질문1) 유연한 통계학습방법과 덜 유연한 방법중 어느것이 일반적으로 성능이 더 나은지 말해보자. \*



a) 유연한 방법이 일반적으로 성능이 더 나음. 데이터가 많기 때문에 과적합의 위험이 적어짐.
b) 덜 유연한 방법이 일반적으로 성능이 더 나음. 데이터양이 적기때문에 과적합의 위험이 크고 모델이 노이즈의 영향을 받을 가능성이 높음.

질문2) 각 시나리오가 분류 혹은 회귀문제 인지 설명하고, 가장 관심 있는것이 추론인지 혹은 예측인지 명시해보자. n과 p 또한 생각해보자.



1. 회귀/추론/n = 500, p = 3 (회사의 이익, 직원수, 업종)
2. 분류/예측/n=20, p = 13 (판매가격, 마케팅 예산, 경쟁사 가격, 10개 다른 변수 등)

질문3) 회귀 또는 분류에 대한 기법 중, 매우 유연한 방법이 덜 유연한 방법에 비해 가지는 장점과 단점은 무엇인가? 어떤 경우에 더 유연한 기법을 선호하고, 또 어떤 경우에 덜 유연한 기법을 선호하는지 생각해보자.

1. 유연한방법과 덜유연한 방법은 편향과 분산의 trade off가 있음.
유연한 방법은 편향이 낮아 성능이 좋아지나 분산이 커져 과적합의 위험이 있음
덜유연한 방법은 분산은 낮지만 편향이 높아져 과소적합의 위험(성능 low)의 위험이 있음
2. 유연한 기법이 선호 될 때: 데이터가 비선형의 관계를 갖고있으며 데이터의 양이 많을 때.
덜유연한 기법이 선호될 때: 데이터가 선형의 관계를 갖고 있으며, 데이터의 양이 많지 않을 때

질문4) 통계학습 기법에서 모수적 기법과 비모수적 기법의 차이는 무엇인가? 회귀 또는 분류에서 모수적 기법과 비모수적 기법의 장점과 단점은 무엇인가?

통계학습의 목적은 f를 통해 예측 f를 추정하는 것이며, 최대한 정확한 f를 찾을 때 사용되는 대부분의 모델링은 2가지 기법에 근거함.

모수적 기법: (회귀)
기본적으로 f는 데이터에 대해 선형적이라고 가정하며 함수의 형태를 가정함.
이 함수의 형태를 기반하여 파라미터(=가중치, 회귀계수)를 추정함.

장점: 심플, 빠르고 데이터가 선형관계가 있다면 성능도 좋게 나옴. 해석력이 좋음.

단점: f가 실제 형태와 보통 맞지않을 것(대부분 선형관계를 띄고있지 않음)이므로 추정 정확성이 낮고 결론적으로 정확도가 떨어짐

비모수적 기법:

f의 함수 형태에 대해 가정하지 않기 때문에 더 넓은 범위의 f형태에 적합될 가능성이 있음.

장점: 비선형모델에 성능이 잘나오기 때문에 실제 환경에서 많이 적용됨

단점: f에 대한 정확한 추정을 위해 많은 데이터 양이 필요하며 과적합 문제가 존재하며 해석력이 낮음.

질문5) 80페이지의 표 3.4에 주어진 p값에 대응하는 귀무가설을 기술하고 어떠한 결론을 내릴수 있는지 설명하여라. 설명은 sales, TV, radio, newspaper에 대하여 해야 한다.



**[회귀식]**

sales = 2.939+0.046\*TV+radio\*0.189-0.001\*newspaper

1. Tv, radio의 p값이 0.05보다 작으므로 귀무가설(=변수들이 sales에 영향을 주지 않는다는 가정)을 기각함. 따라서 TV는(unit이 1 증가할때) +0.046만큼 sales에 영향을 주며 radio(unit 1증가시) +0.0189만큼 sales에 영향을 주고 이는 통계적으로 유의미함.
2. Newspaper의 pvalue는 0.05보다 크므로 귀무가설을 기각하지 못함.

질문6) KNN분류기와 KNN 회귀방법 사이의 차이를 설명하여라

KNN분류기는 K개의 가장 가까운 이웃들 중에서 발견된 가장 일반적인 그룹을 사용해 정성적인 그룹(qualitative)으로 분류 함.
KNN 회귀방법은 K개의 가장 가까운 이웃 결과를 평균하여 정량적인(quantitative) 추정을 함.