# D&A M/L Session 1주차 과제 리포트

빅데이터 경영통계전공 20192774 신기섭

1.편향과 분산이 동시에 줄어들 수 없는 이유(Bias-Variance Tradeoff)를 조사해서 리포트 작성

Bias-Variance Tradeoff를 설명하기 전에 먼저 학습 알고리즘의 에러는 총 3가지로써 noise, bias, variance로 구성되어 있다. 먼저 noise란 데이터의 본질적인 에러로 우리가 수정할 수 없는 에러를 뜻한다. 다음으로 bias와 variance는 모델에 따라 변화하는 에러로써 우리는 이 두가지를 조정함으로써 에러를 줄여간다. 우리가 이 두가지를 고려해서 에러를 줄이는데 이때 발생하는 문제가 바로 Bias-Variance Tradeoff이다. 이를 설명하기 전에 먼저 bias 와 variance를 설명하자면 bias는 데이터의 모든 정보를 고려하지 않으면서 학습이 이루어지는 것을 말한다. 이 bias를 조정하면 알고리즘의 평균 정확도가 변화한다. Variance란 bias와는 다르게 데이터의 에러까지 학습을 하는 것이다. 이를 조정하면 특정 데이터에 알고리즘이 얼마나 민감한지를 알 수가 있다. 위 설명에도 알 수 있듯이 bias와 variance는 서로 반대된 성향을 가지고 있다. Bias를 줄이기 위해 더 많은 데이터를 고려하게 되면 오히려 데이터의 에러까지 고려를 하게 되면서 variance가 높아지게 된다. 반대로 variance를 줄이기 위해 데이터의 학습량을 줄이게 되면 bias가 올라가게 되는 딜레마에 빠지게 된다. 이 수식은 에러를 식으로 표현한 것이다. 이 수식에서 f는 데이터를 완벽히 표현하는 함수이고 y는 우리가 학습시키는 모델을 의미하는데 E{y}를 f와 같게 하면서 bias를 줄이게 되면 variance가 noise와 같게 된다. 반대로 variance를 0으로 만들면 bias가 크게 증가하게 될 것이다. 따라서 우리는 bias와 variance를 동시에 줄일 수는 없는 것이다.

$$\underbrace{\mathbb{E}\left\{(f-\mathbb{E}\left\{y\right\})^2\right\}}_{\text{bias}^2}+\underbrace{\mathbb{E}\left\{(\mathbb{E}\left\{y\right\}-y)^2\right\}}_{\text{variance}}+\underbrace{\mathbb{E}\left\{\epsilon^2\right\}}_{\text{noise}}$$

2.지도학습(회귀/분류)의 모델과 모델에 대한 간단한 설명이 포함된 리포트 작성

지도학습의 회귀 모델에는 선형 회귀 분석, 결정 트리 모델, 랜덤 포레스트 모델 등이 있다. 선형 회귀 분석이란 회귀 형태의 데이터들을 선형 예측 함수를 이용해 직선 또는 곡선의 형태로 예측하는 것이다. 다음으로 결정 트리 란 데이터를 여러 기준을 정해 분류를 해서 비슷한 분류끼리 따로 모아 조합을 만드는 것이다. 맨 처음에 초기에 분류를 root node(뿌리 마디)라 하고 중간 마디를 intermediate node라 한다. 이러한 분류를 거치면 여러 개수의 terminal node(끝 마디)가 생성이 되는데 각각의 끝마디에는 비슷한 종류의 데이터들끼리 모여 있다. 모든 끝 마디의 데이터 개수를 총합하면 뿌리 마디에 있는 데이터의 개수와 동일한데, 이는 어느 한 개의 데이터가 여러 끝마디에 중복해서 들어가지 않는다는 것이다. 이 모델이 학습을 하는 과정에는 재귀적 분기와 가지치기가 존재한다. 다음으로 랜덤 포레스트 모델은 결정 트리를 여러 개를 동시에 적용해 학습 성능을 높이는 방법이다.

다음으로 지도학습의 분류 모델에는 k-최근접 이웃, 로지스틱 회귀, 서포트 벡터 머신 등이 존재한다. 먼저 k-최근접 이웃 이란 데이터를 가장 예측할 때 가장 가까이에 있는 데이터가 무엇인

가를 중심으로 데이터의 종류를 정하는 알고리즘이다. 여기서 k 가 의미하는 것은 주변에 몇 개의 데이터를 보고 판단할지에 대한 숫자이다. 이 알고리즘에서의 핵심은 이 k를 어떠한 숫자로 정하는 지다. 다음으로 로지스틱 회귀 란 선형회귀의 방법을 범주형 데이터를 대상으로 이루어지는 기법이다. 마지막으로 서포트 벡터 머신 이란 두 카테고리를 정한 다음 데이터를 분석해 해당 데이터가 어떠한 카테고리에 들어가는지 파악하고 데이터를 분류하는 모델이다.