

# Introduction to AIC

## Reference

- <https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced>

## 1. What is AIC?

*The Akaike information criterion (AIC) is an **estimator** of **out-of-sample** prediction error and thereby **relative** quality of statistical models for a **given set of data***

위키피디아는 AIC를 아래와 같이 정의한다. 중요하다고 생각 되는 부분을 굵게 표시 했는데, 첫 째로 AIC는 estimator라는 점이다. 그리고 AIC를 절대 지표로 사용할 수 없다. 즉, AIC는 항상 다른 AIC 값들과 비교하여 상대적(relative)으로 사용된다. 마지막으로 동일한 데이터에 대해서 계산된 AIC를 비교할 수 있다.

AIC는 ML에서 테스트할 데이터가 없을 때, 가장 많이 사용된다고 한다. 예를 들어, 데이터 세트가 매우 적은 경우 이 데이터를 훈련, 테스트 데이터로 또 나눈다면 데이터 갯수가 적어 과적합의 우려가 있다. 이때 주어진 데이터를 모두 사용하여 여러 모델을 훈련시키고 AIC로 성능을 비교하는 것이다. 또는 시계열 데이터에서, 가장 최근의 데이터를 사용하여 모델을 훈련하였다면 더 최근의 데이터는 존재하지 않으므로 이때 AIC를 사용할 수 있다.

AIC는 아래와 같이 정의된다.

$$AIC = -2\log(L) + 2k$$

AIC의 의미를 살펴보면 다음과 같다. AIC는 log likelihood에 기반하여 정의되는데, likelihood가 클수록 주어진 데이터에 모델이 잘 적합한다는 뜻이므로 AIC는 작을 수록 좋을 것이다. 그런데 사실 likelihood는 적합되는 모수가 증가할수록 커질 수밖에 없다. 따라서 likelihood만으로 모델을 비교하면 모든 모수를 포함하는 full model이 가장 좋은 결론을 얻는다. 이러한 이유로 penalty의 역할을 하는  $2k$ 를 추가한 것이다. 모수가 커질 수록  $-2L$ 은 작아지지만  $2k$ 는 커지므로 서로 상쇄하는 효과를 기대할 수 있다.

## 2. When should use AIC?

1. 동일한 데이터에 대해 적합한 모델들을 비교할 때
2. 모델들의 동일한 결과를 비교할 때
3. infinite sample size일 때

infinite sample size 조건이 충족될 때, AIC는 estimate가 아닌 실제 correct answer로 수렴한다고 한다. infinite까지는 아니어도 충분한 sample이 있으면 관계 없다고 한다. 하지만 사실 sample의 수가 적을 때 AIC를 사용한다. 이러한 이유로 AIC를 보정한 AICc가 나왔다. AICc는 적은 수의 샘플을 가질 때 정확한 결과를 보여준다.

경험적인 결과로 데이터 수를  $n$ , 모수의 수를  $k$ 라고 할 때 이들의 비가 40보다 작으면 AICc를 쓰는 것이 좋다고 한다. 이에 대한 수학적 설명은 여기를 참고하자.

AIC의 또 다른 장점 중 하나는 LRT처럼 굳이 nested model일 필요가 없다는 뜻이다. 따라서 LRT를 통해서 비교할 수 없는 모델을 AIC를 통해서 비교할 수 있다.

### 3. How to interpret the results?

앞서 AIC는 likelihood를 이용하여 정의되었기 때문에 작을 수록 좋다고 언급했었다. 그런데 가장 좋은 모델을 선택할 때, 이런 기준을 사용할 수도 있지만 AIC는 *probabilistic ranking of the models that are likely to minimize the information loss* 이기 때문에 아래와 같은 방법을 사용할 수도 있다. 위키피디아의 예시를 살펴보자.

세 모델에 대해서 얻은 100, 102, 110의 AIC 값으로 아래를 계산할 수 있다.

$$P = \exp((AIC_{min} - AIC_i)/2)$$

$AIC_{min} = 100$  이므로 102, 100 AIC 값에 대해서  $P$ 를 계산하면 0.368, 0.007이다. 이는 AIC가 100인 모델보다 각각 0.368, 0.007배 더 좋은 모델일 수 있다고 해석한다. 0.007이 가장 작으므로 AIC가 110인 모델은 제외하고 나머지 모델로 1:0.368의 비율로 weighted sum을 하거나 또는 정확도가 가장 중요한 목적이 아니라면 그냥 가장 낮은 AIC를 가지는 모델을 택할 수도 있다.

### 4. Pitfalls of AIC

다시 한번 상기하자면, AIC는 모델에 대한 절대적인 지표가 아니라 상대적인 지표이다. 이는 곧, 모든 모델이 poorly 적합 되었을 수도 있으므로 다른 절대적인 지표(MAPE)와 함께 계산해야 한다. 또한 AIC는 계산하기 상대적으로 간단한데, 계산이 복잡한 대신 더 정확한 결과를 도출하는 DIC, WAIC, LOO-CV 등도 있으니 상황에 따라서 적합한 metric을 선택해야 한다. 이에 대해서는 여기를 참조하자.