16장 시계열 데이터 분석

1. 서론

시계열 예측 - 숫자를 다루는 대부분의 조직에서 행해짐.

이런 장들에서는 주로 시간에 따른 순서가 중요치 않은 횡단면 데이터들에 대해 예측·분류

→ 이번 장에서는 시간에 따라 변화하는 시계열 데이터 분석에 중점

최근 기술의 발달로 시계열 데이터의 수집주기가 매우 짧아짐. → 데이터가 매우 짧은 주기로 생성되더라도 예측에는 항상 유용하지 않을 수 있다.

수집하고자 하는 데이터의 시간주기를 결정하기 위해서는 반드시 예측하고자 하는 시간의 주기와 데이터에 내재되어있는 잡음을 고려해야한다. ex) 식료품점의 다음날 판매량 예측 - 분 단위의 데이터는 가장 불규칙한 시간과 그렇지 않은 시간의 분산을 포함하고 있음 (잡음)

→ 시간을 좀 더 긴 주기로 변환해서 분석

2. 탐색 vs 예측 모델

횡단면 데이터에서와 마찬가지로 시계열 데이터모델링도 { 탐색 / 예측 } 로 구분 가능

탐색적인 측면에서 시계열 분석의 두 목적 - 외부인자와 관련하여 계절적인 패턴, 추세와 같은 요소를 설명할 수 있는 모델을 결정 → 의사결정이나 정책수립 등에 활용 가능

예측적인 측면에서 시계열 분석의 두 목적 - 데이터의 미래 값을 예측

↳ 탐색적인 면과 예측적인 면의 다른 점은 각각이 다른 종류의 방법으로 구현, 모델링 방식의 차이

3. 비즈니스에서 주로 사용되는 예측기법

비즈니스 분야에서 많이 사용되는 두 가지 예측기법 { 회귀분석 기법 - 시계열 데이터로부터 모델 파라미터 추정 / 평활법 - 데이터로 부터 직접 패턴을 추출 }

- 기법들의 결합

예측력을 높이기 위해 여러 기법들을 결합 ← 13장의 '앙상블'과 비슷한 맥락
보통 2단계를 거침

① 실제 데이터를 첫 번째 모델에 적용하여 미래값에 대한 예측치를 생성.

② 첫 번째 모델로부터 얻어진 잔차를 두 번째 모델에 적용하여 미래의 오차값을 예측 ← 예측된 오차정보는 첫 번째 모델의 예측오차를 보정하는 효과

⇒ 기법들을 결합함으로써 시계열 데이터에 내재되어있는 다양한 특성 최대한 반영 가능

여러 방법들로 얻은 값들의 평균을 이용하면 보다 정확, 분산이 작은 결과

4. 시계열 요소

보통 4가지 요소로 설명 : 수준(level), 추세(Trend), 계절변동(Seasonality), 잡음(noise)

수준 - 시계열의 평균값 추세 - 한 시점에서 다음 시점으로의 평균적인 변화

계절변동 - 짧은 기간 동안의 주기적인 패턴 잡음 - 무작위적인 변동

↳ 시간 그래프를 그리면 이해하기 쉬움.

시계열 데이터를 좀 더 자세히 볼 수 있는 방법

"확대 (Zoom-in)" - 장시간에 걸쳐 연버된 시계열 데이터를 작은 구간으로 나누어 확대 ⇒ 내재되어 있는 패턴 확인 가능

"스케일 변화 (change scale of series)" - 데이터의 스케일을 바꿈으로서 전반적인 경향을 좀 더 정확히 볼 수 있음. (선형 → 로그 등)

"추세선 추가 (Add Trend line)" - 추세 확인 가능. 여러 종류의 추세선 (선형, 지수곡선, 실거식 곡선)들을 시도하고 가장 적합한 것을 찾는다.

"계절변동 제거 (Supress Seasonality)" - 계절변동을 제거함으로서 데이터의 전반적인 추세를 좀 더 뚜렷하게 볼 수 있음.

↳ 데이터의 시간 스케일을 늘린다. (월분 → 연간 등)

몇몇 예측기법들은 시계열 형태에 관한 가정을 기반으로 모델링.

↳ '추세'에 대한 가정 - 시계열 데이터의 전체 혹은 일부가 선형 또는 지수곡선을 따른다.

↳ 많은 통계모델들이 노이즈가 정규분포를 따른다는 가정 하에 수립

⇒ 데이터가 가정에 잘 맞을 경우 신뢰성있는 예측값을 얻을 수 있음.

'데이터 기반 예측기법' - 통계적 가정이 적용되지 않고 데이터의 패턴이 시시각각 바뀌는 점우 유용하게 사용될 수 있음.

↳ 간단하고 구축 시 계산량이 적다는 장점도 있음.

모델 기반 / 데이터 기반 방법의 선택기준 - '전역패턴'을 찾느냐 '국소패턴'을 찾느냐에 따라 달라질 수 있음.

전역패턴 - 전체 시계열 내에서 크게 변동이 없는 패턴

국소패턴 - 짧은 시간의 변동을 담고 있는 패턴

모델 기반 - 일반적으로 전역 패턴이 존재하는 시계열 데이터 예측에 사용 → 패턴 추정 시 전체 구간의 데이터를 이용하기 때문에

↳ 국소패턴의 점우 패턴이 바뀌는 시점을 그때그때 지정해줘야함. ⇒ 데이터 기반 기법이 유용

⇒ 시계열 그래프는 시계열 요소를 발견하는데 사용될 수 있음과 동시에 추세와 계절변동에 내재되어있는 전역 혹은 국소패턴을 찾는데도 사용될 수 있다.

5. 데이터 분할 및 성능평가

시계열 데이터의 점우 원칙적로 분할 시 연속되는 시간이 단절되는 점우가 생길 수 있음.

↳ 시간순서를 고려하여 분할 필요 (앞 시점의 데이터로 모델 구축, 뒤 시점의 데이터로 평가)

모델 평가는 RMSE, MAPE를 주로 사용.

- 경쟁 모델의 성능 : 단순예측

성능 평가 시 '단순예측 (naive forecast)'와 먼저 비교실시.

↳ 가장 최근 값으로 미래를 예측. 매우 단순하지만 의외로 좋은 성능을 보일 수도 있다.

- 미래 예측값 생성

시계열 데이터의 미래값 예측 - 학습과 검증을 합쳐 하나의 통합된 데이터를 만들고 구축된 모델을 통합 데이터에 다시 적용함으로써 예측값을 얻는다.

통합데이터 사용의 장점

① 최근 시점의 검증 데이터는 유용한 예측정보를 담고있다.

② 많은 데이터를 이용함으로써 보다 정확한 모델 구축 가능

③ 학습데이터 만으로 예측을 한다면 항상 검증 데이터의 가장 먼 미래의 값 밖에 예측할 수 밖에 없다.

↳ 정확도가 떨어짐.