# 18장 평활법

## 1. 서론

평활법 – 시계열 예측방법 중 하나. 시계열 구성 요소들의 통계적 가정을 근거로 구축되는 회귀분석 방법과는 달리, 특별한 모델 형태 없이 데이터로 부터 직접 예측값을 추정하는 방법.

데이터 기반 방법의 장점 – 시계열 패턴이 시시각각 바뀌는 상황에 유용하게 적용할 수 있다는 것.

평활법의 특징 – 시계열 내 잡음을 제거함으로써 숨겨진 패턴을 찾을 수 있음.

기본개념 – 관측치의 평균을 구하여 미래의 값을 예측

↳ 평활상수나 평활 방법에 따라 몇 개의 관측치 평균을 구할 것인지, 평균은 어떻게 구할 것인지 혹은 얼마나 자주 평균을 취할 것인지가 결정됨.

## 2. 이동평균법

용어자체에서 가늠할 수 있듯이 일정구간 내의 관측치들의 평균을 이용하여 예측하는 방법.

이동평균법을 수행하기 위해서는 윈도우의 폭 $w$를 정하고, 이에 따라 각 윈도우 내 $w$개의 연속된 값들의 평균을 구하여 미래값 예측.

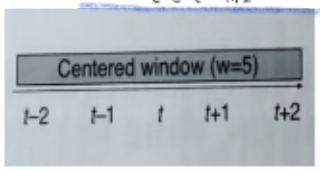
일반적으로 두 가지 종류 ┌ 전후이동평균 (Centered moving average) – 평균을 취하는 과정에서 잡음 제거 ⇒ 전체적인 추세를 살펴보는데 유용
└ 이전이동평균 (Trailing moving average) – 예측에 주로 사용

↳ 이 두 방법은 윈도우가 시계열의 어떤 부분에 위치하는지에 따라 다르다.

- 시각화를 위한 전후이동평균법

$t$ 시점에서의 이동평균값 ($MA_t$)은 시점 $t$를 중심으로 $w$개 값의 평균을 취함으로써 구할 수 있다.

$$MA_t = (Y_{t-(w-1)/2} + \cdots + Y_{t-1} + Y_t + Y_{t+1} + \cdots + Y_{t+(w-1)/2})/w$$




윈도우의 폭이 5, $t=3$에서의 이동평균은 $t=1,2,3,4,5$ 시점의 평균값

이동평균법을 적용하기 전 우선 윈도우의 폭 $w$를 결정해야함.

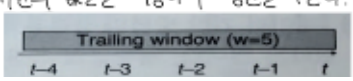
↳ 계절변동을 줄이고 추세가 보다 잘 드러나도록 시각화하는 것이 주목적이기 때문에, 기본적으로 $w$는 계절의 주기로 결정

- 예측을 위한 이전이동평균법

전후이동평균 – 구하고자 하는 시점을 기준으로 과거와 미래 데이터의 평균으로 결정되기 때문에 예측문제에는 적용하기 어렵다.

⇒ 예측 시에는 이전이동평균법 사용

이전이동평균법 – 윈도우의 폭 ($w$)을 정하고 이에 맞게 최근 $w$기간의 값들을 이용하여 평균을 구한다.




폭 $k$기간 후의 예측값 $F_{t+k}$ ($k=1,2,3,\cdots$)은 최근 $w$개의 평균값

$$F_{t+1} = (Y_t + Y_{t-1} + \cdots + Y_{t-w+1}) / w$$

이동평균법은 계절변동을 반영하지 못하기 때문에 월별 데이터를 예측하기에는 불충분. ⇒ 결론적으로 이동평균은 단순평균값을 이용하기 때문에 증가 및 감소추세를 반영하지 못하고 과대 혹은 과소 예측하게 되는 한계점을 가지고 있다.

↳ 추세가 존재하는 시계열 데이터를 예측할 때에도 위와 같은 오류발생.

일반적으로 이동평균법을 통한 예측은 계절변동이나 추세가 없는 경우 적용

↳ 이러한 제약때문에 실용적이지 않아 보이지만, 회귀모델의 잔차를 이용하는 방식과 같이 추세나 계절변동이 제거된 데이터에는 적용가능

⇒ 예측 시 추세나 계절변동이 제거된 데이터에 이동평균법을 적용하고 후에 추세와 계절변동을 다시 반영할 수 있다.

- 윈도우의 폭($w$) 결정

이동평균 시 사용자가 미리 지정해야하는 유일한 하이퍼모수.

이동평균법의 모수인 $w$를 결정하는 것은 과소평활과 과대평활 사이의 균형을 맞추는 작업.

↳ k-최근접 이웃 알고리즘에서 k를 결정하는 것과 마찬가지

시각화(전후이동평균)의 경우 ┌ $w$값을 크게 할수록 전역적인 추세가 반영
                              └ 작게 한다면 국소적인 추세가 드러나게 된다.

⇒ 여러 값의 $w$를 시도해본 것이 전역적/국소적인 추세를 확인할 수 있는 좋은 방법

예측(이전이동평균)의 경우 - $w$의 결정은 시계열 데이터의 전반적인 변화가 언제 있었는지가 중요하기 때문에 해당 데이터에 관한 배경지식이 요구됨.

↳ 예측성능을 검증하기 위한 또 다른 방법은 $w$값을 변화시켜보면서 그때마다 결과를 비교.

↳ 그러나 이를 통해 야기될 수 있는 모델의 과적합은 주의

## 3. 단순지수평활법

지수평활법 - 실제 업무에서 가장 많이 활용되는 예측방법 중 하나. → 이 기법의 유연성, 자동화, 빠른 연산, 정확한 예측력 때문에

단순지수평활법 - 이동평균법과 비슷하지만 단순히 $w$개의 평균을 구하는 것이 아니라, 과거 데이터들의 가중평균을 이용한다는 차이

↳ 가중평균을 계산하는데 사용되는 가중치들은 지수분포 형태로 부여 ⇒ 즉, 최근 데이터에 가장 큰 가중치부여, 과거로 갈수록 지수분포형태로 감소하는 가중치

↳ 주목할 점은 아무리 오래된 관측치라도 작게나마 가중치가 주어진다는 것.

이동평균법과 마찬가지로 단순지수평활법도 추세나 계절변동이 없는 시계열의 예측에 적용 (한 시계열에서 추세나 계절변동이 제거된 시계열 데이터의 경우 적용 가능)

지수평활법을 통한 시점 $t+1$에서의 예측값($F_{t+1}$)

$$F_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots$$ → 과거의 모든 데이터의 가중평균 확인 가능하나 지수적으로 감소

↳ α는 평활상수로서 0과 1 사이의 값을 갖는다.

$F_{t+1} = F_t + \alpha E_t$ 와 같은 식으로 표현되기도 함.

⇒ 즉, t+1 시점의 예측값($F_{t+1}$)은 t 시점에서의 예측값 $F_t$와 예측오차 $E_t$의 가중합으로 구해짐

• 이 식의 계산은 전체 시계열을 모두 사용하지않고 이전 기간의 예측값과 예측오차만을 사용

⇒ 데이터 저장이나 연산시간에서도 이점을 가짐 → 실제 문제는 실시간으로 데이터를 예측하거나 여러 개의 시계열을 동시에 예측.

↳ 저장공간이나 연산시간의 절약은 매우 중요

지수평활법으로 계산되는 예측값들은 다음 시점의 예측값과 동일

↳ 시계열 데이터에 추세와 계절변동이 없다고 가정했기 때문에

$F_{t+k} = F_{t+1}$ → k기간 앞의 예측값은 바로 다음 기간의 예측값과 차이가 없다.

- 평활상수 α 선택

사용자가 설정해야할 평활상수 α - 예측 시 과거 데이터에 대한 비중을 얼마나 부여할지를 조절하는 모수

↳ 1에 가까운 값을 사용하면 최근 관측치에 더 큰 비중, 0에 가까운 값은 과거값에 더 큰 비중

⇒ α의 선택은 요구되는 평활의 정도에 따라, 예측하는데 얼마만큼의 과거 데이터가 관련되어있는지에 따라 달라진다.

최적의 평활상수를 결정하는 방법은 딱히 존재하지 않지만 보통 다양한 값들을 시도하여 그중 검증데이터의 예측오류를 최소화하는 값을 정하는 방법이 사용

여기서 주의해야할 점은 "최적의 α"를 선택하는 것은 모델의 과적합이나 미래 데이터의 예측에 있어 부정확한 결과를 도출할 수 있다는 점. 평활상수에 대한 지식이 없을 때는 통상 0.1-0.2 사이의 값을 사용

추세나 계절변동이 없는 데이터에 사용되는 단순지수평활법은 회귀식을 통해 얻은 잔차를 이용해서 예측에 활용 가능
↳ 보통 잔차는 추세나 계절변동이 없음

- 이동평균법과 지수평활법의 관계

이동평균법 - 윈도우의 폭(ω), 지수평활법 - 평활상수(α)를 설정해야함.

↳ 두 하이퍼 모수는 최근 정보가 과거 정보에 비해 얼마나 중요하게 고려되는지를 결정해주는 역할.

ω와 α 사이에는 관계가 있음. $\omega = 2/\alpha - 1$ 인 경우 이동평균법과 지수평활법은 유사한 결과

## 4. 고급지수평활법

이동평균, 단순지수평활 - 추세나 계절변동이 없는 경우에 사용

↳ 회귀모델 등을 사용하여 추세나 계절변동 제거 후에도 사용 가능

- 추세가 존재하는 시계열

추세를 가진 시계열의 경우 이중지수평활법(double exponential smoothing) 적용 가능 → 회귀모델과는 달리 시간에 따라 계속 변하는 추세에 적용 가능

주어진 데이터를 사용하여 국소추세를 추정하고 새로운 데이터가 주어짐으로써 추세를 갱신

이중지수평활법을 이용한 k기간 이후의 예측값은 시간 t에서의 수준($L_t$)과 추세($T_t$)의 합

$$F_{t+k} = L_t + kT_t$$

추세가 존재하기 때문에, 단순지수평활법과 달리 1, 2, 3, ··· 기간 이후 예측값이 모두 다르게 된다.

수준과 추세는 아래 식을 통해 반복적으로 갱신

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

시간 t에서 수준 $L_t$는 시점 t에서 실제 값과 추세정보를 통해 보정된 이전 수준 값($L_{t-1}$)의 가중평균을 통해 결정

시간 t에서 추세 $T_t$는 t-1 시점에서의 추세와 수준의 차이값의 가중평균을 통해 구할 수 있다.

※ 두 개의 모수 $\alpha, \beta$를 사용자가 결정. 단순지수평활법과 마찬가지로 (0,1) 사이의 값을 가지며 클수록 최근 정보에 많은 가중치

- 추세와 계절변동이 모두 존재하는 시계열

추세와 계절변동이 모두 있는 시계열의 경우 홀트-윈터 지수평활법 이용

↳ 이중지수평활법을 좀 더 확장한 개념. k 기간 이후를 예측하면서 특정 기간 내의 계절변동도 함께 고려

M기간 마다 계절변동이 존재한다고 가정하면 예측값은

$$F_{t+k} = (L_t + kT_t) S_{t+k-m}$$

주의할 점은 시간 t에서의 예측을 가능하게 하기위해 시계열이 적어도 한번의 완전한 계절 순환주기를 포함해야함. 즉 $t > M$

홀트-윈터 지수평활법은 수준, 추세, 계절요인을 반영할 수 있는 예측방법. 이 세 요소는 정보가 추가됨에 따라 계속 추정, 갱신됨

$$L_t = \frac{\alpha Y_t}{S_{t-M}} + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \frac{\gamma Y_t}{L_t} + (1-\gamma)S_{t-M}$$

첫 번째 식은 이중지수평활법과 유사. t 시점 값을 그대로 사용하는 것이 아니라 계절적요소를 포함하여 조정된 값을 사용한다는 차이

조정된 값은 $Y_t$를 계절지수($S_{t-M}$)로 나누어서 구할 수 있다.

두 번째 식은 이중지수평활법과 동일.

마지막 식은 계절지수. 이전 주기의 계절지수와 현재의 추세를 고려하여 조정된 값의 가중평균을 통해 계산 ↗ 계절에 따른 값의 차이를 비율로서 표현

- 계절변동만 포함된 시계열

추세는 없이 계절변동만 포함된 시계열 → "홀트-윈터 지수평활법"에서 추세 부분 식을 제외한 채로 적용.