

Açıklama ve uyarılar: Ödev *el yazısı ile hazırlanacak* ve yanıtlar pdf dosyasına dönüştürüldükten sonra en geç 7 Aralık tarihinde uzak.etu.edu.tr sistemi üzerinden teslim edilecektir. E-posta vb farklı yollarla gönderilen ya da geç teslim edilen ödevler dikkate alınmayacaktır.

Sorular

1. K-kat çapraz-geçerleme yöntemine yönelik olarak aşağıdaki soruları yanıtlayınız:

(a) K-kat çapraz-geçerleme nasıl uygulanır? İzlenecek *adımlar şeklinde* açıklayınız.

Yanıt: K-kat çapraz-geçerleme yöntemi aşağıdaki adımları içerir:

1. Eldeki veri seti birbiriyle örtüşmeyecek şekilde ve yaklaşık aynı büyüklükte olan k parçaya bölünür.
2. Parçalardan biri test veri seti olarak tutulur. Kalan $k - 1$ parça ise birlikte eğitim seti olarak kullanılarak tahmin yapılır.
3. Test için ayrılan parça kullanılarak tahmine ait hata oranı hesaplanır.
4. Her bir parça sırayla test seti olacak şekilde yukarıdaki işlem yinelenir.
5. Elde edilen k adet test hata oranının ortalaması hesaplanır.

(b) K-kat çapraz-geçerleme yönteminin *bir-eksiltmeli çapraz-geçerleme* yaklaşımına göre üstünlükleri ve sakıncaları nelerdir?

Yanıt: Bir-eksiltmeli ÇG yönteminde yanlılık olabilecek en düşük derecededir, ancak varyans yüksektir. K-kat ÇG’de ise yanlılık daha yüksek olmakla birlikte varyans düşüktür. K-kat ÇG bu şekilde test seti ile geçerleme seti büyüklükleri arasında bir denge kurarak uygulamada daha iyi sonuçlar üretir. Ayrıca k-kat ÇG’de hesaplama yoğunluğu da bir-eksiltmeli ÇG’ye göre (doğrusal regresyon hariç) daha düşüktür.

(c) K-kat çapraz-geçerleme yönteminin *geçerleme seti* yaklaşımına göre üstünlükleri ve sakıncaları nelerdir?

Yanıt: Geçerleme seti yönteminde sonuçlar veri setinin nasıl ikiye bölündüğüne bağlıdır. Dolayısıyla sonuçlar yüksek varyans gösterir. Ayrıca geçerleme setinde verilerin önemli bir bölümü kullanılmadığı için hesaplanan hata oranı gerçek değerden yüksek olabilir. Bu iki nedenle k-kat ÇG yöntemi, geçerleme seti yaklaşımından daha üstündür.

2. Özyetim yöntemine yönelik olarak aşağıdaki soruları yanıtlayınız:

(a) Yatay kesit verilerinde özyetim nasıl uygulanır? *Adımlar şeklinde* açıklayınız.

Yanıt: Yatay kesit verilerinde özyetim yöntemi aşağıdaki adımları içerir:

1. Eldeki n büyüklüğündeki veri setinden n adet gözlem çekilir. Ancak her seferinde çekilen değer yerine geri koyulur.
2. Yukarıdaki işlem çok sayıda yinelenerek topluluk adı verilen Ω adet özyetim çoğaltması üretilir.
3. Üretilen çoğaltmalar kullanılarak çok sayıda tahmin hesaplanır.
4. Bu tahminler birleştirilerek ilgi duyulan değere ilişkin yüksek kesinlikli güven aralıkları ve kestirimler elde edilir.

(b) Bir özyetim çoğaltmasında $n = 4$ büyüklüğündeki asıl veri setine ait 4. gözlemin özyetimdeki *ilk gözlem olma* olasılığı nedir? Açıklayınız.

Yanıt: Özyetim çoğaltmaları hesaplanırken gözlemler baştaki asıl veri setinden rastsal olarak seçilir. Dolayısıyla belli bir gözlemin özyetim setindeki ilk gözlem olma olasılığı $1/n = 1/4$ 'tür.

(c) Bir özyetim çoğaltmasında ilk gözlem asıl veri setindeki 4. gözlem olarak belirlenmişse, $n = 4$ büyüklüğündeki asıl veri setine ait 4. gözlemin özyetimdeki *ikinci gözlem olma* olasılığı nedir? Açıklayınız.

Yanıt: Özyetim çoğaltmaları hesaplanırken her seferinde çekilen değer yerine geri koyulur. Dolayısıyla belli bir gözlem daha önce seçilmişse bile tekrar seçilme olasılığı $1/n = 1/4$ olur.

(d) Büyüklüğü $n = 4$ olan veri setindeki 4. gözlemin özyetim çoğaltmasında *hiç yer almama* olasılığı nedir? Açıklayınız.

Yanıt: Belli bir gözlemin özyetim çoğaltmasındaki 4. gözlem olmama olasılığı $1 - 1/4$ 'tür. Asıl veri setinden toplam $n = 4$ kez gözlem seçileceği için çarpım kuralını uygulayarak belli bir gözlemin hiç seçilmeme olasılığını $(1 - 1/n)^n = (1 - 1/4)^4 = 81/256$ olarak buluruz.

(e) Büyüklüğü $n = 4$ olan veri setindeki 4. gözlemin özyetim çoğaltmasında *en az 1 kez yer alma* olasılığı nedir? Açıklayınız.

Yanıt: Belli bir gözlemin özyetim çoğaltmasında en az 1 kez yer alma olasılığı, hiç yer almama olasılığının tersidir. Bir önceki soruyu temel alırsak $n = 4$ için bu olasılığı $1 - 81/256 = 175/256$ şeklinde hesaplayabiliriz.

TOBB - Ekonomi ve Teknoloji Üniversitesi
İKT457 – Ekonomi ve Finans için Yapay Zeka I, Dördüncü Ödevi

3. Bu soruyu yanıtlamak için değişken adları aşağıda verilen ve büyüklüğü $n = 400$ olan “Veriseti” adlı R veri çerçevesini temel alınız:

```
> names(veriseti)
[1] "Y"  "X"
>
```

- (a) Veri setinden 200 gözlemlik bir rastsal alt örneklem belirleyen “egitim” adlı bir nesne oluşturacak R komut(lar)ını yazınız.

Yanıt:

```
egitim <- sample(400, 200)
```

- (b) Önceki soruda belirlenen eğitim setine $Y = \beta_1 + \beta_2 X$ şeklindeki modeli uygulayacak ve daha sonra geriye kalan test setini kullanarak hata kareleri ortalamasını hesaplayacak R komut(lar)ını yazınız.

Yanıt:

```
lm.fit <- lm(Y~X, subset=egitim)
mean((Y - predict(lm.fit, Veriseti))[-egitim]^2)
(Dikkat: “lm.fit” nesnesine farklı ad verilebilir.)
```

- (c) Önceki sorudaki modele ait test hata oranını *bir-eksiltmeli çapraz-geçerleme* yöntemiyle hesaplayacak R komut(lar)ını yazınız.

Yanıt:

```
glm.fit <- glm(Y~X, data=Veriseti)
cv.err <- cv.glm(Veriseti, glm.fit)
cv.err$delta
(Dikkat: “glm.fit” ve “cv.err” nesnelere farklı ad verilebilir.)
```

- (d) Önceki sorudaki modele ait test hata oranını *10-kat çapraz-geçerleme* yöntemiyle hesaplayacak R komut(lar)ını yazınız.

Yanıt:

```
glm.fit <- glm(Y~X, data=Veriseti)
cv.err <- cv.glm(Veriseti, glm.fit, K=10)
cv.err$delta
(Dikkat: “glm.fit” ve “cv.err” nesnelere farklı ad verilebilir.)
```