6월18일~6월23일

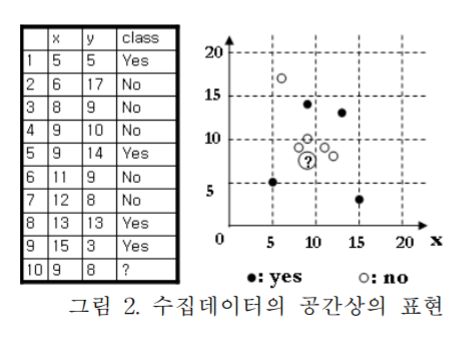
∙ 사용자 모델링을 위해서는 사용자의 성향 및 행위 등의 다양한 정보를 수집하여 분석에 이용한다. 하지만 사용자로부터 얻은 데이터는 기계나 환경에서 수집된 데이터 보다 패턴을 찾기 힘들어 모델링하기 어렵다. 그 이유는 사용자는 사용자의 현재 상태와 상황에 따라 다양한 견과를 보이며, 일관성을 유자 하지 않은 경우가 있기 때문이다. 사용자 모델링을 위해서는 분산되어 있는 데이터에서 노이즈를 선별하고 연관성 있는 데이터를 분류할 수 있는 기술이 필요하다. 이번에는 사용자로부터 수집된 데이터를 k-NN(Nearest Neighbor) 기법을 이용하여 노이즈를 선별하는 방법에 대해 공부하였다.

**∙ k-NN 기법을 이용한 노이즈 선별 방법**

K-NN 기법을 이용한 노이즈 선별 방법은 각 인스턴스 간에 거리를 기반으로 계산한다. 예를 들어 그림 2와 같이 10개의 인스턴스가 있다고 가정하자 각 인스턴스가 가지고 있는 속성은 x와 y 두개이며 정수형의 값을 갖는다. 각 속성의 종속변수인 Class는 Yes와 NO 두가지 값을 가질 수 있다. 그림 2는 10개의 인스턴스를 좌표상에 표현한 것이다.

그림에서 보면 알 수 있듯이 10번째 인스턴스의 Class를 알 수 없다. 그렇다면 여기서 10번째 Class를 무엇이라 할 수 있을까? 아마도 No라고 대답하는 사람이 대다수 일 것이다. 그 이유는 현재 알지 못하는 10번째 인스턴스의 주변 분포가 Yes보다는 Np가 더 많기 때문이다. 만약 주어진 데이터 10번째 인스턴스의 Class가 Yes라고 주어졌다고 가정하자.

주변 분포를 볼 때 No가 더 많으므로, 10번째 데이터를 노이즈로 결정할 수 있을 것이다. 이처럼 주어진 전체 데이터에서 한 인스턴스의 주변 분포를 측정하여 자기 자신의 노이즈 여부를 판단하는 것이 k-NN 기법을 이용한 노이즈 선별 방법이다.



이번주는 k-NN 기법으로 노이즈 선별방법에 대하여 공부하였다. 다음주에는 k-NN 기법 노이즈 선별을 위한 고려사항과 노이즈 선별을 이용한 학습자 진단 과정에 대해서 공부할 예정이다.