Harf Değerleri

Bir veri kümesinin özetlenmesinde geleneksel olarak örneklem ortalaması

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

ve standart sapması

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

kullanılır.

Bu tahmin ediciler normallik varsayımı altında en etkin tahmin edicilerdir. Ancak veri setinin dağılımı normal değilse ya da veri setinde aykırı değerler varsa bu tahmin edicilerin etkinlikleri hızla düşer. Bu nedenle örneklem ortalaması ve standart sapması dayanıklı (robust) tahmin ediciler değillerdir.

Harf değerleri,

- Veri kümesinin konumunun ve yayılma miktarının robust tahmin edicilerini elde etmeye ve
- Veri kümesindeki aykırı değerleri belirlemeye

yardımcı olur.

Harf değerleri veri kümesini özetlemek için sınıflama ve sıralamayı kullanmaktadır.

Sınıflama ve Sıralama:

$X_1, X_2, \ldots, X_n$ bağımsız ve aynı dağılımlı rastgele değişkenler olsun. Bu rastgele değişkenler küçükten büyüğe doğru sıralandığında $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ sıralı istatistik olarak adlandırılır. Burada, $X_{(i)},\ i = 1, \ldots, n,$ i. sıralı istatistiktir.

Derinlik (Depth): Bir örneklemdeki gözlem değerinin derinliği, o gözlemin yukarıdan aşağı olan sırası ile aşağıdan yukarı olan sırasının küçük olanıdır.

Rank: Gözlem değerinin rankı o gözlemin yukarıdan aşağı olan sırasıdır.

Sıra istatistiklerinin en popüler olanı medyandır.

$$Medyan = \frac{1}{2}\left(X_{(k)} + X_{(k+1)}\right) \; ; \qquad n = 2k$$

$$Medyan = \; X_{(k)} \; ; \qquad n = 2k - 1$$

$$Medyanın\ derinliği = \frac{n+1}{2}$$

Örnek: $n = 3$ ise $Medyanın\ derinliği = \frac{3+1}{2} = 2$ ve $Medyan = X_{(2)}$ olur.

$n = 4$ ise $Medyanın\ derinliği = \frac{4+1}{2} = 2.5$ ve $Medyan = \frac{1}{2}\left(X_{(2)} + X_{(3)}\right)$ olur.

Not: Uç değerler veri kümesindeki en küçük ve en büyük gözlem değerleridir. Aykırı değerlerle karıştırılmaması gerekmektedir.

Tanım: Bir veri kümesinde veri kümesinin ortadaki %50'lik kısmını kapsayan değere alt dörtlük ve üst dörtlük denir.

$$Dörtlüğün\ derinliği = \frac{[medyanın\ derinliği] + 1}{2}$$

$[x]$; x'i aşmayan en büyük tam sayı şeklinde tanımlanmıştır.

Konum parametresinin aykırı değerlerden etkilenmeyen bir diğer tahmin edicisi $trimean$ dir.

$$trimean = \frac{1}{4}(alt\ dörtlük) + \frac{1}{2}(medyan) + \frac{1}{4}(üst\ dörtlük)$$
$$= \frac{1}{4}F_L + \frac{1}{2}M + \frac{1}{4}F_U$$

şeklinde tanımlanır. Trimean çarpık dağılımlarda da etkin sonuçlar verir.

Verilerin konumunu tahmin etmek için yaygın olarak medyan ve trimean kullanılırken veri kümesinin yayılımını tahmin etmek için basit bir robust tahmin edicisi olan *Dörtlük Yayılımı* ($Fourth\ spread$) kullanılır. Dörtlük Yayılımı

$$Dörtlük\ Yayılımı = üst\ dörtlük - alt\ dörtlük$$
$$d_F = F_U - F_L$$

olarak ifade edilir.

Aykırı değerlerin belirlenmesi: Dörtlük yayılımı (d_F) kullanılarak elde edilen kesim (cut off) noktaları (c_L ve c_U) dışında kalan gözlemler aykırı değer olarak belirlenir.

Burada

$$alt\ kesim\ noktası: c_L = F_L - 1.5d_F$$

ve

$$üst\ kesim\ noktası: c_U = F_U + 1.5d_F$$

şeklinde hesaplanır.

Eğer gözlem değeri c_L değerinden küçük veya c_U değerinden büyük ise bu gözlem değeri aykırı değer olarak belirlenir.

Örnek: $N(\mu, \sigma^2)$ dağılımı için

$$c_L = \mu - 2.698\sigma \quad \text{ve} \quad c_U = \mu + 2.698\sigma$$

olarak hesaplanır. Buradan,

$$(c_L, c_u) = (\mu - 2.698\sigma, \mu + 2.698\sigma)$$

aralığı dışında kalan değerler aykırı değer olarak belirlenir.

$N(0,1)$ dağılımı için

$(c_L, c_u) = (-2.698, 2.698)$ olup $P(X < c_L) = P(X > c_U) = 0.0036$ olarak hesaplanır. Bu durumda, veri setindeki gözlemlerin $2 \times 0.0036 = 0.0072$'sinin aykırı değer olduğu söylenir. Veri setinde $n = 400$ gözlem olsaydı; $400 \times 0.0072 = 2.88$ olurdu. Bu durumda, tüm gözlemlerin 2 ya da 3 tanesinin aykırı değer olması beklenirdi.

Not: Bir veri setinde, gözlem sayısının %5'i ile %10'u kadarının aykırı değer olma olasılığı vardır.

Tanım: Medyan, dörtlükler ve uç değerler 5 değerli özet olarak adlandırılır.

Örnek: Aşağıdaki veri seti için 5 değerli özeti hesaplayalım.

$$28, 43, 87, 47, 49, 36, 57, 65, 27, 59, 91, 102, 95$$

İlk önce veriler küçükten büyüğe doğru sıralanır.

i	1	2	3	4	5	6	7	8	9	10	11	12	13
$X_{(i)}$	27	28	36	43	47	49	57	59	65	87	91	95	102

$n = 13$

$Medyanın\ derinliği = \dfrac{13+1}{2} = 7$

$Medyan = X_{(7)} = 57$

$Dörtlüğün\ derinliği = \dfrac{[7]+1}{2} = 4$

$Alt\ dörtlük = X_{(4)} = 43$

$Üst\ dörtlük = X_{(10)} = 87$

$Uç\ değerler: X_{(1)} = 27,\ X_{(13)} = 102$

$Fourth\ spread = 87 - 43 = 44$

$alt\ kesim\ noktası: c_L = 43 - (1.5)(44) = -23$

$üst\ kesim\ noktası: c_U = 87 + (1.5)(44) = 153$

Bu veri setinde, $c_L = -23$ değerinden küçük ve $c_U = 153$ değerinden büyük gözlem bulunmadığından aykırı değer yoktur.

Daha büyük veri kümelerinde özetleyici değer çiftlerini eklemeye devam edebiliriz. Örneğin 8'lik, 16'lık v.b. Sekizliğin derinliği

$$sekizliğin\ derinliği = \frac{[dörtlüğün\ derinliği]+1}{2}$$

olarak hesaplanır.

Önceki örnekte,

$$sekizliğin\ derinliği = \frac{[4]+1}{2} = \frac{5}{2} = 2.5$$
$$alt\ sekizlik = \frac{X_{(2)} + X_{(3)}}{2} = \frac{28+36}{2} = 32$$
$$üst\ sekizlik = \frac{X_{(11)} + X_{(12)}}{2} = \frac{91+95}{2} = 93$$

olacaktır. Sekizliklerin de yer aldığı özetlere 7 değerli özet denir.

Formülü genelleştirirsek, genel formül,

$$\frac{[önceki\ derinlik]+1}{2}$$

şeklinde ifade edilir.

Not: Derinlik 1'e ulaştığında işlem durdurulur, çünkü uç değerlere ulaşılmış olur.