<용어체크>

최소제곱법

단일 또는 다중 회귀 식에서 발생하는 오차를 최소하 하기 위한 방법으로 잔차의 제곱을 구해 최소값을 갖도록 함수를 구성하는 것을 말한다. 그러나 최소 제곱법은 특이값에 취약하다는 단점을 가지고 있다.

정규화

최소제곱법의 단점을 보완하기 위해 가중치를 이용하는 방법 중 하나로 최소제곱법으로 구성한 방정식에 페널티를 부여하는 방법이다. L1, L2 정규화가 있으며 L1 정규화는 종속 변수에 가중치 계수 w의 절대값을 페널티로 더하는 방법이며, L2 정규화는 최소제곱법의 종속 변수인 잔차 제곱의 합에 가중치 계수인 w의 제곱의 합을 페널티로 추가한 것이다.

유사도

회귀 분석할 때 상관 관계를 이용하여 두 변수값 쌍이 얼마나 비슷한가를 측정하는 방법이다. 수학적인 유사도 개념으로 코사인 유사도, 상관 함수, 편집 거리, 자카드 계수 등을 사용한다.

<학습내용>

가중 회귀분석

유사도

<학습목표>

최소제곱법 수정을 통한 가중 회귀분석을 설명할 수 있다.

유사도를 통한 상관관계를 설명할 수 있다.

Q. 검색 엔진에 검색어를 넣어 주는데 가끔 오타가 나서 엉뚱한 검색어를 입력한 경우가 많습니다. 그런데도 검색 엔진에서는 오타가 있는 검색어 대신 원래 검색하고자 하는 단어로 바꿔 검색 결과를 보여 줍니다. 어떤 식으로 가능한 것인지요?

: 질문과 같은 경우가 상당히 많습니다. 요즘은 모바일 환경에서 검색하는 경우도 있다 보니 가상 키보드가 손에 익지 않아 오타가 나는 경우가 많죠. 그런데 어떻게 오타를 보고 제대로 된 검색어로 바꿔 검색을 한다 던지 아니면 관련 검색어를 제시한다 던지 할까요. 오타가 난 검색어와 원래 제대로 된 검색 어 간의 틀린 부분을 찾습니다. 틀린 부분을 바꿔 주면 되겠죠. 틀렸다, 아니다를 어떻게 분석하는 가 하면 바로 단어 간의 유사도를 분석하여 틀린 단어와 원래의 단어 간의 거리를 측정합니다. 거리가 바로 유사도가 되는 것이죠. 단어 간의 거리를 측정하여 가장 가까운 값을 갖는 단어가 추천 단어가 되는 것입니다. 유사도가 바로 거리의 개념을 사용해서 응용되는 예가 되겠습니다.

가중 회귀분석

- LOWESS, L1 정규화, L2 정규화 등을 사용한다.
- 단순 최소제곱법은 특이값에 취약하여 최소제곱법에 가중치를 더하여 최소제곱법 보정한다.
- LOWESS 분석은 독립 변수의 값에서 멀어져 있는 점의 기울기를 조정하여 특이점으로 인한 영향을 무시하도록 보정한다.
- L2 정규화는 일반적인 회귀 모델로 계산한다.
- L1 정규화는 볼록 최적화의 추정 알고리즘을 사용한다.

유사도

- 회귀 분석 시 상관 관계를 확인하는 것이다.
- 코사인 유사도는 두 변수 간의 떨어진 정도를 코사인 함수로 표현하여 같은 방향으로 진행하는 경우는 유사도가 높은 것으로 분석한다.
- 유사도를 거리 개념으로 이해하여 분석하는 편집 거리, 레벤슈타인 거리, 해밍 거리, 유클리드 거리, 마할라노비스 거리, 자카드 계수 등을 사용한다.
- 문서 분석 및 문자 들 간의 연관성을 거리값으로 산출한다.