데이터 사이언스 – 1

* **데이터 과학이 무엇인가?**
* 데이터를 분석하여 유용하고 새로운 정보, 지식을 추출하는 기술
* **데이터 과학이 어떻게 활용되는가?**
* 통계학, 인공지능, 기계학습이 융합된 분야
* 다른 분야와는 달리 기술 자체보다는 기술의 활용에 중점을 둔다.
* **데이터 과학의 두가지 분석방법인 기술 분석과 예측 분석이 어떻게 다른가?**
* **기술 분석** : 많은 양의 숫자나 문자로 표시되어 있는 데이터를 분석하여 그 속에 담겨있는 정보를 **사람이 알기 쉬운 방식으로 설명**해주는 것

(Ex) 1. 두반의 성적을 비교하기 위하여 각 반마다 시험점수의 평균을 계산

* 요약(abstraction) : 원래의 객체가 양적/질적으로 크고 복잡할 경우에 이 객체로부터 필요한 요소나 속성만을 추출하여 이 객체를 대표할 수 있는 것을 만드는 과정을 요약이라 한다

2. 일일의 기온 변화를 그래프로 플롯

* 그래프 : 꺾은선 그래프로 보여준다면 더 이해하기 쉬움
* **예측 분석** : 축적되어 있는 데이터를 기반으로 **미래를 예측할 수 있는 예측 모델**을 만드는 것

(Ex) 1. 오늘의 경제 상황을 기반으로 내일의 주식가격을 예측한다.

2. 생산 공정에서 발생하는 불량품을 검출한다.

3. 알파고

* **간단한 사례를 통해 예측모델 설명**
* 소쩍새가 울면 풍년이 든다 등등 : **직관력이나 통찰력에 의존**한 예측 분석
* 데이터과학에서는 컴퓨터와 분석알고리즘을 과거의 사례-데이터를 적용하여 상황과 결과를 연관짓는 **패턴을 발견하여 예측모델을 생성**한다.
* 패턴 : 반복적으로 발생하는 현상으로써, 원인과 결과를 연결 짓는 관계를 말한다.
* 데이터과학에서 모델을 만들기 위한 과정을 **데이터 마이닝 과정**이라 한다.
* **데이터 마이닝 과정** : 여러가지 분석 알고리즘을 활용하여 **축적된 데이터를 분석하고 명시적인 모델을 생성하는 체계적인 과정**을 말한다.
* **예측 모델의 학습에 사용되는 데이터를 ‘학습용 데이터셋’ / ‘트레이닝 데이터셋’ 이라 부른다.** 이 때, 사람이 이해할 수 있는 명시적인 규칙 형태로 패턴을 만들기도 하지만, 수학적이거나 컴퓨터적인 데이터 구조를 이용하여 패턴을 만들기도 한다.
* 모델이 ‘학습용 데이터셋’에 대하여 학습을 마치고 난 다음에는 모델을 활용할 차례인데, 이 때, **완성된 모델에 새로운 사례를 입력하여 모델이 새로운 사례에 대한 결과를 예측하게 하는 것**이다.
* 여기에서 **예측모델의 입력에 사용되는 새로운 사례는 용돈을 요구할 당시의 상황만 들어있을 뿐이고, 시도의 결과는 들어있지 않다.**
* 우리의 용돈 예측 모델은 **결과에 대한 예측 값으로 성공 또는 실패라는 값**을 내놓게 된다.
* 모델을 완성한 뒤에는 모델의 성능을 평가할 수 있다. **모델의 성능평가를 위하여 사용되는 데이터 셋을 ‘시험용 데이터 셋’**이라 부른다.
* 학습이 끝나서 완성된 예측 모델은 공학적인 관점에서 입출력을 가지는 하나의 시스템으로 생각할 수 있다. 시계이(가) 표시된 사진

  자동 생성된 설명
* 입력변수를 통하여 새로운 사례를 예측 모델(시스템)에 입력하면 예측 모델(시스템)은 출력변수를 통하여 이 사례에 대한 예측값을 출력한다.
* 예측모델은 시스템의 형태 즉, 수학의 함수나 프로그램의 형태로 제작한다.

**∴ (정리)**텍스트, 표지판이(가) 표시된 사진

자동 생성된 설명

* 데이터 과학에서 예측 모델과 관련된 분석 방법을 ‘예측 분석’이라 하고, 예측 분석은 두 단계로 나뉘어진다.

1. 모델 생성 단계 : 과거의 사례를 예측 모델이 학습하는 단계

* 모델이 과거의 데이터를 학습하여 지식을 추출

\*데이터란 과거에 실제 발생했던 사실/사례들을 의미한다.

\*지식이란 사물 혹은 사건들간의 관계를 말한다.

* 논리학에서 여러 개의 개별적 사례들로부터 일반적인 규칙을 추출해내는 과정과 상응한다. = Induction

1. 모델 적용 단계 : 학습이 끝난 예측모델에 새로운 사례를 입력하여 모델이 새로운 사례에 대한 결과를 예측하도록 하는 단계

* 알아낸 지식을 적용하는 단계
* 논리학에서 일반적인 규칙들로부터 개별적인 사례들을 유추하는 과정과 상응한다. = Deduction
* **예측 모델의 성능측정**
* 모델을 완성한 다음에는 모델의 성능을 평가할 수 있다. 여러 개의 모델중에 최적의 성능을 가지는 모델을 선택해야 한다.
* 일반적으로 예측모델의 성능은 실행속도 보다는 **예측의 정확도를 기준으로 평가**한다.
* 개체, 시계이(가) 표시된 사진

  자동 생성된 설명

예측정확도는 시험용 데이터셋에 들어있는 사례들을 모델에 적용한 후 시험용 사례들 중 정분류된 사례들의 비율로 계산할 수 있다. 이 비율을 % 또는 [0, 1] 범위의 소수로 나타내는 것이 일반적이다.

\*정분류 : 데이터셋에 실제 기록된 목표변수값과 모델이 계산한 예측값이 동일한경우 (**실제 목표변수값 = 예측된 목표변수값**)

\*오분류 : 그렇지 않은 경우를 의미한다.

* **데이터와 데이터에 관련된 용어 설명**
* 데이터란 과거의 사례들을 표의 형식으로 정리하여 모아 놓은 것을 말한다.
* 행 : 사례 = 레코드, 객체, 데이터포인트, 벡터
* 열 : 변수 = 속성, 필드, attribute
* 데이터에서는 결측치(빠진 값)가 있을 수 있다.
* 컴퓨터과학은 융합 분야이므로 여러 분야에서 사용하던 용어를 그대로 이어받아 사용하는 경우가 많아서, 하나의 사물을 여러 개의 다른 이름으로 부르기도 한다.
* 예측 분석에 사용되는 데이터의 경우 변수들을 크게 두 가지 종류로 나눌 수 있다.
* 예측(독립)변수 : 예측모델의 입력으로 사용되는 변수
* 목표(종속)변수 : 예측모델이 출력하는 변수 / 예측하고자 하는 변수
* 반드시 하나의 목표변수가 정해져 있어야 한다.
* 목표변수 외의 다른 변수들은 예측변수가 된다.
* 완성된 예측 모델을 활용하는 단계에서 예측 모델의 입력되는 사례에 대해서는 목표 변수값을 모르는 경우가 있는데, 예측모델의 예측 변수값을 입력해주면 예측모델은 이 값을 재료로 하여 목표변수값에 대한 예측값을 계산하여 준다.
* 예측분석에 주어진 데이터셋을 두 가지 셋으로 구분한다.
* 학습 데이터 셋 : 모델을 생성하기 위한 재료로 사용되는 데이터 셋
* 시험 데이터 셋 : 학습이 완료된 모델의 성능 시험을 위하여 사용하는 데이터 셋
* 예측 모델의 정확한 성능 평가를 위하여 두 셋이 중복되지 않도록 구성해야 한다.
* 변수의 값이 **종류를 나타내느냐 또는 양을 나타내느냐에 따라 범주형과 수치형의 두종류**로 나눌 수 있다.
* **범주형(명목형)** : 변수값이 **종류 또는 이름을 나타낸다.** 따라서, 변수가 가질 수 있는 값이 **유한**하다.
* **수치형** : 변수값이 **양, 정도를 나타낸다.** 일반적으로 **정수, 실수 등의 숫자로 표현**이 된다. 따라서, 변수가 가질 수 있는 값이 **무한**하다.

\***이 때, 숫자로 표현되어 있다고 해서 모두 수치형은 아니다.** 만약, 직업변수를 나타내기 위하여 각각 다른 직업을 1,2,3의 정수로 나타냈다면, 이 숫자는 직업의 종류를 나타내므로 직업변수는 범주형이라고 할 수 있다.

* 정형 데이터 : 표의 형식으로 나타내어진 데이터
* 비정형 데이터 : 텍스트, 영상, 사운드 등의 멀티미디어 데이터 / 문서, 웹페이지, 시스템 및 웹로그, 복합 데이터 – 전문적인 처리 과정을 거쳐 정형 데이터 형식으로 변환하여 처리