9-2. Machine Learning

홍형경

chariehong@gmail.com

2020.01

1. Machine Learning 이란?

· 머신 + 러닝(학습)

· 기계(컴퓨터)가 데이터를 학습해 유의미한 예측 결과를 산출

· 기존에는 사람(프로그래머)가 업무 로직 분석 후 프로그램을 개발

· 데이터 양이 많아지고 사람이 파악하기 어려운 숨어 있는 로직을 머신 러닝을 통해 도출, 예측

· 제공된 데이터를 컴퓨터가 제시된 알고리즘으로 학습, 훈련, 미래 데이터에 대한 예측

· 지도 학습, 비지도 학습, 강화 학습이 있음

1. Machine Learning 이란?

(1) 지도 학습 (Supervised Learning)

· 분석 대상 데이터에 이미 레이블(Label) – 정답 -이 존재

· 훈련 데이터로 학습 후, 나머지 데이터로 테스트

· K-최근접 이웃, 선형 회귀, 로지스틱 회귀, SVM(서포트 벡터 머신), 결정 트리와 랜덤 포레스트, 신경망

1. Machine Learning 이란?

(2) 비지도 학습 (Unsupervised Learning)

· 분석 대상 데이터에 레이블(Label)이 없음

· K-평균, 계층군집분석, 기댓값 최대화, 연관규칙

2. 연관 규칙

· 장바구니 분석

· 맥주와 기저귀

- 맥주가 판매된 날 기저귀도 많이 판매 되었다.
- 아이 아빠들이 기저귀 사러 왔다가 맥주도 같이 구매
- 맥주 옆에 기저귀 진열해 매출 증대

· A 상품을 구매 → B 상품 구매 : 두 상품간 연관성 있음

· 상품이 수 백, 수 천개인데 어떻게 비교? → APIRIOR 알고리즘

2. 연관 규칙

· 장바구니 분석

· 맥주와 기저귀
  - 맥주가 판매된 날 기저귀도 많이 판매 되었다.
  - 아이 아빠들이 기저귀 사러 왔다가 맥주도 같이 구매
  - 맥주 옆에 기저귀 진열해 매출 증대

· A 상품을 구매 → B 상품 구매 : 두 상품간 연관성 있음

· 상품이 수 백, 수천개인데 어떻게 비교? → APRIORI 알고리즘

2. 연관 규칙

ID	ITEMS
1	달걀,라면,참치캔
2	라면,햇반
3	라면,콜라
4	달걀,라면,햇반
5	달걀,콜라
6	라면,콜라
7	라면,햇반
8	달걀,라면,참치캔,콜라
9	달걀,라면,콜라
10	양파

2. 연관 규칙

ID	달걀	라면	참치캔	햇반	콜라	양파
1	1	1	1	0	0	0
2	0	1	0	1	0	0
3	0	1	0	0	1	0
4	1	1	0	1	0	0
5	1	0	0	0	1	0
6	0	1	0	0	1	0
7	0	1	0	1	0	0
8	1	1	1	0	1	0
9	1	1	0	0	1	0
10	0	0	0	0	0	1

2. 연관 규칙

· 지지도 (Support)

- 전체 거래항목 중 상품 A와 상품 B를 동시에 포함하여 거래하는 비율

- 지지도 = P(A∩B) : A와 B가 동시에 포함된 거래 수 / 전체 거래 수

- 달걀과 라면의 지지도

= 4 / 10 = 0.4

2. 연관 규칙

· 신뢰도 (Confidence)

- 상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비중

- 상품 A를 구매 했을 때 상품 B를 구매할 확률

- 신뢰도 = $P(A \cap B) / P(A)$: A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수

- 달걀과 라면의 신뢰도

  = 4 / 5 = 0.8

2. 연관 규칙

· 향상도 (Lift)

- 상품 A의 거래 중 B 상품이 포함된 거래 비율 / 전체 거래 중 상품 B가 거래된 비율
  → A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B의 확률 증가 비율
  → A 구매 시 B 구매 확률이, A 고려 안했을 때 B 구매 확률에 비해 얼마나 향상되는가

- 향상도 = $P(A \cap B) / P(A)*P(B) = P(B|A) / P(B)$
  = A와 B가 동시에 일어난 횟수 / A, B가 독립된 사건일 때 A,B가 동시에 일어날 확률

- 달걀과 라면의 향상도
  = $P(\text{달걀}|\text{라면}) / (P(\text{달걀}) * P(\text{라면}))$
  = $0.4 / ((5/10) * (8/10)) = 0.4 / (0.5 * 0.8) = 1$

2. 연관 규칙

· 향상도 (Lift)

- 향상도 = 1 : 서로 독립적 관계, 우연에 의한 연관성

- 향상도 > 1 : 양의 상관 관계

- 향상도 < 1 : 음의 상관 관계

3. 연관 규칙 시연

· 오라클 DBMS_DATA_MINING 패키지 이용

· 시각화는 오픈소스인 Apache Zeppelin

