Box-Cox ve Yeo-Johnson Dönüşümleri

Atıl Samancıoğlu

1 Giriş

Makine öğrenmesinde ve istatistiksel modellemede, birçok algoritma verinin dağılımının normal (gauss) olmasını bekler. Ancak ham veriler çoğunlukla çarpık (skewed) dağılır. Bu durumda dönüşümler kullanılarak veri daha simetrik hale getirilir. Bu dökümanda, bu amaçla yaygın olarak kullanılan **Box-Cox** ve **Yeo-Johnson** dönüşümlerini anlatacağız.

2 Neden Dönüşüm Uygulanır?

- Çarpık dağılımları daha normal dağılıma yaklaştırmak.
- Regresyon modellerinde hataların normal dağılmasını sağlamak.
- Lineer modellerin daha iyi çalışmasını sağlamak.
- Varyansı stabilize etmek.

Not

Box-Cox dönüşümü sadece **pozitif değerler** üzerinde çalışır. Yeo-Johnson ise hem pozitif hem de negatif değerlerle çalışabilir.

3 Box-Cox Dönüşümü

Matematiksel Formül

Bir gözlem y için Box-Cox dönüşümü aşağıdaki şekilde tanımlanır:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{eğer } \lambda \neq 0 \\ \ln(y), & \text{eğer } \lambda = 0 \end{cases}$$

Burada λ bir hiperparametredir ve en uygun değer genellikle veri üzerinden maksimum log-likelihood ile bulunur.

Örnek

Elimizdeki değerler şu olsun: $y = [1, 2, 4, 8, 16]$

- $\lambda = 0$ için: $\log(y)$ uygulanır: $\log([1, 2, 4, 8, 16]) = [0, 0.693, 1.386, 2.079, 2.773]$
- $\lambda = 0.5$ için:
$$\frac{\sqrt{y} - 1}{0.5} = 2 \times (\sqrt{y} - 1)$$
- $y = 4$ için: $\sqrt{4} = 2 \to 2 \times (2 - 1) = 2$

Tahminleri Geri Dönüştürme (Inverse Transform)

$$y = \begin{cases} (\lambda \cdot y^{(\lambda)} + 1)^{1/\lambda}, & \lambda \neq 0 \\ \exp(y^{(\lambda)}), & \lambda = 0 \end{cases}$$

Uyarı

Box-Cox sadece $y > 0$ olan durumlarda kullanılabilir. Eğer verinizde sıfır veya negatif değer varsa, Box-Cox uygulanmadan önce sabit bir sayı eklemek gerekir.

4 Yeo-Johnson Dönüşümü

Box-Cox dönüşümünün negatif sayılarda da uygulanabilen genişletilmiş halidir.

Matematiksel Formül

$$y^{(\lambda)} = \begin{cases} \frac{[(y+1)^{\lambda} - 1]}{\lambda}, & y \geq 0, \lambda \neq 0 \\ \ln(y + 1), & y \geq 0, \lambda = 0 \\ \frac{-[(-y+1)^{2-\lambda} - 1]}{2-\lambda}, & y < 0, \lambda \neq 2 \\ -\ln(-y + 1), & y < 0, \lambda = 2 \end{cases}$$

Örnek

Elimizdeki değerler: $y = [-3, -1, 0, 1, 3]$

- $\lambda = 0$ için:

  $y = [-3, -1, 0, 1, 3] \Rightarrow y^{(\lambda)} = [-\ln(4), -\ln(2), 0, \ln(2), \ln(4)] \approx [-1.386, -0.693, 0, 0.693, 1.386]$

Geri Dönüştürme (Inverse Yeo-Johnson)

Formüllerin tersine çevrilmesi ile yapılır. Örneğin:

- $y \geq 0, \lambda \neq 0$ için:
$$y = (y^{(\lambda)} \cdot \lambda + 1)^{1/\lambda} - 1$$
- $y < 0, \lambda \neq 2$ için:
$$y = 1 - [-(y^{(\lambda)} \cdot (2 - \lambda) + 1)]^{1/(2-\lambda)}$$

5 Hangisi Ne Zaman Kullanılır?

- Eğer tüm değerleriniz pozitifse: **Box-Cox** önerilir.
- Eğer verilerinizde negatif veya sıfır değerler varsa: **Yeo-Johnson** dönüşümünü tercih etmelisiniz.

Not

Sklearn kütüphanesinde `PowerTransformer` sınıfı, her iki yöntemi de destekler:

- `method='box-cox'`: sadece pozitif veriler için.
- `method='yeo-johnson'`: tüm reel değerler için.

6 Sonuç

Box-Cox ve Yeo-Johnson dönüşümleri, çarpık veri dağılımlarını daha simetrik hale getirerek modellerin performansını artırabilir. Özellikle lineer modellerde ve varsayımları normal dağılım gerektiren durumlarda önemli avantaj sağlarlar. Tahmin sonrası inverse transform uygulanarak orijinal ölçekle yorumlama yapılabilir.