출석수업 과제물(평가결과물) 표지(온라인제출용)

교과목명 : 데이터과학입문

학 번 : 202135-367895

성 명 : 김태정

강 의 실 : 지역대학 호

연 락 처 : 010-4172-4516

- 이하 과제 작성

1 번문제

빅데이터 시대의 데이터과학의 발전의 요인은 여러가지가 있다. 요인들을 크게 임의로 정리해보면 1 번째 데이터를 수집할 수 있으며, 2 번째 데이터를 가공할 수 있으며, 3 번째 데이터를 활용할 수 있다. 수집, 가공, 활용적인 측면에서 기존의 경우 항상 한가지씩은 부족한 면이 있었기에 발전할 수 없었다. 그러나 시간이 지나서 모든 부분에서 발전이 있었기에 데이터 과학은 더욱이 발전하게 될 수 있었다.

데이터 수집의 발전은 그 중에서 가장 원초적으로 중요한 기술이라고 할 수 있다. 가공 기술과 활용 기술의 경우에도 결국 데이터가 있어야 할 수 있기 때문이다. 생각보다 AI 기술들은 오래되었다. 머신러닝의 모태가 되는 퍼셉트론의 경우 1957 년 코넬 항공 연구소에서 개발되었으니 오히려 역사로만 따지면 활용기술들은 결코 최근 기술이 아니라고 할 수 있다.

문제는 데이터의 수집과 가공이였다. 데이터는 단순히 많다고 좋은 것이 아니라 실제 서비스에 사용될 수 있는 구조적 데이터(정형 데이터)가 중요하다. 과거에는 이 정형데이터를 만드는데 비용이 너무 많이 들었다. 비용이 많이 들은 이유는 정형데이터의 경우 처음부터 정형화된 방법으로 데이터를 수집하거나(조사 연구, 실험 등) 다량의 비정형 데이터를 정형화 시켜야하는데 막대한 빅데이터를 정형화 시키는 일도 쉽지 않았다. 또한 설령 기존의 방법들로 해결하려고 해도 많은 시간들이 걸리는

작업이였고 세상이 빠르게 변하는 현대에서 사용자의 니즈에 즉각적으로 반응하기에는 부족한 점이 많았다.

하지만 현대에 이 문제들을 해결할 수 있게 되었다. 먼저 가공 기술들의 변화이다. 2006 년에 하둡이 등장하면서 빅데이터들을 어떻게 분산 스토리지에 저장하여 사용할 수 있을지에 대한 방법이 제시 되었고 여러 하둡 에코시스템과 스파크의 등장으로 가공기술들은 비약적인 발전을 이룩하게 된다. 가공 기술들이 발전되었고 활용 기술은 그전부터 있었으므로 이제는 많은 정형 데이터를 저렴한 비용과 시간으로 얻을 수 있느냐는 것만 남게 되었다. 이는 시간이 지나면서 생각보다 쉽게 해결되었는데 IT 기업들이 늘어나고 플랫폼 기술들의 상용화가 지속되면서 실시간 생성 데이터와 로그성 데이터들이 비약적으로 쌓이기 시작하였다. 사람들의 로그인 정보를 바탕으로 해서 SNS 데이터들, 클라우드에 올라간 사진들을 비롯하여 사실상 모든 사람들이 들고 다니는 스마트폰의 GPS 와 카드 사용내역들의 데이터가 쏟아졌다. 또한 IoT 의 발전으로 챗봇 등의 대화 내역, 자동차나 운동기구들의 움직임 역시 확보할 수 있게 되었다. 과거에는 데이터를 얻기 위해서 여론조사나 실험들을 통하여 통제된 정형화된 데이터만 얻을 수 있었다면 현재는 이러한 비정형 시계열 데이터들을 확보할 수 있게 되었다. 현재 까지도 인간이 상상할 수 있는 대부부분의 방법으로 데이터를 수집하고 있다. 골프채 내부에도 가속도 센서가 들어가며 로봇 수술기구에도 동작 센서들로 데이터를 수집하고 있다. 물론 아직 도입 안한 분야들이 있겠지만 이러한 부분 역시 빠른 시간내에 매워질 것은 자명하다. 이러한 변화가 현재 빅데이터 시대에 데이터 과학이 한걸음 더 발전할 수 있게 되는 밑 바탕이 되었다.

데이터 수집의 미래에 대해서 확실히 나아갈 것 같은 부분은 “가상 비정형 데이터의 생성”이 있다고 생각한다. 기존에는 비정형 데이터를 수집했지만 앞으로는 비정형 데이터자체를 임의로 만들어 낼 수 있을 것이라고 예측해본다. 사회 조사 및 실험에서 아쉬운 점은 실세계에서 테스트 해볼 수 없는 일들이 아주 많다는 것이다. 가령 현재 상태에서 금리를 올리게 나은지, 아니면 금리를 내리는게 나은지는 실제로 해보기 전까지 알 수 없다. 그래서 인류는 항상 한가지 선택을 해왔다. 물론 긴 역사에서 두 사례가 다 존재하는 경우가 없는 것은 아니나 각 상황에 적용되는 요소(Factor)가 다르기에 일반화시키기가 쉽지 않다. 또한 역사를 관찰할 경우 과거의 생활상을 파악하는게 여간 어려운게 아니다. 그래서 과거의 인구나 물가 등을 추론하는 것은 쉽지 않다. 이러한 한계를 해결하기위해서 가상 세계를 만들고 거기서 시뮬레이션해서 비정형데이터를 추출하는 기술이 발전할 것으로 예상된다. 실제 메타버스의 가장 중요한

화두중의 하나가 베타버스 데이터 수집으로 이를 통하여 인류의 데이터과학의 미래가 한층 더 발전해 나갈 것으로 예상된다.

2 번문제

모수적 모형 접근방법

해당 방법은 입력값 x 와 출력값 y 가 주어졌을 때 해당 식을 y = a*x + b 꼴과 같이 모수(Parameter)와 관련된 식을 나타내어서 a, b 를 찾는 방법을 모수적 모형 접근방법이라고 한다.

이때 모수들을 찾는 방법은 과거 데이터로부터 적합한 값들을 이용해서 찾는데 대표적인 방법으로는 최소 자승법과 최대우도 추정법등이 있다. 최소 자승법은 다른 말로는 최소 제곱법으로 실제 y 값에 예상된 y 값(a*x + b)의 오차의 제곱합이 최소가 되는 해를 찾는 방법이 최소 제곱법이다. 최대 우도 추정법의 경우 우도(표본의 값이 등장할 확률)의 최대값을 구하는 것으로 나타낸다.

모형으로는 선형 회귀 모형과 로지스틱 회귀 모형이 존재한다. 선형 회귀 모형은 y = a*x + b 꼴로 직선으로 나타내며 -무한과 무한 사이의 값을 가지며 로지스틱 회귀 모형은 y = 1/(1+e^(-μ))로 나타내며 0 과 1 사이의 값을 가진다.

모수적 모형 접근방법의 경우 보다시피 단일 함수로 되어 있기에 데이터가 갖추어 져 있을 때 제일 먼저 시도해 볼 수 있는 방법이며 결과 역시 간단하게 나오기에 사용하기 쉽다. 복잡성이 높지 않거나 변수끼리의 상관관계가 명확하다면 시도해보기 좋은 방법이다.

반면 가정을 하기 힘들거나 식에 맞지 않는 데이터라면, 혹은 변수끼리의 상관관계가 명확하지 않다면 시도하기 힘들며 설령 시도하더라도 산출된 데이터가 정확도나 성능이 낮게 나올 수 있다.

알고리즘 접근방법

비모수적 접근방법이라고도 하며 입력값 x 와 출력값 y 가 주어졌을 때 해당 관계를 알기 어려운 상황에서 사용되는 방법들을 총칭해서 말한다. 즉 중간에 사용하는 모수들을 알수 없다. 모수들을 알 수 없기 때문에 중간에는 특정 알고리즘, 즉 모델들을 사용하여 값을 추론하게 된다. 이 때 알고리즘에 의해서 데이터를 학습하게 된다. 기존의 x 의 값에 y 라는 값들이 나왔다고 가정하면 새로운 x1 의 값이 들어갔을 때는 y1 이 나올거라고 예상하는 것이다.

해당 알고리즘 접근 방법으로는 의사 결정 트리나 신경망, 랜덤 포레스트등의 기법들이 사용되며 중간중간에는 모수적 모형 접근방법이 동원될 수도 있다.

알고리즘 접근방법에서는 데이터 복잡성이 높아도 사용 가능 하다는 장점이 있다. 다만 알고리즘에 기반해서 분석하기에 알고리즘 자체가 틀린다면 완전 잘못된 값을 도출될 수 있다. 또한 학습 데이터가 실제 세계의 방식과 똑같을 거라고 잘못 가정하는 과적합문제가 발생할 수 있다.

3 번문제

빅데이터는 많은 데이터를 수집하는 방법과 그 데이터를 가공하는 능력이 뒷받침 되어야만 활용될 수 있다. 현대에는 클라우드와 컴퓨팅 자원의 발전, 기법들의 발전으로 이 장벽이 해소되었다. 이러한 변화에 우리를 둘러싼 환경은 크게 바뀌기 시작했다.

빅데이터 시대의 도래로 인해 체감되는 가장 큰 변화는 “개인 맞춤 서비스”가 가능해졌다는 것이라고 생각한다. 예를 들어 빅데이터 시대 전에 광고들은 대부분 옥외 광고의 형태를 띄고 있었다. 이는 불특정 다수에게 공개하는 것이다. TV 광고는 어느정도 시청률을 바탕으로 타겟팅을 할 수 있었지만 사람의 취향이 다양했기에 완전한 맞춤형이라고는 볼 수 없었다. 하지만 빅데이터 시대로 넘어오면서 개개인의 검색 기록, 메신저 기록, 어플 사용 이력들의 데이터를 확보할 수 있게 되었고 이를 바탕으로 맞춤형 광고를 할 수 있게 되었다. 이러한 현상은 광고 분야에 국한된 것이 아니라 모든 분야에서 드러난다. 유튜브나 트위치등의 서비스에서 추천 역시 내가 시청했던 데이터를 바탕으로 비슷한 특징의 유저와 비교해서 추천해준다. 이러한 일들은 많은 데이터가 쌓여야만 할 수 있다. 빅데이터 시대가 되면서 이러한 장벽이 무너졌다.

또한 음성인식과 글쓰기 인식 등의 인식 분야에도 많은 변화가 이루어 졌다. 과거에는 이를 룰베이스로 해결하려고 하였다. 문제는 음성인식이나 글쓰기 인식률은 룰베이스로 하기에는 여러가지 난제들이 있었다. 사람들이 반드시 정자체로 글을 쓰지 않으며 사투리나 억양등이 일정하지 않는다는 문제점이 있었다. 이러한 문제를 잡아내려면 그에 합당한 많은 데이터가 필요하였다. 빅데이터가 수집, 가공이 활성화되면서 머신러닝과 딥러닝을 적용할 수 있게 되어 이 부분이 해소가 되었다. 이는 나비효과처럼 변화를 몰고와 인공지능 챗봇의 활성화를 불러오고 손이 자유롭게 되어 더 많은 작업을 할 수 있게되는 효과를 가져왔다.

여러가지 선순환을 가져오지만 개인의 사생활 정보를 수집해야 하는 아주 큰 문제가 있다. 빅데이터도 결국에 없는 데이터를 사용할 순 없으므로 데이터를 수집해야 하는데 이 방식이 SNS 나 검색기록 등 사생활 침해요소가 다분한 데이터들이 많다. 가령

사진인식율을 높이기 위해서 클라우드 내의 사진들을 들여 본다면 민감한 사진이나 사용자를 식별할 수 있는 사진을 들여다볼 수 있다는 것이다. 그리고 이러한 데이터를 만약 불순한 집단, 해커나 독재국가의 정부 등의 손에 들어간다면 아주 큰 문제를 일으킬 수 있다.

이러한 시류속에 데이터 과학자들의 책임은 더욱 무거워져 간다. 과거 데이터 과학자는 그냥 수학과 통계만 잘하면 됐다면 지금은 그 도메인 지식 자체를 이해하고 활용할 줄 알아야 하며 자신의 생각을 구현할 수 있는 정도의 최소한의 개발 지식을 갖추어야한다. 그리고 높은 수준의 윤리의식을 갖추어야한다. 자신이 다루는 데이터가 남들에게는 소중한 데이터임을 깨닫고 적법한 절차를 갖추고 보안에 신경 써서 다루어야 한다.

한국과 미국 모두 데이터를 다루는 플랫폼 기업들이 강세를 보이고 있다. 그 만큼 빅데이터의 중요성과 이를 다루는 기술이 주목받고 있다. 빅데이터 산업 및 이를 바탕으로 하는 플랫폼 산업은 인력이 전부인 우리나라의 미래 먹거리 산업이 될 것이다. 이러한 시대의 흐름에 힘입어 데이터 과학자는 앞으로 우리나라의 산업을 선도하고 더 편리한 세상을 맞이할 방향타 역할을 하게 될 것이다.

4 번문제

데이터 품질의 정의는 사전적 정의와 실무적 정의가 있다. 사전적 정의는 데이터를 사용하기에 적절한 환경을 구현하는 것이다. 즉 적합성, 적시성, 정확성, 완전성, 적절성, 접근 가능성을 의미한다. 데이터 품질의 실무적 정의는 지식 및 정보와 관련된 업무에 종사하는 사람들이 데이터를 활용하여 업무를 효과적으로 수행하기 위한 데이터의 기대수준이라 할 수 있다. 정의 자체는 아무래도 좀 딱딱한 면이 있지만 간략하게 설명하자면 활용이 용이할수록 높은 품질의 데이터라고 할 수 있다.

데이터를 수집하는 데는 데이터 과학자의 역할은 과거에 비해서 오히려 줄었다. 과거에 데이터를 얻는 방식은 실험과 조사에 기반을 두고 있었다. 이런 방식은 시간과 노력이 많이 들어가는 반면 얻어진 결과는 충분히 부호화 되며 통제되었고 사용하기 용이한 정형데이터의 형태로 나왔다.

반면 최근 데이터 수집에서 데이터 과학자의 역할은 줄어들었다고 생각된다. 이는 데이터 수집과정에서 데이터 과학자의 손길이 안 닿게 되는 경우도 있고 못 닿는 경우도 있다. 가령 이미 주어진 환경이라면 추가적인 데이터 수집은 어려울 것이다. 이부분에서는 어플리케이션 개발자와 기획자가 데이터의 수집에 더 연관이 있는 경우가 많다. 이러한 방식으로 얻게 되는 데이터는 비정형데이터로 로그성 데이터이며 결측치도

많고 속성이 부정확한 경우가 많다. 하지만 컴퓨터는 기본적으로 데이터를 정형화된 데이터만 사용할 수 있다. 비정형 데이터는 낮은 품질의 데이터와 궤를 같이한다고 할 수 있다. 이러한 데이터의 품질을 높이는 것은 중요한 과제라고 생각한다.

높은 품질, 좋은 품질의 데이터는 "충분히 가공된 데이터"라고도 할 수 있다. 그러나 현대에 빅데이터 시대가 도래하면서 많은 데이터를 얻게 되었지만 데이터의 품질은 낮게 되었다. 하지만 이러한 변화하는 흐름에도 불구하고 데이터 품질에 관해서 기업들이 크게 생각하지 못하는 경향이 있다. 이는 빅데이터의 흐름이 빠르게 변하는 것과 무관하지 않을 것이라고 생각이 든다. 과거의 방식들로 얻은 것은 처음부터 정형데이터에 가까웠기에 가공에 큰 공을 안 들였던 그 관성이 지금까지 남아 있는 것으로 판단된다.

잘못된 데이터로는 잘못된 값 밖에 얻을 수 없다. 그렇기에 빅데이터 시대에서는 데이터 품질의 관리에도 신경을 쓰는 것이 중요하다. 현존 시스템을 진단하고 우리의 도메인에 따라서 어떠한 데이터가 필요할지 생각을 해야 하며 이에 맞춰서 오류를 불러일으킬 값들을 제거한다. 또한 법과 규제에 맞춰서 추가적인 정보를 수집 혹은 제거를 해야 한다.

빅데이터 시대에 모든 사람들이 빅데이터를 모으는 것이 중요한 것은 알지만 이를 가공하는 것 역시 중요하다는 것을 간과하는 경우가 많다. 사실 시대를 막론하고 데이터는 사용할 수 있는 형태로 가공을 하는 것, 즉 품질을 유지하는 것은 중요하다. 다만 지금같이 낮은 품질의 데이터들이 범람하는 시기에 어떻게 하면 이를 보완해서 품질을 높일 수 있는 지에 대해서 많은 고민을 해야 한다.