[Data Science] 데이터 사이언스 개념 - 3.과적합과 모델 선택 — 나무늘보의 개발 블로그

노트북:	첫 번째 노트북
만든 날짜:	2021-01-16 오후 5:01
URL:	https://continuous-development.tistory.com/212

Data Science

[Data Science] 데이터 사이언스 개념 - 3.과적합과 모델 선택

2021. 1. 9. 04:29 수정 삭제 공개

3.과적합과 모델 선택

1.기댓값과 분산

기댓값 - 확률변수가 취하는 값을 확률로 가증치를 둔 평균값
확률변수의 기댓값은 다음처럼 정의 된다.

이산형 확률변수의 기대값

$$E(X)=\sum_{k=1}^{n}x_kP(X=x_k)$$

연속형 확률변수의 기대값

$$E(X)=\int_{-\infty}^{\infty}xf(x)dx$$

E[g(X)] = $\int_{-\infty}^{\infty}g(x)f(x)dx$

$$E(\bar{X})=\mu \qquad Var(\bar{X})=\frac{\sigma^2}{n}$$

$$E(\hat{p})=p \qquad Var(\hat{p})=\frac{p(1-p)}{n}$$

분산- 확률 변수가 취하는 값이 어느정도 퍼져 있는지 나타낸 것

$$\mathrm{Var}[X]=\mathrm{E}[X^2]-(\mathrm{E}[X])^2=\mathrm{E}[X^2]-\mu^2$$

$$\mathrm{E}[X^2]=\mu^2+\mathrm{Var}[X]$$

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[(X-\mu)^2] \\ &= \mathrm{E}[X^2-2\mu X+\mu^2] \\ &= \mathrm{E}[X^2]-2\mu\mathrm{E}[X]+\mu^2 \\ &= \mathrm{E}[X^2]-2\mu^2+\mu^2 \\ &= \mathrm{E}[X^2]-\mu^2\end{aligned}$$

평균이나 분산은 주사위의 예처럼 이산변수(서로 분리된 정해진 단위의 값만 갖는 변수)가 아니라, 연속변수로도 정의할수 있다.

2.배리언스(variance) = 분산

분산 - 흩어진 정도를 평가하기 위해 학습 데이터간의 분산을 계산 한 것

$$
\begin{aligned}
\operatorname{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 + \mathrm{E}[X]^2 - 2\mathrm{E}[X]X\right] \\
&= \mathrm{E}\left[X^2\right] + \mathrm{E}[X]^2 - 2\mathrm{E}[X]\mathrm{E}[X] \quad \text{(by the linearity of the expected value)} \\
&= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2
\end{aligned}
$$

3.편향-분산 분해

어떤 평가점 X0에 대해 시험데이터 상의 기대 평균제곱오차는 반드시 다음 3가지로 분해 할 수있다.

편향의 제곱

배리언스

오차항의 분산

여기서는 1과 2가 바꼈다.

$$
\begin{aligned}
\mathbb{E}[(y_0 - h(x_0))^2] = \mathbb{E}[(h(x_0) - \bar{h}(x_0))^2]&\mathbf{(Variance)} \\
+ (\bar{h}(x_0) - f(x_0))^2&\mathbf{(bias)} \\
+ \mathbb{E}[(y_0 - f(x_0))^2]&\mathbf{(Intrinsic)}
\end{aligned}
$$

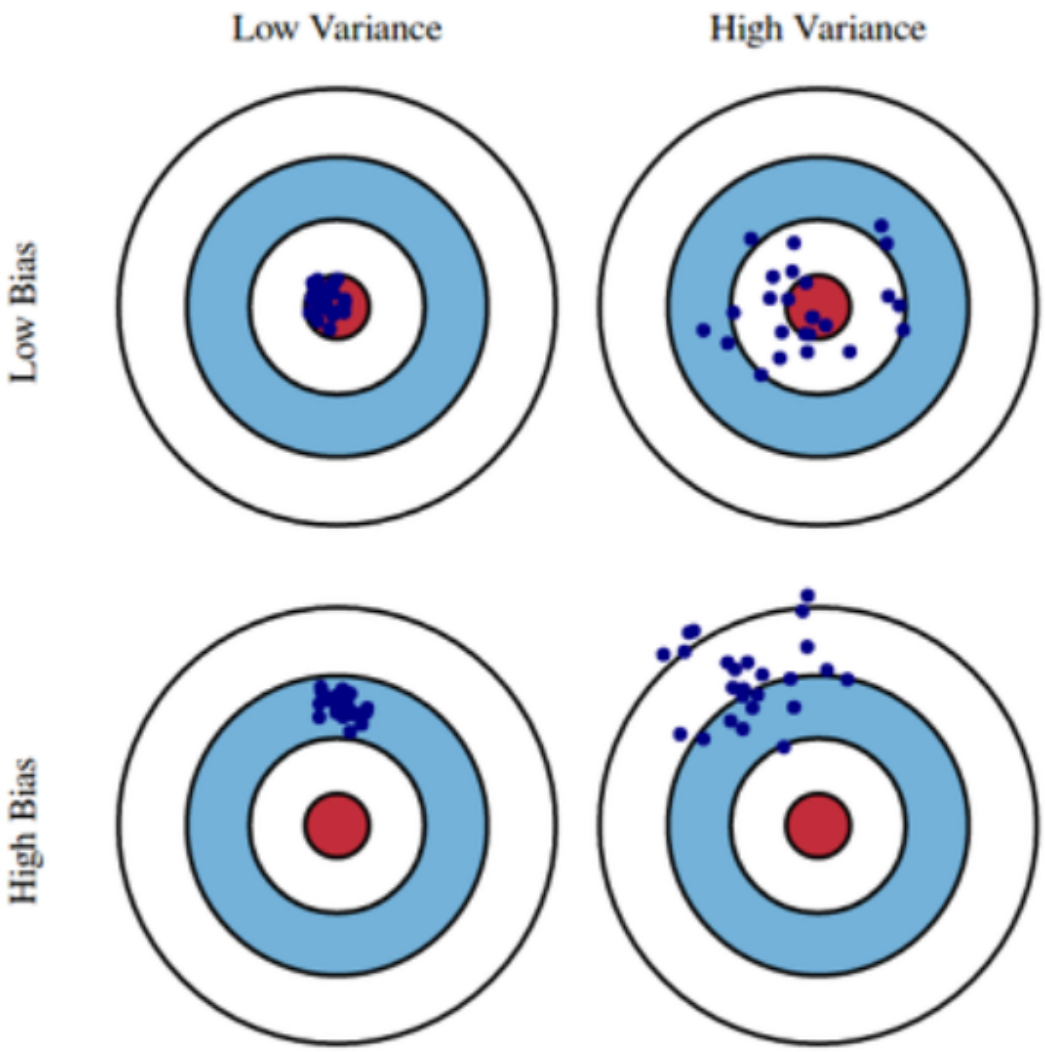
where $\bar{Z} = \mathbb{E}[Z]$

이것을 편향 - 분산(바이어스-배리언스)분해라고 부른다.

1항은 실제함수와 추정에 사용하는 함수의 차이를 제곱 오차로 수치화 한 평향
유연성이 결여된 함수라면 이 부분은 양수가 된다.
반대로 비모수적이고 유연한 방법인 경우 편향이 0 에 가까워진다.

2항은 배리언스다.
배리언스는 유연한 함수가 될수록 높아진다.

3항은 랜덤 노이즈의 분산
이부분은 함수와 관계없이 존재하며 줄일 수 없는 존재이다.



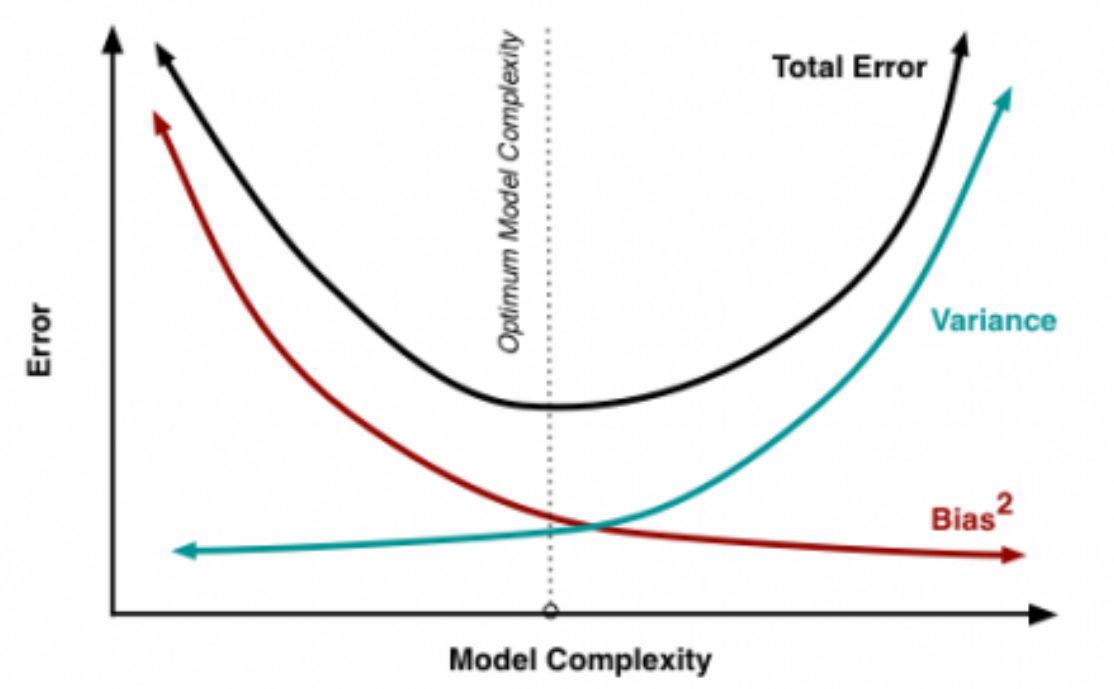
1번 과녁은 편향(정답에 가까움)은 낮고 분산(모여있다)도 낮다.

2번 과녁은 편향은 낮고 분산은 높다.
3번 과녁은 편향은 높고 분산은 낮다.
4번 과녁은 편향도 높고 분산도 높다.

참조: https://opentutorials.org/module/3653/22071

4.편향-분산 트레이드오프

편향-분산 트레이드오프란 유연성을 높여 근사오차를 낮추려고 할수록 배리언스가 상승하는 상관관계를 가진다.
그래서 우리는 편향과 분산을 봤을때 어느지점에서 오류가 최소화 되는 지점인 지를 찾아야한다.
그 지점이 바로 편향-분산 트레이드오프다.

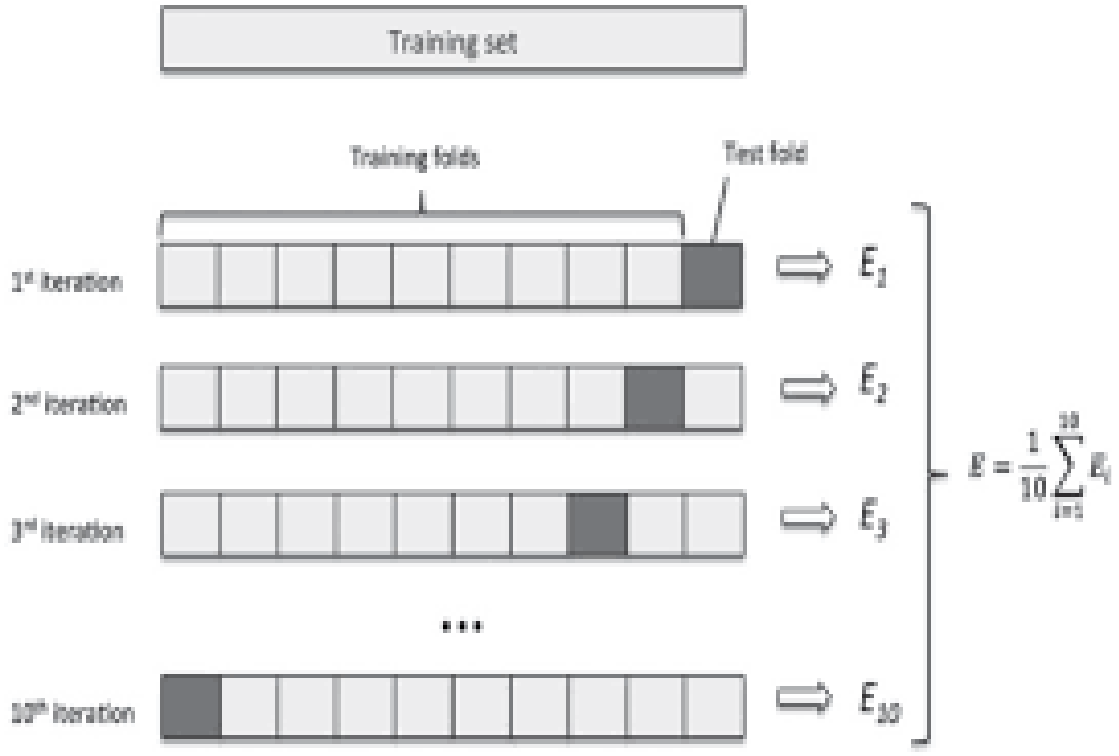


참조:https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-12-%ED%8E%B8%ED%96%A5Bias%EC%99%80-%

EB%B6%84%EC%82%B0Variance-Trade-off

5.교차검증법(Cross Validation)

N개의 데이터를 K분활하고 (k-1)개의 데이터를 학습 데이터로 취급해 모델을 추정한다.
이 과정을 k번 반복해 평균오차의 평균값을 취하면 시험 오류의 근삿값을 계산할 수 있다.
이 방법을 교차검증법(Cross Validation)이라고 한다.
학습할 때와 시험할 때 데이터를 생성하는 배후 모델이 같다고 가정할 수 있으면 어떤 상황에서도 사용할 수 있다.



'**Data Science**' 카테고리의 다른 글

과적합 교차 검증법 모델 선택 편향 분산 분해 편향 분산 트레이드오프

나아무늘보

혼자 끄적끄적하는 블로그 입니다.