고급 통계학

| **@KMOOC**

| **부산대학교 김충락 교수님**

| **http://www.kocw.net/home/cview.do?cid=f653717700cd7176**

통계학에서의 회귀 분석

분석하고자 하는 변수들 간의 유의미한 상관 관계를 알아낼 수 있다.

단순회귀모형

연속형인 두 변수를 가정할 때, 이들의 관계, 즉 상관 관계를 표현한다. 이 때 각 변수는 **독립 변수**와 **종속 변수**로 구분된다.

```
(x1, y1), (x2, y2), (x3, y3), ... (x?, y?) | (X, Y)

Y = A + BX + E
```

A : 절편(*y-intercept*)

B : 기울기

E : 오차(잔차)

최적 직선식 추정

변수 간 상관 관계를 알아보기 위해 단순회귀모형의 모수를 추정한다.

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} \left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right) \right)}{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta} \cdot \overline{x}$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$$

E 는 실제 값과 추정 값의 차이로 계산된다.

회귀 분석에서의 분산 분석

분산 분석을 통해, B 가 0이 아니라고 판단될 때, 회귀 모형이 유의하다고 할 수 있다.

분산 분석은 **변동**을 중심으로 여러 연산을 통해 판단하게 되는데, 여기서 **변동**이란 실제 값과 Y 값의 차이이다. 여기에 E 가 포함된다.

ANOVA Test Table

cuemath THE MATH EXPERT

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	SSB = $\Sigma\, n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	SSE = $\Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

Sum of Squares(SS) 가 **변동**의 제곱합을 의미하며, 여기서 파생된 여러 요소들로 분산분석표가 구성된다.

Degrees of Freedom 은 **자유도**로, A 를 제외한 독립 변수의 개수이다.

F-value 는 **검정 통계량**이며 이것이 회귀 모형이 얼마나 유의미한지 판단할 수 있는 지표(+**p-value**)이다.

평가

회귀 분석을 시행할 때 E 는 독립, 그리고 정규 분포를 따른다고 가정한다. 회귀 모형의 이 **오차**에 대한 조건 몇 가지(생략)를 만족하는 것으로 회귀 분석의 결과를 평가할 수 있다.

이외의 방법으로는 이상치 분석, 영향치(제거되면 회귀 모형에 큰 영향을 주는 것들) 분석, 샤피로 검정, 등분산 검정, 독립성 검정, 신뢰 구간과 예측 구간 탐색 등이 있다.

분류에서의 통계학

회귀 분석은 양적 변수, 즉 연속적인 변수들을 주로 다루지만 분류 문제에선 질적 변수를 다루게 된다. 분류 문제에서 핵심이 되는 동작들은 대부분이 통계학에 기반하고 있다.

데이터

- 독립성
- 통계적 검정
- 차원 핸들링
- 특성 분석

분류기

- 베이즈 정리 & 나이브 베이즈
- 최대우도법
- 로지스틱

평가

- 신뢰 구간
- 혼동 행렬
- ROC & AUC

군집 분석

통계학의 관점에선 **다변량 분석**(**다변수 분석**)에 포함되며 기술 통계학에 많이 활용된다.

본격적으로 체계화되기 시작한 건 머신 러닝의 비지도 학습으로 **군집화**(**Clustering**) 개념을 적용할 수 있게 되면서부터이다.

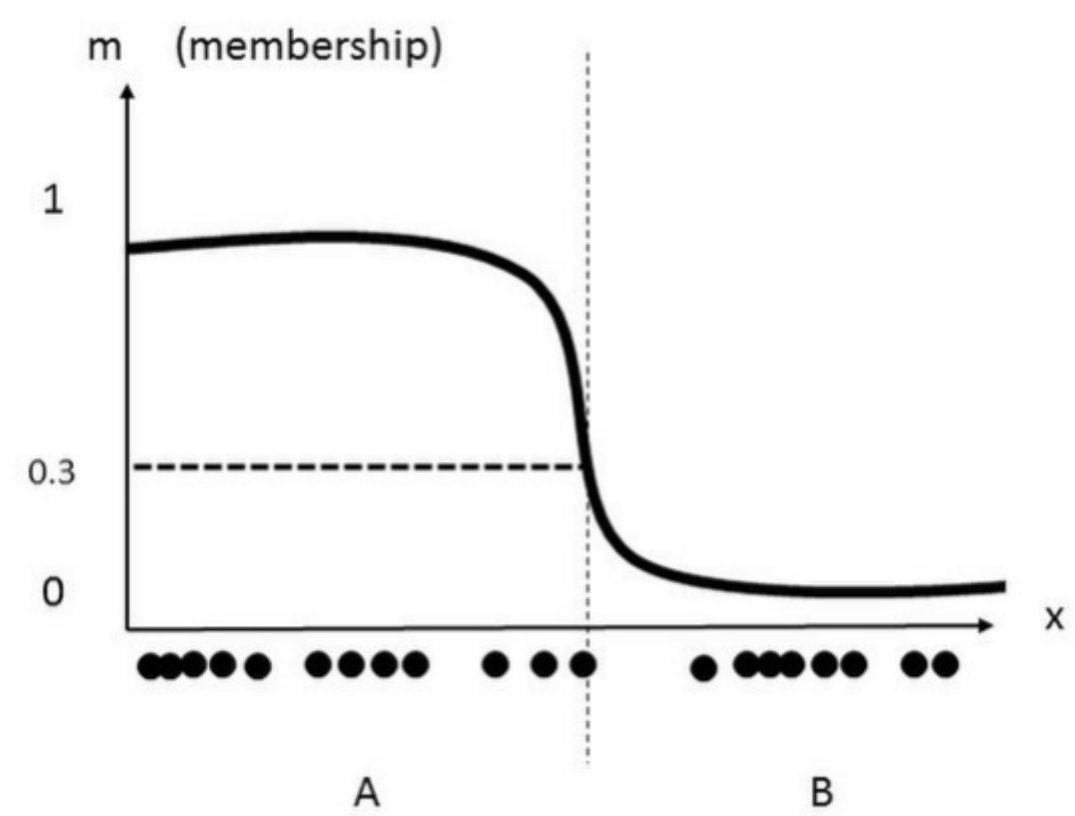
퍼지 군집화

펴지 이론

```
추상적인 상황, 애매모호한 기준을 다루기 위해 고안된 이론이다.

고안된 초창기엔 쓸데없는 이론이라며 욕을 많이 먹었지만 실제 상황에 적용해 얻을 수 있는 이득이 많아 활발하게 연구되고 있다.
```

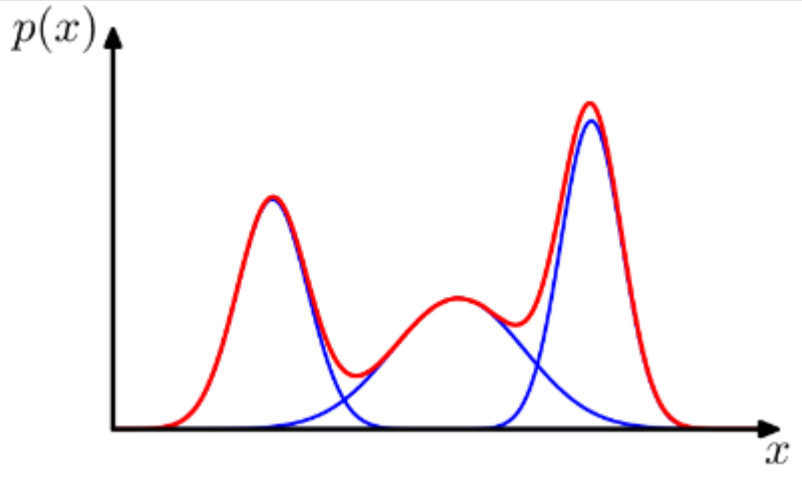
퍼지 군집화의 핵심은, 어떤 표본이 어느 군집에 속하는지를 확률적으로 판단한다는 것이다. 판단의 척도를 *Membership* 이라고 칭한다.



즉, 어떤 표본이 특정 군집에 속할지의 여부를 단순히 0과 1로 판정하는 기존의 군집화와 다르게, 0과 1 사이의 확률로 정의한다는 것이다.

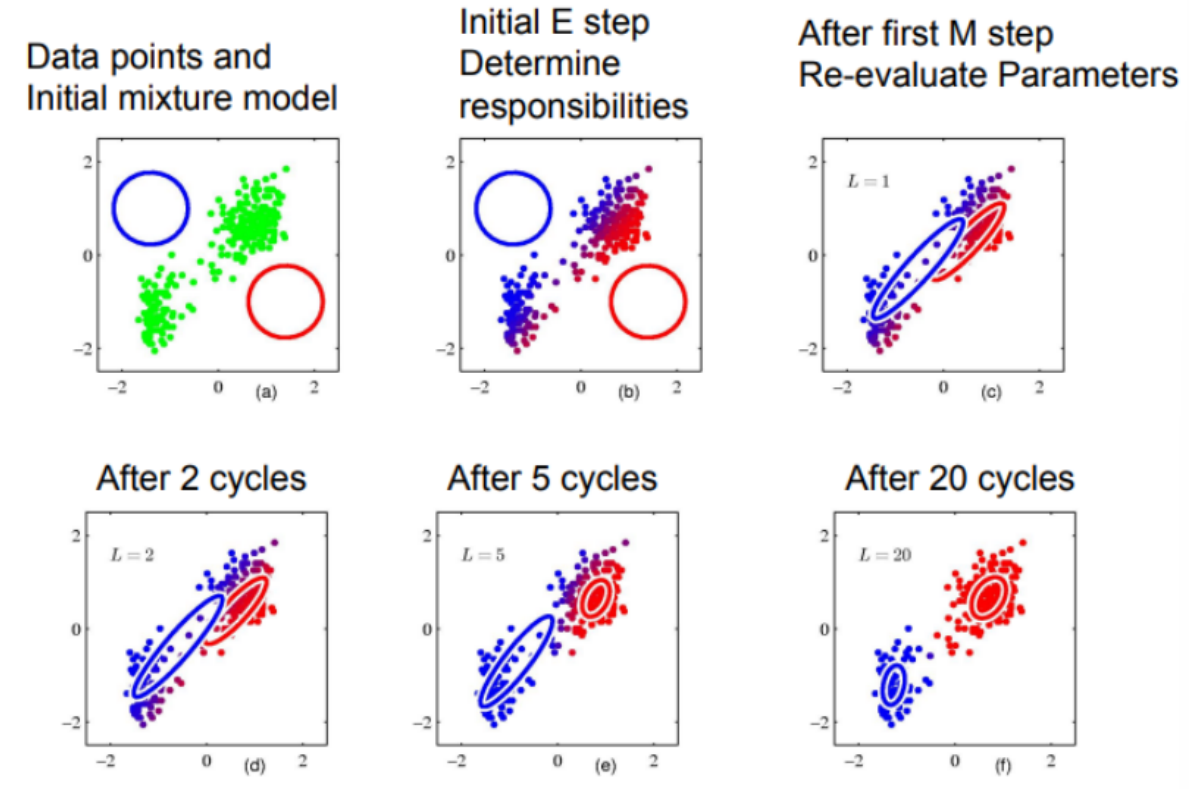
혼합 분포 군집화

여러 분포가 쇽률적으로 선형 결합한 분포를 **혼합 분포**라 칭한다.



혼합 분포 그래프의 각 분포는 하나의 모형이며, 표본을 군집화하는 과정에서 각 표본이 어떤 모형에 속할지를, 즉 각 표본이 어느 분포를 따르는지를 확률적으로 구분한다.

EM Algorithm



최대우도법과 추정, 그리고 갱신 과정으로 하나의 사이클이 돌아가면 위 그림과 같이 수렴된 군집화 결과를 얻을 수 있다.