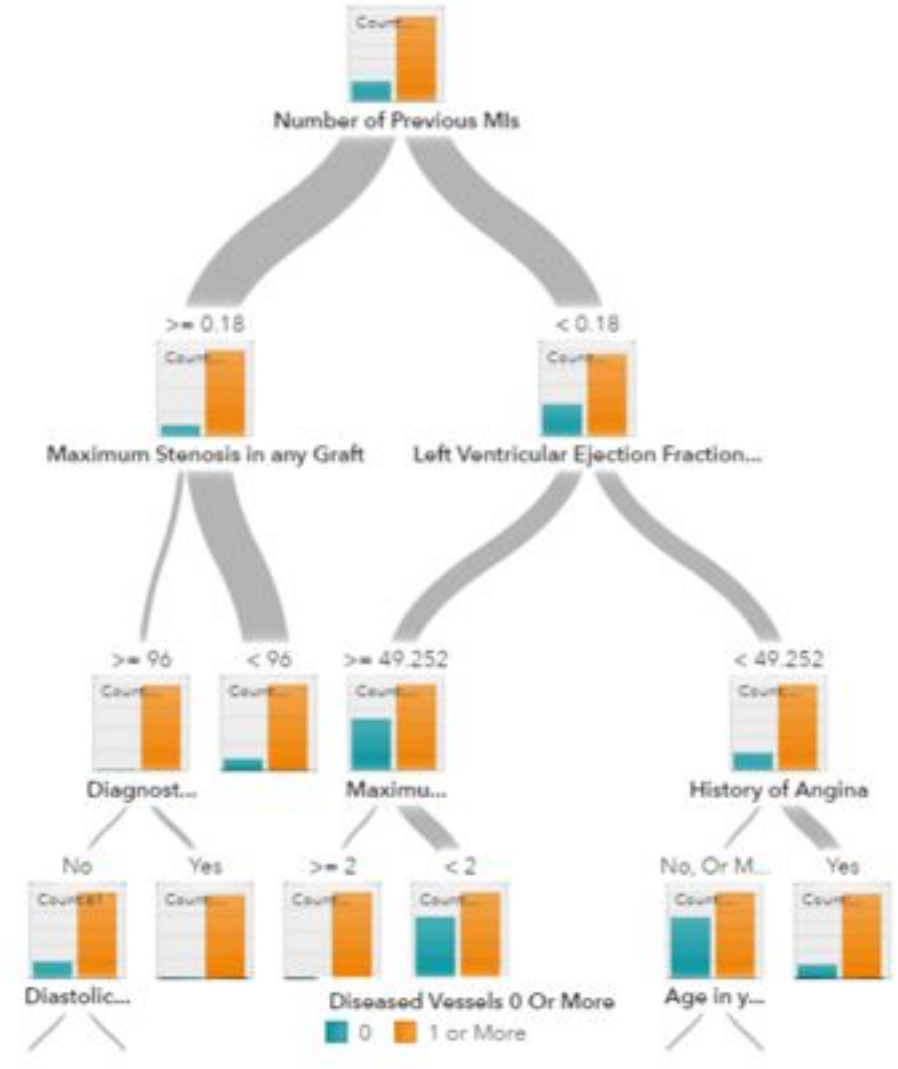
**커널 트릭(기법)은 분리 가능한 비선형 함수를 고차원의 분리 가능한 선형 함수로 매핑하기 위해 사용된다.** 서포트 벡터 머신(SVM; support vector machine) 학습 알고리즘은 초평면(hyperplane)의 법선 벡터(normal vector) 'w'와 편향 값(bias) 'b'로 표현되는 분류기(classifier)를 찾는다. 이러한 초평면(경계)은 가능한 최대 오차(margin)로 각기 다른 클래스를 분리하는데. 그러면 문제를 제약 조건이 있는(constrained) 최적화 문제로 변환할 수 있다.

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & ||w|| \\ \text{subject to} \quad & y_i\left(w^T X_i - b\right) \geq 1, \; i = 1, \ldots, n. \end{aligned}$$



▲커널 트릭(기법)은 분리 가능한 비선형 함수를 고차원의 분리 가능한 선형 함수로 매핑하기 위해 사용된.

## 3. 트리와 앙상블 트리(ensemble tree)

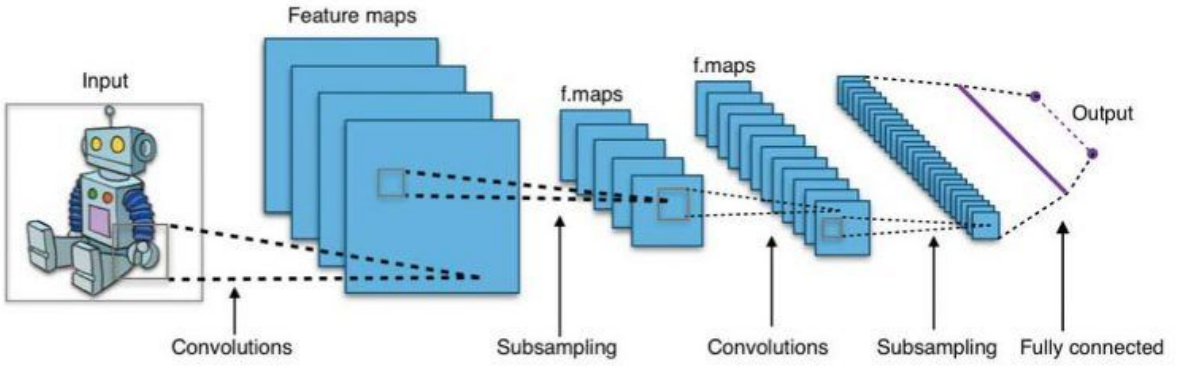



▲예측 모델의 의사결정 트리

의사결정 트리, 랜덤 포레스트(random forest), 그래디언트 부스팅(gradient boosting)은 모두 의사결정 트리를 기반으로 한 알고리즘임. 다양한 종류의 의사결정

트리가 있지만, 모두 동일한 작업을 수행. 즉 특징 공간(feature space)을 거의 같은 레이블로 구별되도록 분리한다. 의사결정 트리는 이해와 구현이 쉽지만 가지를 다 쳐내고 트리의 깊이가 너무 깊어질 경우 데이터를 과적합(overfit)하는 경향이 있다. 랜덤 포레스트와 그래디언트 부스팅은 일반적으로 높은 정확성을 달성하고 과적합 문제를 해결하기 위해 트리 알고리즘을 사용하는 두 가지 방법이다.

## 4. 신경망과 딥러닝



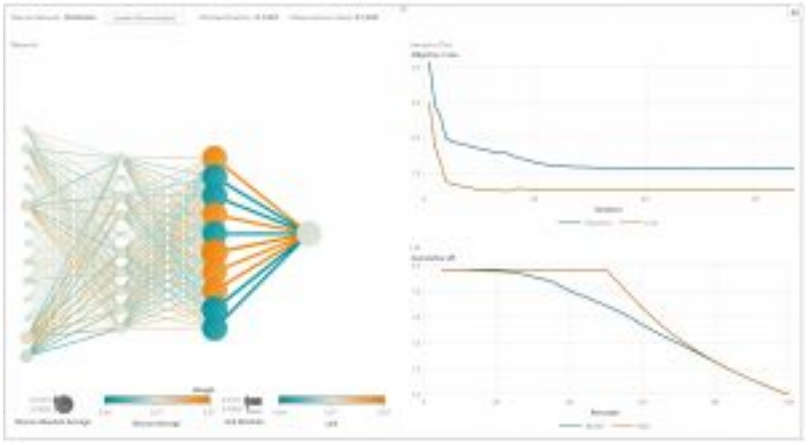

▲컨볼루션 신경망(convolution neural network) 아키텍처(이미지 출처: wikipedia creative commons)

신경망은 병렬 분산 처리 능력 덕분에 1980년대 중반 크게 성장했다. 그러나 신경망 매개 변수를 최적화하기 위해 널리 사용되는 역전파(back-propagation) 학습 알고리즘이 효과가 없어 신경망 연구가 지연됐는데. 이후 컨벡스(볼록) 최적화(convex optimization) 문제가 해결됨으로써 쉽게 학습할 수 있는 서포트 벡터 머신(SVM)과 여타 단순한 다른 모델들이 서서히 머신러닝의 신경망을 대체했다.

최근 몇 년간 새롭게 개선된 비지도 사전 학습(unsupervised pre-training)과 계층별 탐욕 학습(layer-wise greedy training) 등의 학습 기법들은 신경망에 대한 관심을 부활시키는 계기가 되었고. 또 GPU(graphical processing unit)와 MPP(massively parallel processing)와 같이 점차 강력해지는 연산 능력은 신경망을 다시 채택하게

하는 원동력이 됐었다. 신경망 연구가 재개되면서, 수천 개의 계층을 가진 모델이 개발되기 시작했다.

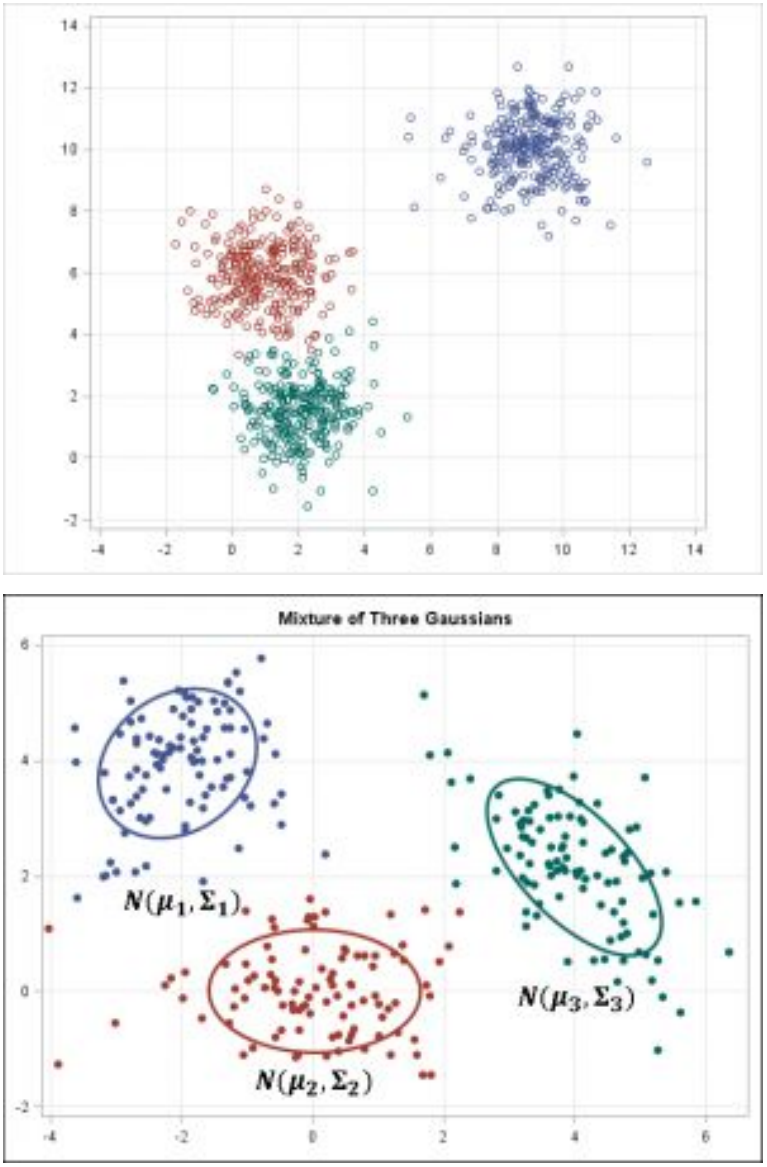
다시 말해, 얕은(shallow) 신경망이 딥러닝 신경망으로 진화한것.! 심층(deep) 신경망은 지도 학습에 매우 성공적이였다. 딥러닝은 음성이나 이미지 인식에 사용될 때 인간만큼 또는 심지어 인간보다 더 나은 성능을 보인다. 또 딥러닝은 특징 추출(feature extraction)과 같은 비지도 학습 과제에 적용될 때 인간의 개입이 훨씬 줄어든 상황에서 원시 이미지(raw images)나 음성으로부터 특징을 추출할 수 있다.



▲신경망

**신경망은 입력 계층(input layer), 은닉 계층(hidden layers), 출력 계층(output layer)의 세 부분으로 구성된다.** 학습 표본(training samples)은 입력 및 출력 계층을 정의하는데요. 출력 계층이 범주형 변수일 때 신경망은 분류 문제를 해결. 출력 계층이 연속 변수일 때 신경망은 회귀 작업을 위해 사용될 수 있다. 또 출력 계층이 입력 계층과 동일할 때 신경망은 고유한 특징을 추출하기 위해 사용될 수 있다. 이때 은닉 계층의 수는 모델 복잡성과 모델링 수용력(capacity)을 결정한다.

## 5. K-평균/K-모드(k-means/k-modes), 가우시안 혼합 모델(GMM; Gaussian mixture model) 클러스터링

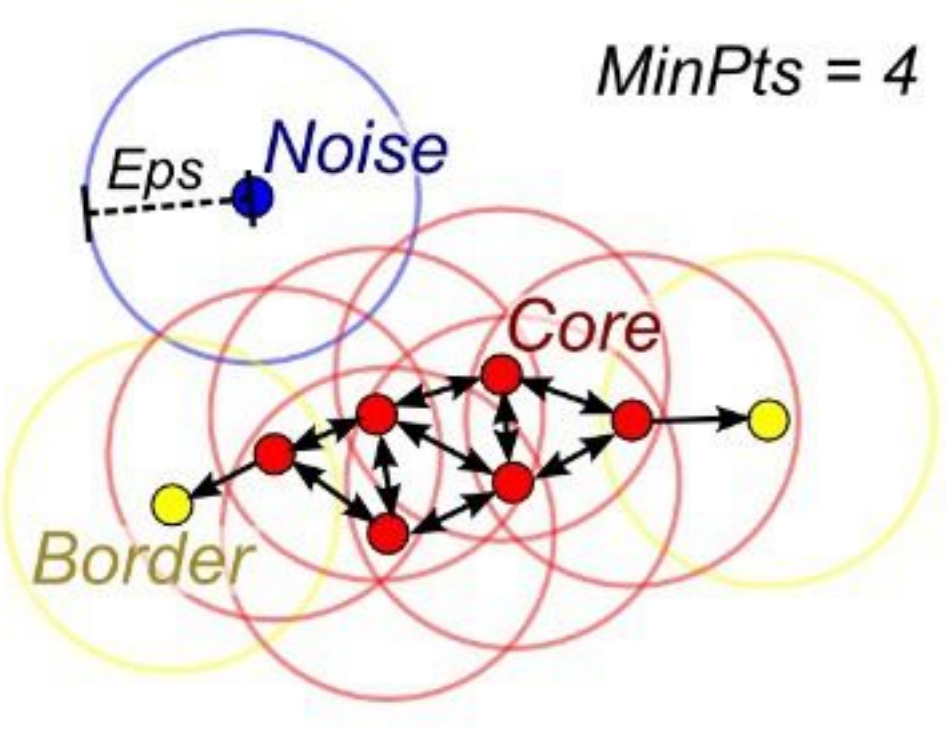



▲K-평균 클러스터링(왼쪽)과 가우시안 혼합 모델

K-평균/K-모드 클러스터링과 GMM 클러스터링의 목표는 n개의 관측치(observations)를 k개의 클러스터로 나누는 것이다. K-평균은 표본을 하나의

클러스터에만 강하게 결속시키는 '하드 할당(hard assignment)'를 정의한다. 반면 GMM은 각 표본이 확률 값을 가짐으로써 어느 한 클러스터에만 결속되지 않는 '소프트 할당(soft assignment)'을 정의하는데. 두 알고리즘 모두 클러스터 k의 수가 주어질 때 클러스터링을 빠르고 단순하게 수행할 수 있다.

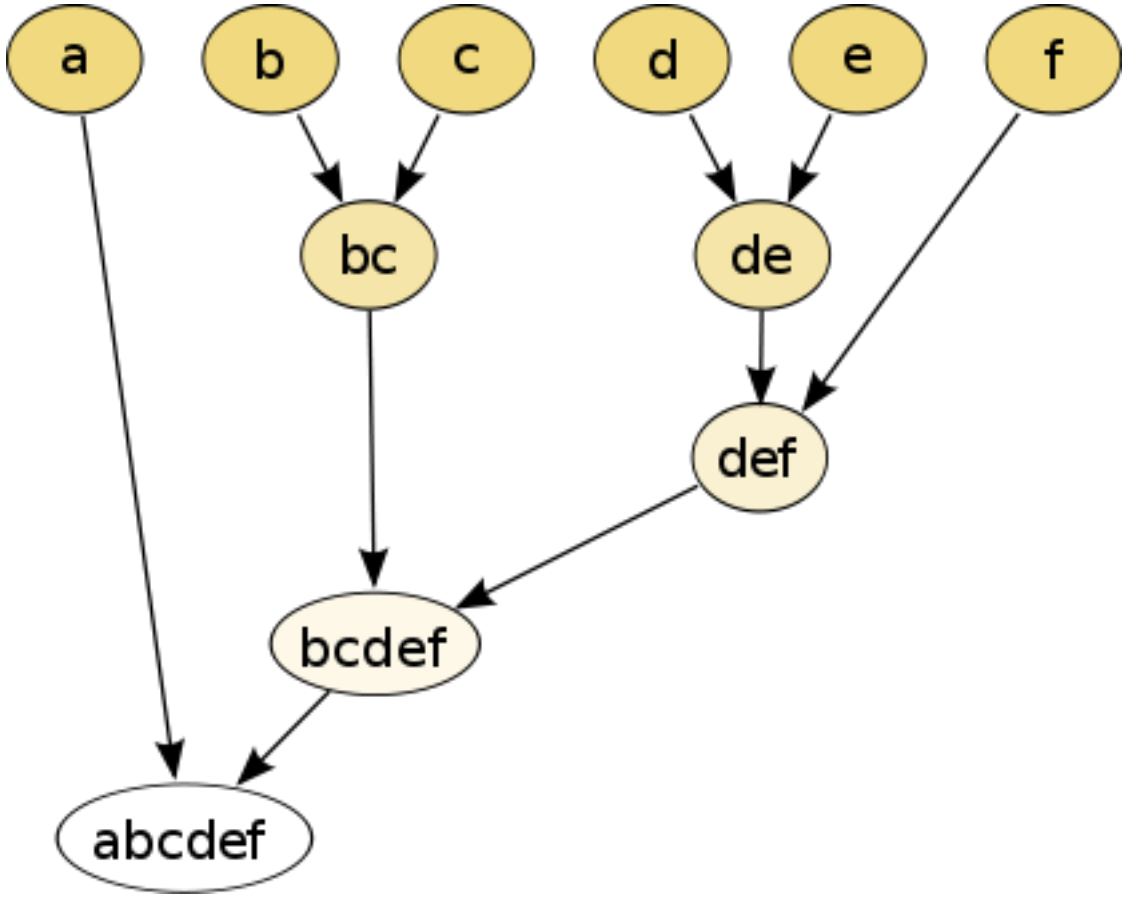
## 6. DBSCAN




▲DBSCAN 도해(이미지 출처: Wikipedia)

클러스터 k의 수가 주어지지 않을 때에는 밀도 확산(density diffusion)을 통해 표본을 연결함으로써 DBSCAN(density-based spatial clustering)을 사용할 수 있다.

## 7. 계층적 군집화(Hierarchical clustering)




계층적 분할은 트리 구조인 덴드로그램(dendrogram)를 이용해 시각화할 수 있다. 각기 다른 K를 사용해 클러스터를 정제하거나 조대화할 수 있는 각기 다른 세분화(granularities) 수준에서 입력과 분할 결과를 확인할 수 있기 때문에 클러스터의 개수가 필요 없다.

## 8. PCA, SVD, LDA

일반적으로 머신러닝 알고리즘에 많은 수의 특징을 직접 투입하는 것은 선호되지 않는다. 일부 특징은 관련이 없거나 '고유한' 차원수가 특징의 수보다 적을 수 있기 때문인데. 따라서 주성분 분석(PCA; principal component analysis), 특이값 분해(SVD; singular value decomposition), 잠재 디리클레 할당(LDA; latent Dirichlet allocation)을 이용해 차원 축소를 수행할 수 있다.

PCA는 원래의 데이터 공간을 저차원의 공간으로 매핑하면서 가능한 많은 정보를 보존하는 비지도 클러스터링 방식. PCA는 기본적으로 데이터 분산(variance)을 가장 많이 보존하는 하위 공간(subspace)을 찾는데. 하위 공간은 데이터의 공분산 매트릭스(covariance matrix)의 지배적인 고유 벡터(eigenvectors)에 따라 정의한다.

SVD는 중앙 데이터 매트릭스의 SVD(특징 vs. 표본)가 PCA로 찾은 것과 동일한 하위 공간을 정의하는 지배적인 왼쪽 특이 벡터(left singular vectors)를 제공한다는 점에서 PCA와 관련되어 있다. 그러나 SVD는 PCA가 할 수 없는 작업을 수행할 수 있기 때문에 훨씬 다재다능한 기법입. 예를 들어, 사용자 대 영화 매트릭스의 SVD는 추천 시스템에서 사용할 수 있는 사용자 프로파일과 영화 프로파일을 추출할 수 있다. 또 SVD는 자연어 처리(NLP; natural language processing) 과정에서 잠재 의미 분석(latent semantic analysis)으로 알려진 주제 모델링(topic modeling) 도구로서 널리 사용된다.

자연어 처리(NLP)와 관련된 기법은 잠재 디리클레 할당(LDA)이다. LDA는 확률적 주제 모델(probabilistic topic model)로 가우시안 혼합 모델(GMM)이 연속 데이터를 가우시안 밀도로 분해하는 것과 비슷한 방식으로 문서를 주제를 기준으로 분리한다. GMM과 다르게 LDA는 이산 데이터(discrete data, 문서 내 단어)를 모델링하고, 주제는 디리클레 분포(Dirichlet distribution)에 따라 연역적(priori)으로 분포돼야 하는 제약이 있다.