Hafta 9 (Unsupervised Learning - Principal Component Analysis - Time Series)

@mebaysan

30/10/2021

İlgili Okuma Listesi:

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/99a52248-29b3-4fc6-903a-dbf2120f29b8/denetimsiz_ogrenme.9.hafta.cars.pdf

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/3b11bc26-5360-4c61-abdd-f4bba8cd1f2a/times_series.9.hafta.pdf

Benim Yazdığım Yazılar:

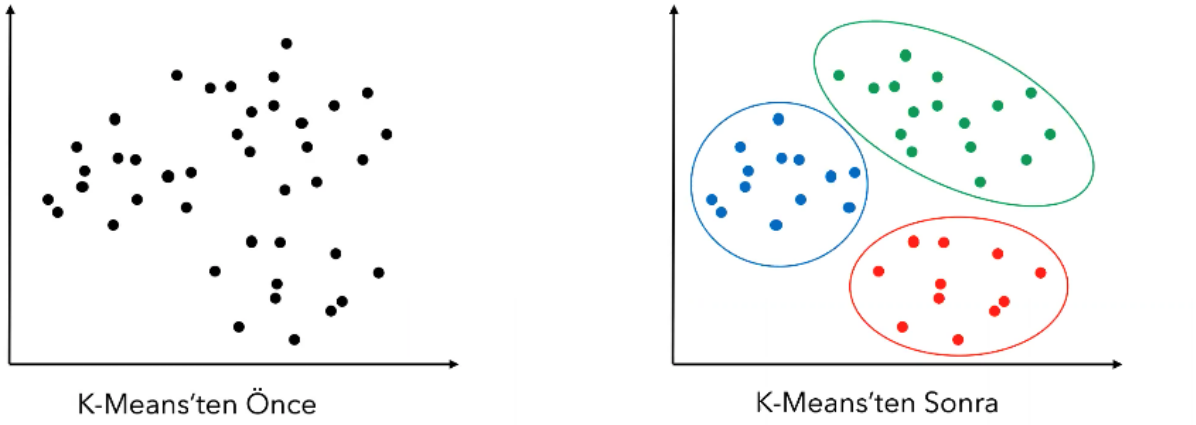
-

Unsupervised Learning (Denetimsiz Öğrenme)

Bağımlı değişkenimiz olmadığı senaryolarda unsupervised learning ile uğraşacağız.

K-Means (K-Ortalamalar)

Amaç: Gözlemleri birbirlerine olan benzerliklerine göre kümelere ayırmaktır.



K-Means Uygulama:

- Küme sayısı belirlenir
- Rastgele k merkez seçilir
- Her gözlem için k merkezlere uzaklıklar hesaplanır
- Her gözlem en yakın olduğu merkeze (kümeye) atanır
- Atama işlemlerinden sonra oluşan kümeler için tekrar merkez hesaplamaları yapılır (iteratif)
- Bu işlem belirlenen bir iterasyon adedince tekrar edilir ve küme içi hata kareler toplamlarının toplamının (total within - cluster variaton) minimum olduğu durumdaki gözlemerin kümelenme yapısı nihai kümelenme olarak seçilir

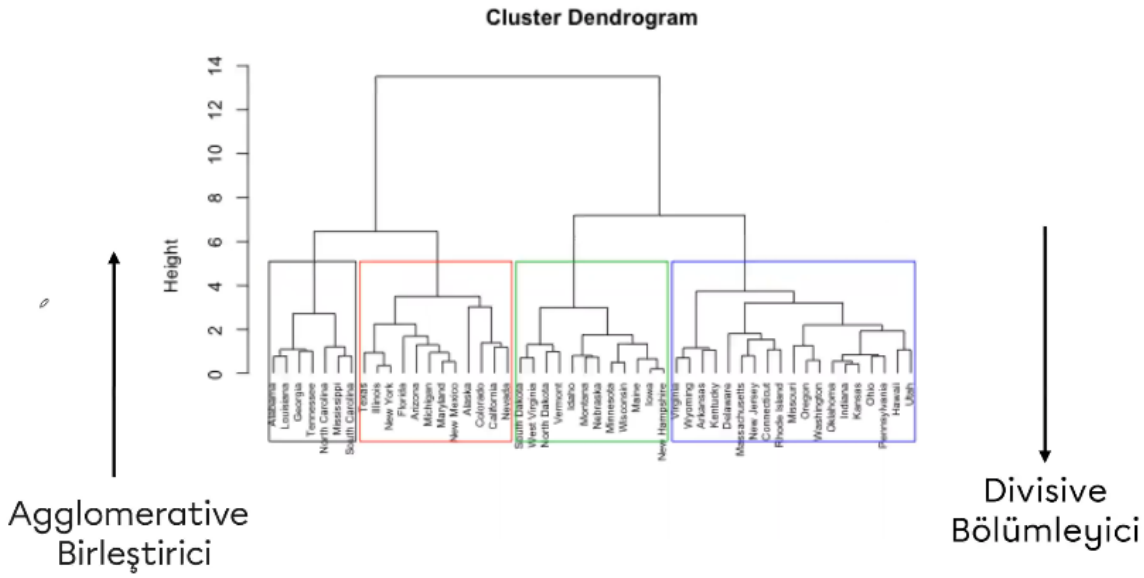
Uzaklık temelli bir yöntem olduğu için aykırı değerlerden etkilenebilmektedir

Hierarchical Cluster Analysis (Hiyerarşik Kümeleme Analizi)

Amaç: K-Means'teki gibi, birbirine benzer olan gözlemleri kümelere ayırmaktır.

Yaygınca kullanılan 2 yöntem:

- Agglomerative (Birleştirici)
  - Bütün gözlem birimleri kendi başına kümeymiş gibi kabul edilip birbirlerine olan uzaklıklarına göre birleştiriliyor
- Divise (Bölümleyici)
  - Bütün gözlem birimleri bir küme olarak kabul edilip daha sonra kümelere ayrılıyor



Agglomerative (Birleştirici) Clustering

Başlangıçta gözlem sayısı kadar küme vardır.

- Veri setinde birbirine en yakın olan iki gözlem bulunur

- Bu iki nokta bir araya getirilerek yeni bir gözlem oluşturulur. Yani artık veri seti ilk birleşimdeki gözlemlerden oluşmaktadır
- Aynı işlem tekrarlanarak yukarı doğru çıkılır. Yani iki kümenin birleşiminden oluşan bu yeni kümeler aynı şekilde birbirlerine benzerliklerine göre tekrar birleştirilir. Bu işlem tüm gözlemler tek bir kümede toplanana kadar bu işlemler tekrar edilir

Birbirine yakın noktalar uzaklık ölçüleri kullanarak belirlenir. Öklit uzaklığı, manhattan uzaklığı, korelasyon vs kullanılır.

Divise (Bölümleyici) Clustering

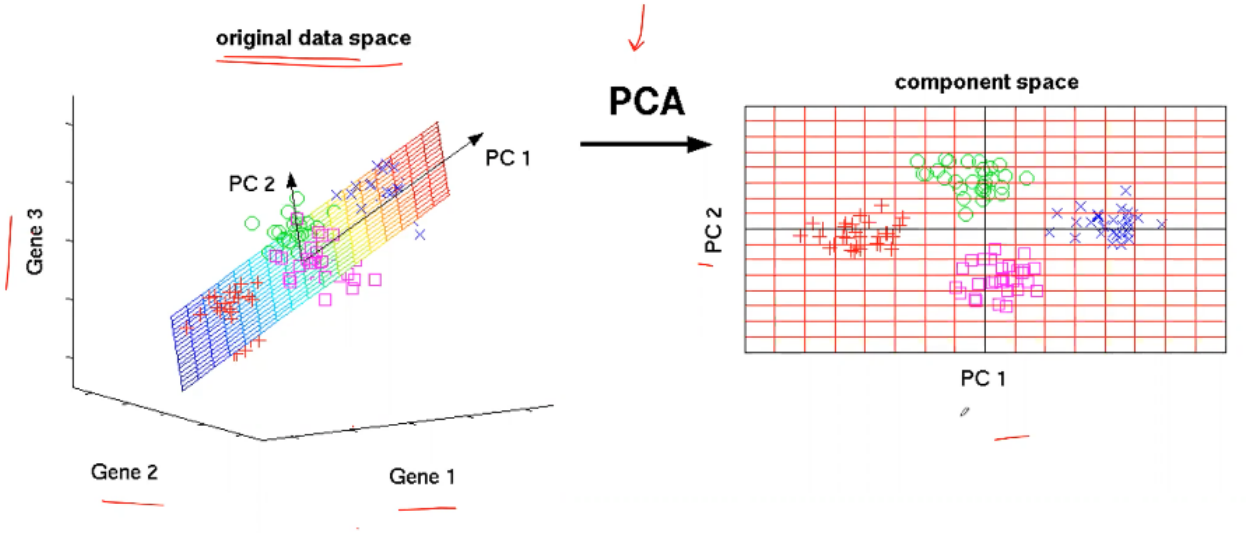
Başlangıçta 1 tane küme vardır. O da tüm veri setidir.

- Tüm gözlemlerin bir arada olduğu küme iki alt kümeye ayrılır
- Oluşan yeni kümeler birbirlerine benzemeyen alt kümelere bölünür
- Aynı işlem gözlem sayısı kadar küme elde edilinceye kadar tekrar edilir

Principal Component Analysis (Temel Bileşen Analizi)

Temel fikir, çok değişkenli verinin ana özelliklerini daha az sayıda değişken/bileşen ile temsil etmektir.

Diğer bir ifade ile; küçük miktarda bilgi kaybını göze alıp değişken boyutunu azaltmaktır.



Çok Boyutluluk Laneti: Değişken sayısının gözlem sayısından fazla olmasıdır.

PCA Kullanım Alanları:

- Çok boyutluluk lanetinden kurtulmak için PCA kullanılır.
- Ekonometrik, iktisadi ve istatistik alanlarında yapılan modellemeler, regresyon problemlerinde çoklu doğrusal bağlantı probleminden kurtulmak için kullanılır.
- Görüntü işleme çalışmalarında gürültüden kurtulmak için kullanılır.
- Makine öğrenmesi yöntemi olarak kullanılabilir. Önce boyut indirgenir üzerine model fit edilir.
- Çok boyutlu veriyi görselleştirmek için kullanılabilir.

Time Series Forecasting (Zaman Serileri)

Sektörde kendine en fazla yer bulan uygulamalardandır.

- Zaman Serisine Giriş ve Temel Kavramlar
- Smoothing Methods

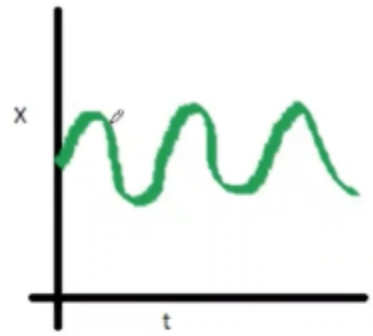
    - Single Exponential Smoothing (a)
    - Double Exponential Smoothing (a, b)
    - Triple Exponential Smoothing a.k.a. Holt-Winters (a, b, g)
- Statistical Methods
    - Ar (p)
    - Ma (q)
    - ARMA (p, q)
    - ARIMA (p, d, q)
    - SARIMA (p, d, q) (P, D, Q) m
- Machine Learning for Time Series Forecasting

Zaman Serisine Giriş ve Temel Kavramlar

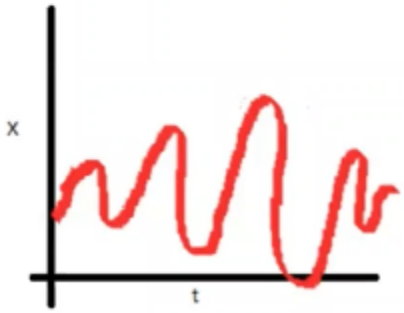
Zamana göre sıralanmış gözlem değerlerinden oluşan verilere "zaman serisi" denir.

Stationary (Durağanlık)

- Bir zaman serisinin istatistiki özelliklerinin zamana göre değişmemesini ifade eder.
- Bir zaman serisinin ortalaması, varyansı ve kovaryansı zaman boyunca sabit kalıyorsa, serinin durağan olduğu ifade edilir.



Stationary series

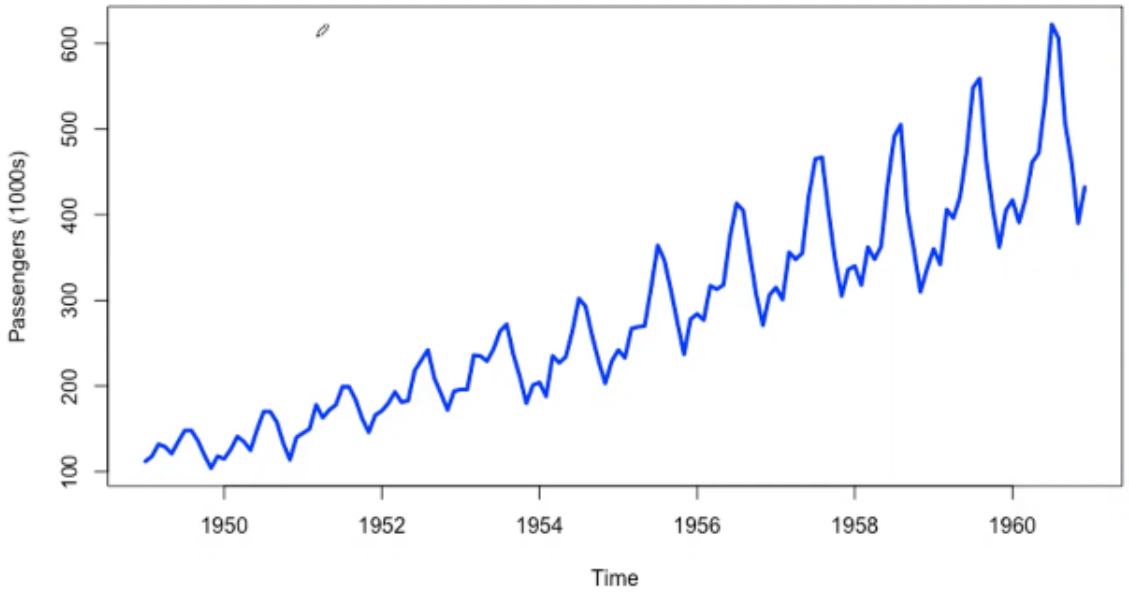


Non-Stationary series

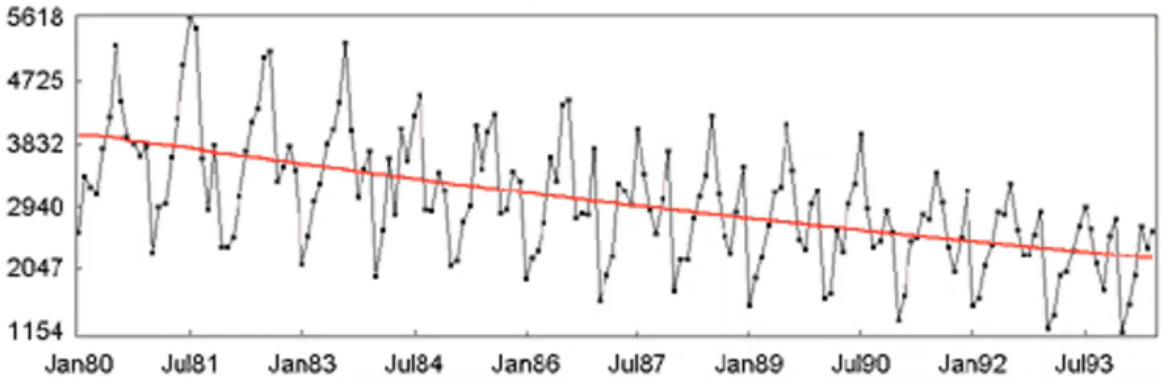
Trend

- Serinin artış ya da azalış durumunu ifade eder.
- Bir zaman serisinin uzun vadedeki artış ya da azalışının gösterdiği yapıya trend denir.

Artan trend'e sahip bir seri:



Azalan trend'e sahip bir seri:

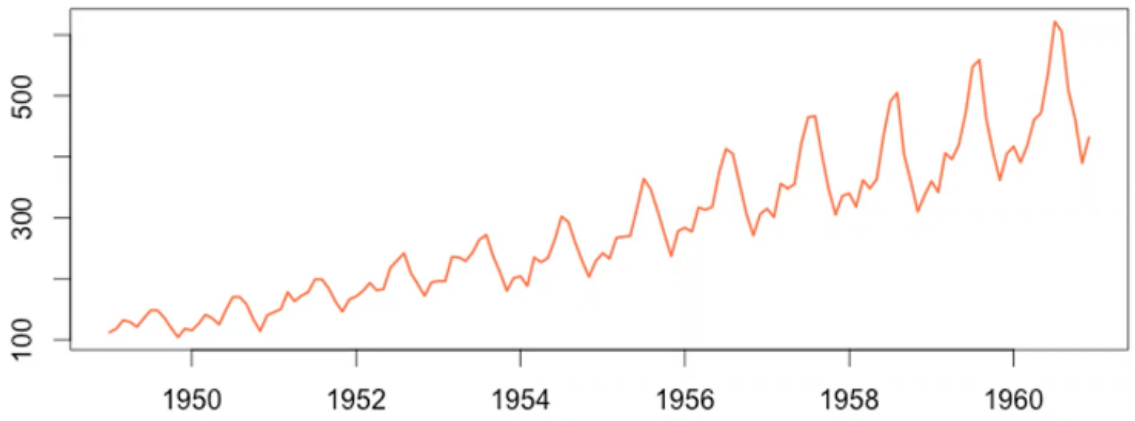


Seasonality (Mevsimsellik)

- Serinin izlediği belirli bir örüntü var mı yok mu?

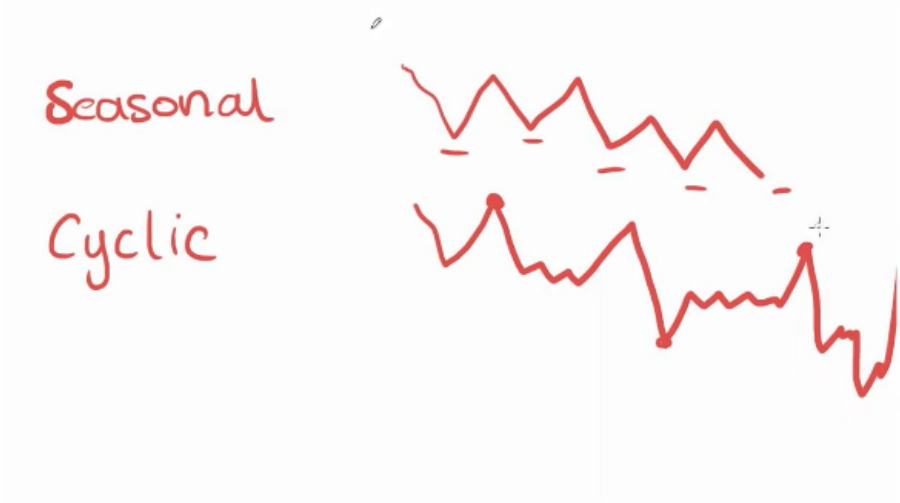
- Zaman serisinin belirli bir davranışı belirli periyorlarla tekrar etmesi durumuna mevsimsellik denir.

Mevsimselliği olan örnek bir seri:



Cycle (Döngüsellik)

- Mevsimselliğe benzer fakat belirli örüntüleri sergileme yönü zayıf olan bir kavram.
- Tekrar eden örüntüler (düzenler) barındırır.
- Mevsimsellikte örüntüyü yakalamak kısa vadelidir: Aylık, haftalık, günlük, yıllık vs. Daha açıktır
- Döngüsellikte örüntüler daha uzun vadelidir. Haftalık vb kısa vadeli yapılarda gözlemlenemez.



Zaman Serisi Modellerinin Doğasını Anlamak

Genel kural: Bir seri kendisinden önceki değerden daha fazla etkilenir.

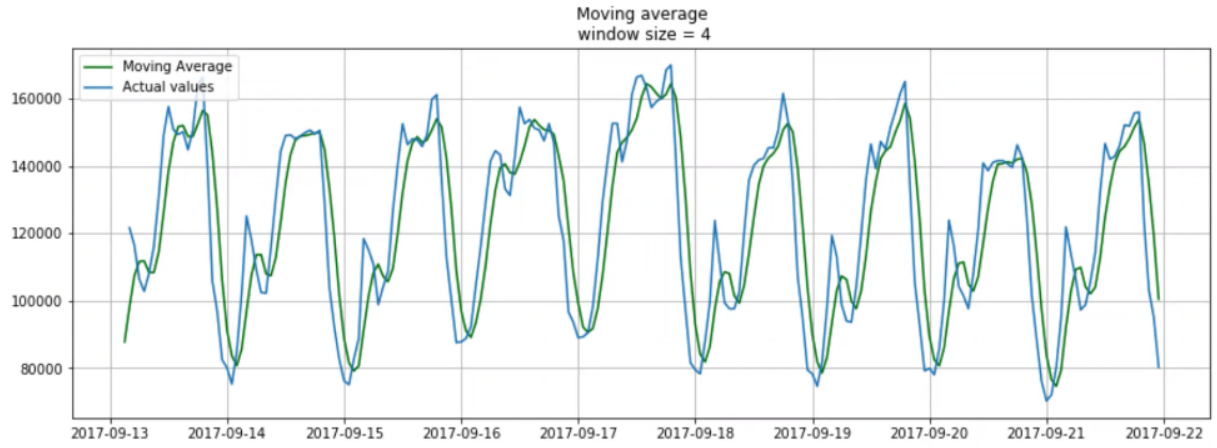
Örnek: 3 Kasım satışları için 2 Kasım tarihinin satışları 1 Kasım'dan daha etkilidir.

Moving Average

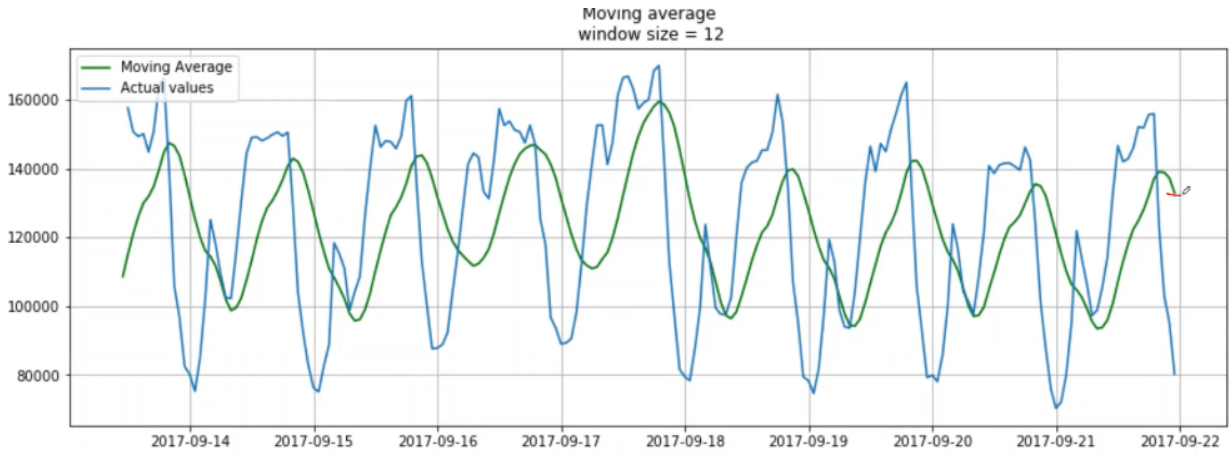
Bir zaman serisinin gelecek değeri kendisinin k adet önceki değerinin ortalamasıdır.

$$\hat{y}_t = \frac{1}{k} \sum_{n=1}^{k} y_{t-n}$$

4 günlük hareketli ortalama:



12 günlük hareketli ortalama:



Weighted Average

Ağırlıklı ortalama. Hareketli ortalamaya benzer. Daha sonralarda olan gözlemlere daha fazla ağırlık vermek fikrini taşır.

$$\hat{y}_t = \sum_{n=1}^{k} \omega_n y_{t+1-n}$$

Time	Ads
2017-09-21 19:00:00	155890
2017-09-21 20:00:00	123395
2017-09-21 21:00:00	103080
2017-09-21 22:00:00	95155
2017-09-21 23:00:00	80285

Time	Ads
2017-09-21 19:00:00	155890
2017-09-21 20:00:00	123395
2017-09-21 21:00:00	103080
2017-09-21 22:00:00	95155
2017-09-21 23:00:00	80285

Moving Average (Son 3 gözlemin ortalaması):

(123395+103080+95155)/3 = 107210

Weighted Average ([0.1, 0.3, 0.6]):

(123395 * 0.1) + (103080*0.3) + (95155*0.6) = 100356

Smoothing (Düzeltme) Yöntemleri

Level: Serinin ortalamasıdır.

Single Exponential Smoothing (SES)

Sadece durağan serilerde çalışır. Trend ve mevsimsellik olmamalı.

Level'i yakalayabiliyor.

Üssel düzeltme yaparak tahminde bulunur.

Gelecek yakın geçmişle daha fazla ilişkilidir varsayımıyla geçmişin etkileri ağırlıklandırılır.

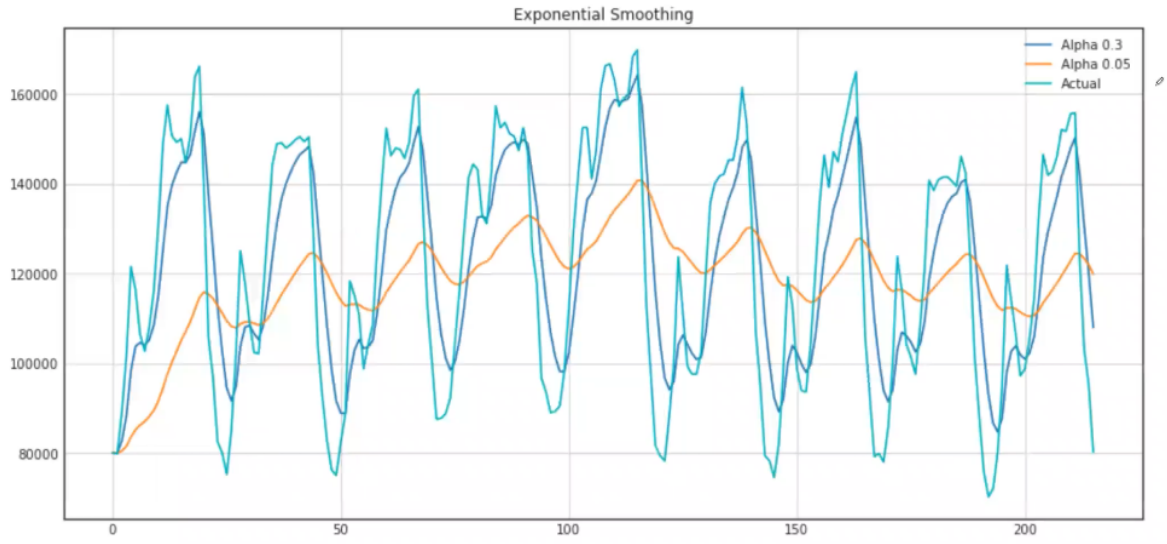
Geçmiş gerçek değerler ve geçmiş tahmin edilen değerlerin üssel olarak ağırlıklandırılmasıyla tahmin yapılır.

Formul:

$$\hat{y}_t = \alpha \cdot y_{t-1} + (1 - \alpha) \cdot \hat{y}_{t-1}$$

- y_hat t → tahmin edilecek zaman birimi
- yt - 1 →bir önceki gerçek değer (learning, bu değerden öğrenir)
- y_hat t - 1 → bir önceki tahmin edilen değer (remember, bu değerden hatırlar)
- a → smoothing factor. düzeltme faktörüdür. 0 ile 1 arasında değişir.
- Yani bir sonraki zaman değerini tahmin etmek için; bir önceki gerçek değere ve bir önceki tahmin edilen değere ihtiyacımız var.

Örnek:



Double Exponential Smoothing (DES)

Mevsimsellik olmamalıdır. Durağan ve trend serilerde çalışır.

Level + Trend'i yakalayabiliyor.

Trend etkisini göz önünde bulundurarak üssel düzeltme yapar.

DES = Level (SES) + Trend

SES'e ek olarak trend de dikkate alınır.

Trend içeren ve mevsimsellik içermeyen tek değişkenli zaman serileri için uygundur.

Formul:

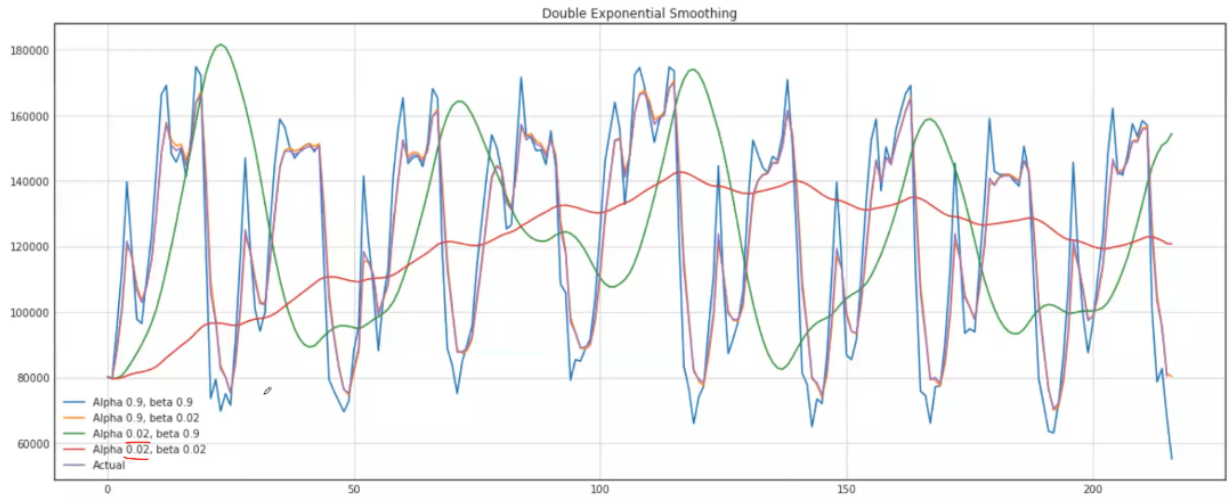
$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}$$

$$\hat{y}_{t+1} = \ell_t + b_t$$

- En sondaki eşitlik bize şunu söyler:
  - Bir sonraki periyot = level + trend
  - Yani: bir sonraki periyot = bir önceki dönemin ortalaması (level) + trend bilgisi

Örnek:



Triple Exponential Smoothing (a.k.a. Holt-Winters)

Level + Trend + Mevsimsellik yakalayabilir. Durağan, trend ve mevsimsel serilerde çalışır.

Level (SES) + Trend + Mevsimellik

Triple Exponential Smoothing en gelişmiş smoothing yöntemidir.

Bu yöntem dinamik olarak level, trend ve mevsimsellik etkilerini değerlendirerek tahmin yapmaktadır.

Trend ve/veya mevsimsellik içeren tek değişkenli serilerde kullanılabilir.

Formul:

$$\ell_t = \alpha(y_t - s_{t-p}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}$$

$$s_t = \gamma(y_t - \ell_t) + (1-\gamma)s_{t-p}$$

$$\hat{y}_{t+m} = \ell_t + mb_t + s_{t-p+1+(m-1)modp}$$

Smoothing Yöntemleri Özet Tablosu

Algorithm	Level	Trend	Seasonality	Tuning Parameters
Single HWES	Yes	No	No	α
Double HWES	Yes	Yes	No	α, β
Triple HWES	Yes	Yes	Yes	α, β, γ

Statistical Methods (İstatistiksel Yöntemler)

AR (Autoregression)

Formül: AR(p)

Burada hiperparametremiz p'dir.

Geçmiş gerçek değerleri modelleyen bir yöntemdir.

AR(1) model : $y_t = a_1 y_{t-1} + \epsilon_t$

AR(2) model : $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t$

AR(p) model : $y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \epsilon_t$

- Önceki zaman adımlarındaki gözlemlerin doğrusal bir kombinasyonu ile tahmin işlemi yapılır.
- Trend ve mevsimsellik içermeyen tek değişkenli zaman serileri için uygundur.
- p: zaman gecikmesi sayısıdır. p = 1 ise bir önceki zaman adımı ile model kurulmuş demek olur.

MA (Moving Average)

Formül: MA(q)

Burada hiperparametremiz q'dur.

Önceki zaman adımlarından elde edilen hataların (Et) doğrusal kombinasyonudur.

$$y_t = m_1\epsilon_{t-1} + \epsilon_t$$

$$y_t = m_1\epsilon_{t-1} + m_2\epsilon_{t-2} + \ldots + m_q\epsilon_{t-q} + \epsilon_t$$

- Önceki zaman adımlarında elde edilen hataların doğrusal bir kombinasyonu ile tahmin yapılır.
- Trend ve mevsimsellik içermeyen tek değişkenli zaman serileri için uygundur.
- q: zaman gecikmesi sayısıdır.

ARMA (AR + MA) (Autoregressive Moving Average)

Formül: ARMA(p, q) = AR(p) + MA(q)

SES'in kardeşidir.

$$y_t = a_1 y_{t-1} + m_1\epsilon_{t-1} + \epsilon_t$$

- Autoregressive Moving Average. AR ve MA yöntemlerini birleştirir.
- Geçmiş değerler ve geçmiş hataların doğrusal bir kombinasyonu ile tahmin yapılır.
- Trend ve mevsimsellik içermeyen tek değişkenli zaman serileri için uygundur.
- p ve q zaman gecikmesi sayılarıdır. p AR, q MA modeli içindir.

ARIMA (Autoregressive Integrated Moving Average)

Formül: ARIMA(p, d, q)

DES'in kardeşidir.

- Önceki zaman adımlarındaki farkı alınmış gözlemlerin ve hataların doğrusal bir kombinasyonu ile tahmin yapılır.
- Tek değişkenli, trendi olan fakat mevsimselliği olmayan seriler için uygundur.
- p: gerçek değer gecikme sayısı (otoregresif derece), p = 2 ise yt-1 ve yt-2 modeldedir.
- d: fark işlemi sayısı (fark derecesi)
- q: hata gecikmesi sayısıdır (hareketli ortalama derecesi)

SARIMA (Seasonal Autoregressive Integrated Moving Average)

Formül: SARIMA(p, d, q) (P, D, Q) m

TES'in kardeşidir

- ARIMA + mevsimselliktir
- Trend ve/veya mevsimsellik içeren tek değişkenli serilerde kullanılabilir.
- p, d, q ARIMA'dan gelen parametreler. Trend elemanlarıdır. ARIMA trend'i modelleyebiliyordu.
- p: gerçek değer gecikme sayısı (otoregresif derece). p = 2 ise yt-1 ve yt-2 modeldedir.

- d: fark işlemi sayısı (fark derecesi)
- q: hata gecikme sayısı (hareketli ortalama derecesi). q = 2 ise et-1 ve et-2 modeldedir.
- P, D, Q mevsimsel gecikme sayıları. Season elemanlarıdır.
- m tek bir mevsimlik dönem için zaman adımı sayısıdır. Mevsimselliğin görülme yapısını ifade eder.

Statistical Methods Özet Tablosu

Durağan	SES, AR, MA, ARMA
Trend	DES, ARIMA, SARIMA
Trend + Mevsimsellik	TES, SARIMA