

1장. 통계적 추론이란?

◆ Introduction

미국 자동차 보험에서는 차량의 색상까지도 고려하여 보험료를 책정한다. 일반적인 생각으로 자동차 색상이 화려할수록 눈에 띄어 사고가 덜 발생할 거 같다. 하지만 파랑색, 녹색 차량이 전체 자동차 사고 중 45%를 차지한다. 우리는 이러한 사실을 통계적인 추론 없이는 알 수 없다. 통계적 추론은 우리에게 주변에서 일어나는 상황에 대해 새로운 통찰력을 가져다주며, 예측을 하는 데 도움을 주기도 한다. 예를 들어 보험회사가 건강보험 상품의 과거 보험금 지급액을 성별, 연령, 상해급수에 따라 나열하면 각 항목에 따른 보험금 지급액 패턴을 파악할 수 있다. 그 후 보험계약자 특성(성별, 연령, 상해급수)을 입력하면 이에 따른 보험금 지급액을 예측할 수 있다. 보통 이는 일반화선형모형(Generalized Linear Model)를 통해 구해진다. 일반화선형모형은 통계적 추론(Statistical Learning)에서 쓰이는 도구 중에 하나다. 이번 장에서는 통계적 추론에 대한 기본개념을 먼저 간단히 알아보도록 하겠다.

1. An Overview of Statistical Learning

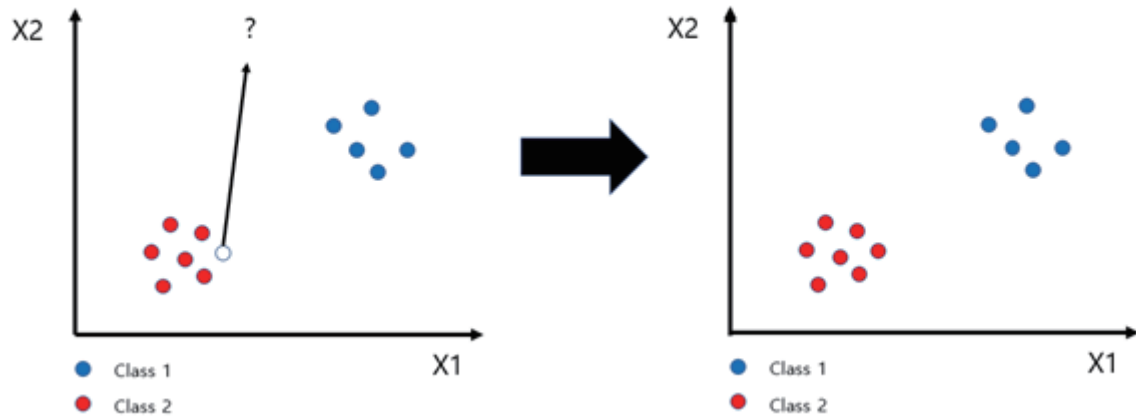
Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised. Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. With unsupervised statistical learning, there are inputs but no supervising output.

(James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013, p. 1)

2. 지도학습(Supervised Learning) vs 비지도학습(Unsupervised Learning)

통계적 추론에서 학습의 종류는 크게 지도학습과 비지도학습으로 나뉜다. 지도학습의 경우, 입력값(inputs)들에 대응되는 출력값(outputs)들이 존재하는 샘플들을 통해 학습에 사용되지 않은 입력에 대응하는 출력을 예측하는데 목적을 둔다. 이는 크게 분류(Classification)와 회귀(Regression)의 방법론이 있다. 지도학습과는 다르게 비지도학습은 입력들만 존재하고 대응되는 출력들은 존재하지 않는 샘플들을 통해 데이터의 구조를 탐색한다. 이제 그림을 통해 지도학습의 예시를 살펴보자.

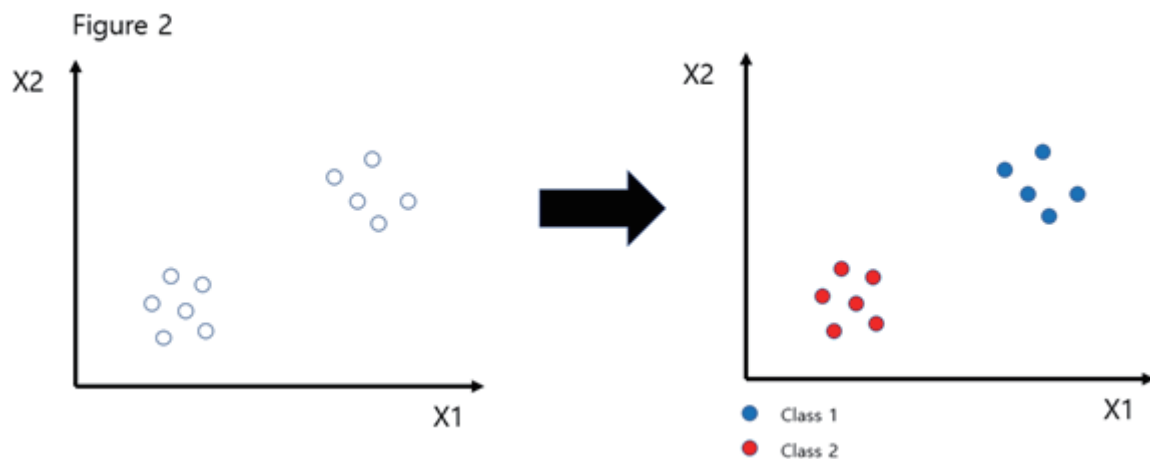
1장. 통계적 추론이란?



[그림1] 지도학습을 통한 Prediction

그림 1을 보면 입력들 X_1, X_2 에 대응되는 출력들 Class 1, Class2들이 명시되어 있다. 이 샘플들을 통해 학습에 사용하지 않은 하얀색 점을 예측할 경우에 Class 2에 속한다는 결과를 얻었다. 이는 위에서 설명한 지도학습 내용과 정확히 일치한다.

이제 그림 2를 통해 비지도학습의 예시를 살펴보자.

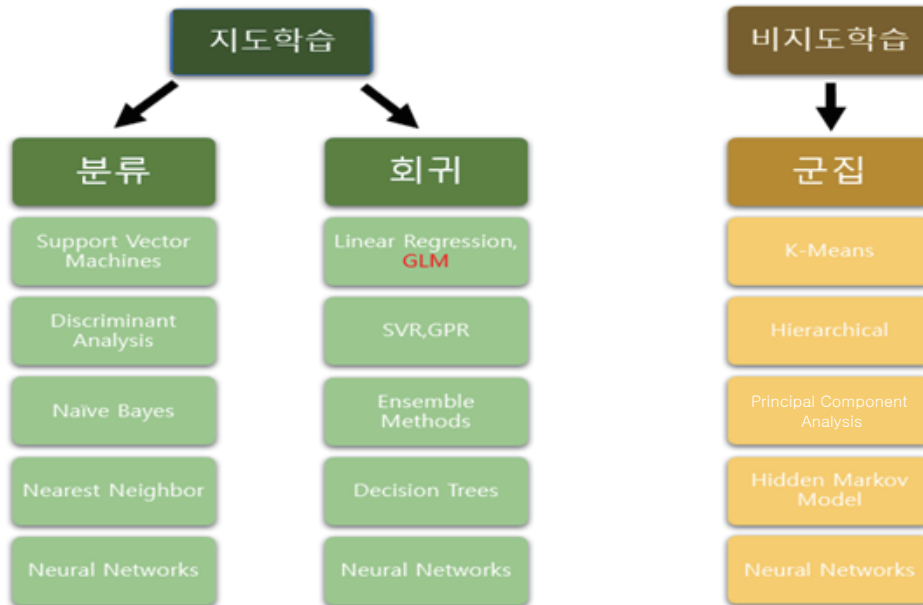


[그림2] 비지도학습

그림 2을 살펴보면 입력들 X_1, X_2 에 대응하는 출력들이 샘플에 존재하지 않는다. 선택한 군집의 알고리즘을 통해 주어진 데이터의 구조를 Class 1과 Class2로 구분하는 결과를 얻었다. 이는 앞서 말한 비지도학습의 설명과 일치한다.

1장. 통계적 추론이란?

그렇다면 일반화선형모형은 지도학습과 비지도학습 중에 어디에 속하는지 구별할 수 있겠는가? Introduction의 예시에서 성별, 연령, 상해급수와 같은 입력값에 대해 보험금 지급액이라는 출력이 존재하였다. 또한 이를 통해 보험회사는 보험계약자의 특성에 따른 보험금 지급액을 예측하는 것을 목적으로 했다. 즉, 일반화선형모형은 지도학습의 방법론 중에 하나이다. 이는 분류, 회귀, 그리고 군집에 방법론의 종류들을 보여주는 그림 3에서 다시 한번 확인 가능하다.



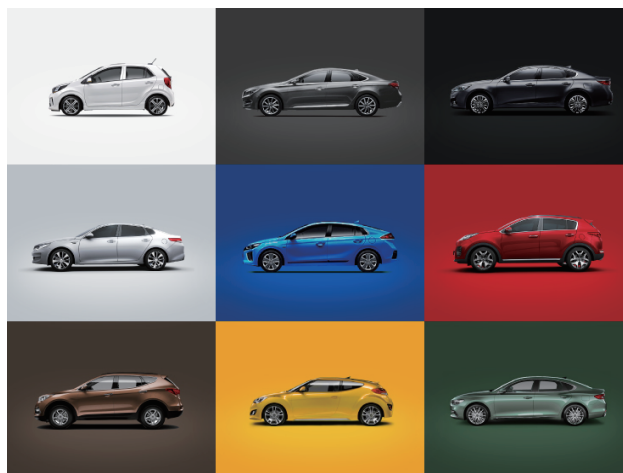
[그림3] 지도학습과 비지도학습의 방법론

3. 예측(Prediction) vs 추론(Inference)

Introduction의 보험금 지급액 예시에선 일반화선형모형을 사용한다고 가정하여 지도학습 방법론의 예측에 관한 쓰임만이 강조되었다. 하지만 지도학습은 차량의 색상에 따른 사고발생여부 예시처럼 추론을 목적으로도 쓰인다. 예측이 학습에 사용되지 않은 입력들에 대한 출력을 구하는 과정이라면 추론은 입력들과 출력에 상관관계에 목적을 둔다. 차량의 색상과 사고 발생 여부의 상관관계에 관심을 두는 과정이 추론이다. 상관관계는 인과관계를 포함한다는 것을 유의해야 한다. 즉, 입력변수와 출력변수 간에 높은 상관관계가 있다고 하여 두 변수가 인과관계에 있다고 결론을 내릴 수가 없다. 예를 들어, 해변에 상어 출몰률과 해변에서의 아이스크림의 매출액의 상관관계는 높게 나온다. 하지만 이는 두 변수 간 인과관계로 인한 것이 아니라 여름이라는 공통된 변수에 의해 영향을 받기 때문이다.



★ 쉬어가는 페이지 ★



사진출처 : HMG 저널

자동차 색에 따른 사고발생률 차이

파란색 가장 위험, 노란색 가장 안전

파란색 차의 교통사고 발생률이 다른 색에 비해 월등히 높다는 결과가 나왔다. 최근 경찰청은 이러한 내용을 담은 ‘자동차 색상에 담긴 불편한 진실’이라는 자료를 공개했다.

경찰청이 자동차 색상에 따른 사고율을 조사해 본 결과, 파란색 차량에 의한 사고가 전체의 25%를 차지하는 것으로 나타났다. 그 다음으로는 녹색 20%, 회색 17%, 흰색 12%, 빨간색 8%, 검정색 4%, 밤색 3%, 황금(노란)색 2%, 기타 9% 등으로 조사됐다.

문제는 소비자들이 선호하는 차량의 색과 사고발생률이 높은 색이 일정부분 일치 한다는 것이다. 참고로 지난해 한국 소비자가 가장 선호한 차의 색깔은 은색(30%), 흰색(25%), 검정색(15%), 회색(12%), 파란색·빨간색·갈색·베이지색(각 4%), 녹색 1%, 기타 5% 순으로 조사된 바 있다. 인기가 많은 흰색, 회색계열 차량은 그만큼 주의가 필요한 것이다.

이처럼 차량 색깔별로 사고율에 차이가 있는 것은 눈의 굴절률과 초점기능 때문이다. 예를 들어 파란색의 경우 빛의 굴절률이 커 망막보다 앞쪽에 상이 맺히는데, 이때 망막은 초점을 맞추려고 수정체를 오목하게 만든다. 즉 파란 물체는 실제보다 더 멀게, 더 작게 느껴지게 되는 것이다.

그에 비해 황금(노란)색은 색채 중에 가장 크게 보이는 성질을 가지고 있고, 망막에 정확히 상이 맺히기 때문에 사고가 적다고 경찰청은 설명했다. 어린이 유치원차량이나 스쿨버스 등이 노란색인 이유가 바로 여기에 있는 것이다.