# Machine Learning

2019년 3월 11일 월요일 오후 6:36

1. **Defining the problem statement**
2. **Collecting the data**
3. **Exploratory data analysis**

    - **데이터 종류 :**

        - 데이터 집합(data set) : 데이터 개체(data object)들의 집합
          ※ 데이터 개체(data object)는 레코드(record), 점(point), 벡터(vecotr), 패턴(pattern), 사례(case), 사건(event), 샘플(sample), 관찰(observation), 개체(entity) 등으로 불리기도 한다.
          데이터 개체는 여러 개의 속성(attribute)으로 기술되는데, 속성이란 데이터 개체들 사이의 차이를 규정할 수 있는 특성이나 특징을 의미한다.
          이러한 속성은 변수(variable), 특성(characteristic), 필드(field), 특징(feature), 차원(dimesion) 등으로 불리기도 한다

          **(1) 질적자료(정성적자료, Qualitative, Categorical)**
          : 범주 또는 순서 형태의 속성을 가지는 자료
          a. 범주형(명목, nominal) 자료: 사람의 피부색, 성별
          b. 순서(서수, ordinal) 자료: 제품의 품질, 등급, 순위

          **(2) 양적자료(정량적자료, Quantitative or Numeric)**

: 관측된 값이 수치 형태의 속성을 가지는 자료

a. 범위형(interval) 자료: 화씨, 섭씨와 같이 수치 간에 차이가 의미를 가지는 자료.

b. 비율(ratio) 자료: 무게와 같이 수치의 차이 뿐만 아니라 비율 또한 의미를 가지는 자료

## 4. Feature engineering

- **데이터 결측치 처리방법1 - 삭제**
  결측값이 발생한 모든 관측치를 삭제하거나 데이터 중 모델에 포함시킬 변수 들 중 관측값이 발생한 모든 관측치를 삭제하는 방법이 있습니다. 그러나 전체삭제 또는 부분삭제는 실제 예측에 영향을 주는 데이터일 경우 Cost에 영향을 미칠 수 있습니다.
  그렇기 때문에 삭제는 결측값이 무작위로 발생한 경우에 사용합니다. 결측값이 무작위로 발생한 것이 아닌데 삭제할 경우 왜곡된 모델이 생성될 수 있습니다.

- **데이터 결측치 처리방법2 - 다른 값으로 대체**
  결측값이 발생한 경우 다른 관측치의 평균, 최빈값, 중간값으로 대체할 수 있습니다. 결측 값이 발생이 다른 변수와 관계가 있는 경우 유용하짐나 그렇지 않은 경우 모델이 왜곡될 가능성이 존재합니다.
- **데이터 결측치 처리방법3 - 예측값 삽입**
  결측값이 없는 관측치를 트레이닝 데이터로 사용해서 예측모델을 만드는 방법입니다. 예측하는 방법은 Regression이나 Logistic Regression을 주로 사용합니다.

- **이상 데이터 처리방법1 - 단순삭제**
  이상데이터가 실수로 발생한 경우에는 해당 값을 삭제하면 됩니다. 예를 들어 단순 오타나 비현실적인 응답 등입니다.
- **이상 데이터 처리방법2 - 다른 값으로 대체**
  절대적인 관측치의 숫자가 작은경우 삭제를 할 경우 관측치의 절대량이 작아지는 문제가 발생합니다. 이럴 경우 결측치 데이터 처리 방법과 같이 다른 값(평균)으로 대체하거나 유사하게 다른 변수를 사용해 예측 모델을 만들어서 사용하는 방법이 있습니다.

- **Feature Engineering**이란, 기존 변수를 사용해서 데이터에 정보를 추가하는 과정입니다. 새로 관측치나 변수를 추가하지 않고도 기존의 데이터를 보다 유용하게 만드는 방법입니다.

- **Feature Engineering 방법1 - SCALING**
  변수의 단위를 변경하고 싶거나, 변수의 분포가 편향되어 있을 경우, 변수 간의 관계가 잘 드러나지 않는 경우에는 변수 변환의 방법을 사용합니다. 방법으로는 Log 함수를 사용하거나 Square root를 사용하는 방법이 있습니다.

- **Feature Engineering 방법2 - BINNING**
  연속형 범주를 범주형 변수로 만드는 방법입니다. 예를 들어 시간 데이터가 수치로 존재하는 경우, 이를 3시간 미만, 4시간~5시간 식으로 범주형으로 변환하는 것입니다. 특별히 정해진 방법이 있는 것이 아니라 분석하는 사람에 따라 다르게 할 수 있습니다.

- **Feature Engineering 방법3 - DUMMY**
  마지막으로 Binning과 반대로 범주형 변수를 연속형 변수로 변환하기 위해 사용하는 것입니다.
  지금까지 데이터 분석을 위한 전처리에 대해 간단하게 살펴보았습니다. 요약하자면 데이터를 분석하여 적절할 모델링을 위해 좋은 데이터를 만드는 과정이라고 볼 수 있습니다. 위 내용만으로 모든 것을 이해할 수 는 없겠지만 아래 자료를 보시면 실제 전처리를 어떻게하는지 감을 잡으실 수 있을 거라 생각합니다.

## 5. Modelling

## 6. Testing

a. **classifier :** 어떤 머신러닝 **알고리즘**을 선택할 것인가.
b. **fit( data , target ) :** 입력 데이터(data)와 결과(target) 과의 **관계**를 위의 알고리즘을 통해 구한다.
c. **predict( example )** : 위에서의 알고리즘을 통해 **예상 결과**를 얻는다.
d. **accuracy ( examples_label , results ) :** 실제 답과 위의 결과값의 **일치 정도**를 구하여 모델을 평가한다.