**Report**

**데이터 마이닝에 대한 개념 파악하기**

**빅데이터 학과**

**20175327 이연희**

1) 데이터 과학(Data Science)이란 무엇인지 정의하세요.

데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야다.

데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다. 데이터의 구체적인 내용이 아닌 서로 다른 성질의 내용이나 형식의 데이터에 공통으로 존재하는 성질, 또는 그것들을 다루기 위한 기술의 개발에 착안점을 둔다는 특징을 가진다.

2) 데이터 과학과 데이터 마이닝(Data Mining)의 관계를 설명하세요.

데이터 과학(Data Science)은 일반적인 컴퓨팅 기술+분산 컴퓨팅 기술+데이터 마이닝+기계학습+데이터 분석(통계, 수학)에 해당하는 학술분야를 아울러 통칭하기 위해서 만들어졌다.

컴퓨터를 활용해서 어떤 가치를 찾는 작업이 분포를 통해서 통찰을 찾는 것에서 추가로 뭔가를 해야 한다면 자동화가 필요하다. 기계학습도 자동화의 일종이다. 데이터 마이닝 또한 데이터에서 가치를 찾는 작업을 일부 자동화한 것이라고 볼 수 있다. 자동화 기술은 구현된 이후에는 편하게 사용할 수 있지만 구현 작업 자체는 매우 어렵다. 기계학습은 빅데이터와 데이터 과학에 있어서는 필수라고 할 수 있으며 이때 기계학습 기술이 포함된 데이터 마이닝 기술 역시 데이터 과학에 있어서 필수라고 할 수 있다.

2) 분류(Classification) 알고리즘이 무엇인지 정의하고 사례 찾아보기

분류는 데이터 분석의 대표적인 기법으로서 데이터 class를 설명하는 모형을 제공해 준다. 데이터에 대한 class label이 없는 상태에서 자동으로 이들을 분류하는 것을 말한다. 분류가 범주형 class를 예측한다면, 반면에 예측은 연속형 속성에 대한 모형을 세움으로써 미래 데이터 경향을 예측하게 된다.

사례로는 “이 타이어는 다음 1,000마일을 가는 동안 터질까요? 예 또는 아니요? /어떤 방식이 더 많은 고객을 이끌까요? 5달러 쿠폰 또는 25% 할인?”등이 있다.

3) 변칙감지(Anomaly Detection, 이상감지) 알고리즘이 무엇인지 정의하고 사례 찾아보기

변칙감지 or 이상감지 알고리즘은 수많은 데이터들을 바탕으로 기존의 데이터들이 가지고 있는 특징들, 즉 '보편적'인 특징들을 갖지 않는 데이터들을 'Anomaly' 한 데이터라고 하는데, 바로 이러한 데이터들을 찾아내는 것을 말한다.

예시로는 “자동차에 압력 계기판이 있는 경우 이 압력 계기판에 정상적인 수치가 표시되는지 알고 싶을 수 있습니다. /인터넷을 모니터링 하는 경우 인터넷의 이 메시지가 일반적인지 확인하는 것이 좋습니다.”와 같이 변칙 감지는 예기치 않거나 비정상적인 이벤트 또는 동작에 플래그를 지정하고 어디서 문제점을 찾아내야 할지에 대해 찾을 수 있다.

4) 회귀(Regression) 알고리즘이 무엇인지 정의하고 사례 찾아보기

회귀 알고리즘은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링 등의 통계적 수치 예측에 이용될 수 있다.

예시로는 수치 예측을 수행할 때 회귀 알고리즘이 쓰인다. “다음 화요일의 기온은 얼마나 될까요? /4사분기 매출은 얼마나 될까요?”와 같은 수치를 요구하는 질문에 답변하는 데 도움이 된다.

5) 클러스터링(Clustering) 알고리즘이 무엇인지 정의하고 사례 찾아보기

주어진 데이터들의 특성을 고려해 데이터 집단(클러스터)을 정의하고 데이터 집단의 대표할 수 있는 대표점을 찾는 것이다. 클러스터란 비슷한 특성을 가진 데이터들의 집단이며 반대로 데이터의 특성이 다르면 다른 클러스터에 속해야 한다. 보다 쉬운 해석을 위해 데이터를 적절한 그룹으로 분리하는 것, 즉 유사한 속성을 가지는 데이터들을 묶어 전체를 분류기준 없이 그룹 또는 군집으로 나누는 작업으로 해석할 수 있다.

사례는 “어떤 시청자들이 같은 종류의 영화를 좋아하나요? /어떤 프린터 모델에서 동일한 방식으로 오류가 발생하나요?” 등이 있다.

6) 강화학습(Reinforcement learning) 무엇인지 정의하고 사례 찾아보기

강화 학습(Reinforcement learning)은 기계 학습의 한 영역이다. 행동심리학에서 영감을 받았으며, 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법이다.

강화 학습 알고리즘은 진행하면서 데이터를 수집하고 시행 착오를 통해 학습한다. 사례로는 “내가 주택의 온도 제어 시스템이라면 온도를 조절할까요? 아니면 현재 상태로 둘까요? /내가 자동 주행 자동차라면 노란색 신호등에서 브레이크를 밟을까요? 아니면 가속 페달을 밟을까요? /로봇 진공청소기의 경우 계속 청소를 할까요? 아니면 충전 스테이션으로 되돌아갈까요?”등이 있다.