1906003132015

Doğal Dil İşleme

BAİBÜ Bilgisayar Müh.

Dr. Öğr. Üyesi İsmail Hakkı Parlak

ismail.parlak@ibu.edu.tr

Oda: 335

Çapraz Doğrulama (Cross Validation)

```
clf = DecisionTreeClassifier(criterion="entropy", max_depth=3)
clf.fit(X_train, y_train)
clf.score(X_test, y_test) -> 0.85

clf = DecisionTreeClassifier(criterion="entropy", max_depth=7)
clf.fit(X_train, y_train)
clf.score(X_test, y_test) -> 0.89

clf = DecisionTreeClassifier(criterion="gini", max_depth=5)
clf.fit(X_train, y_train)
clf.score(X_test, y_test) -> 0.91
```

Modelin test setine ***overfit*** etmesi.

Çapraz Doğrulama (CV)

- Modeller için farklı ayarları (hiper-parametreler) değerlendirirken, örneğin bir KA için manuel olarak ayarlanması gereken `max_depth` parametresi gibi, tahmin edici en iyi şekilde performans gösterene kadar parametrelerde ince ayar yapılabileceğinden, test setinde aşırı uyum (overfitting) riski mevcuttur.
- Bu şekilde, test seti hakkındaki bilgi modele "sızabilir" (leak) ve değerlendirme metrikleri artık genelleme performansı hakkında rapor veremez.
- Bu sorunu çözmek için veri setinin başka bir kısmı "doğrulama seti" (validation set) olarak adlandırılabilir: eğitim, eğitim (train) seti üzerinde devam eder, ardından doğrulama seti üzerinde değerlendirme yapılır ve deney başarılı göründüğünde, son değerlendirme test seti üzerinde yapılabilir.
- Bununla birlikte, mevcut verileri üç kümeye bölerek, modeli öğrenmek için kullanılabilecek örnek sayısını büyük ölçüde azaltırız ve sonuçlar bir rastgele seçime bağlı olabilir.

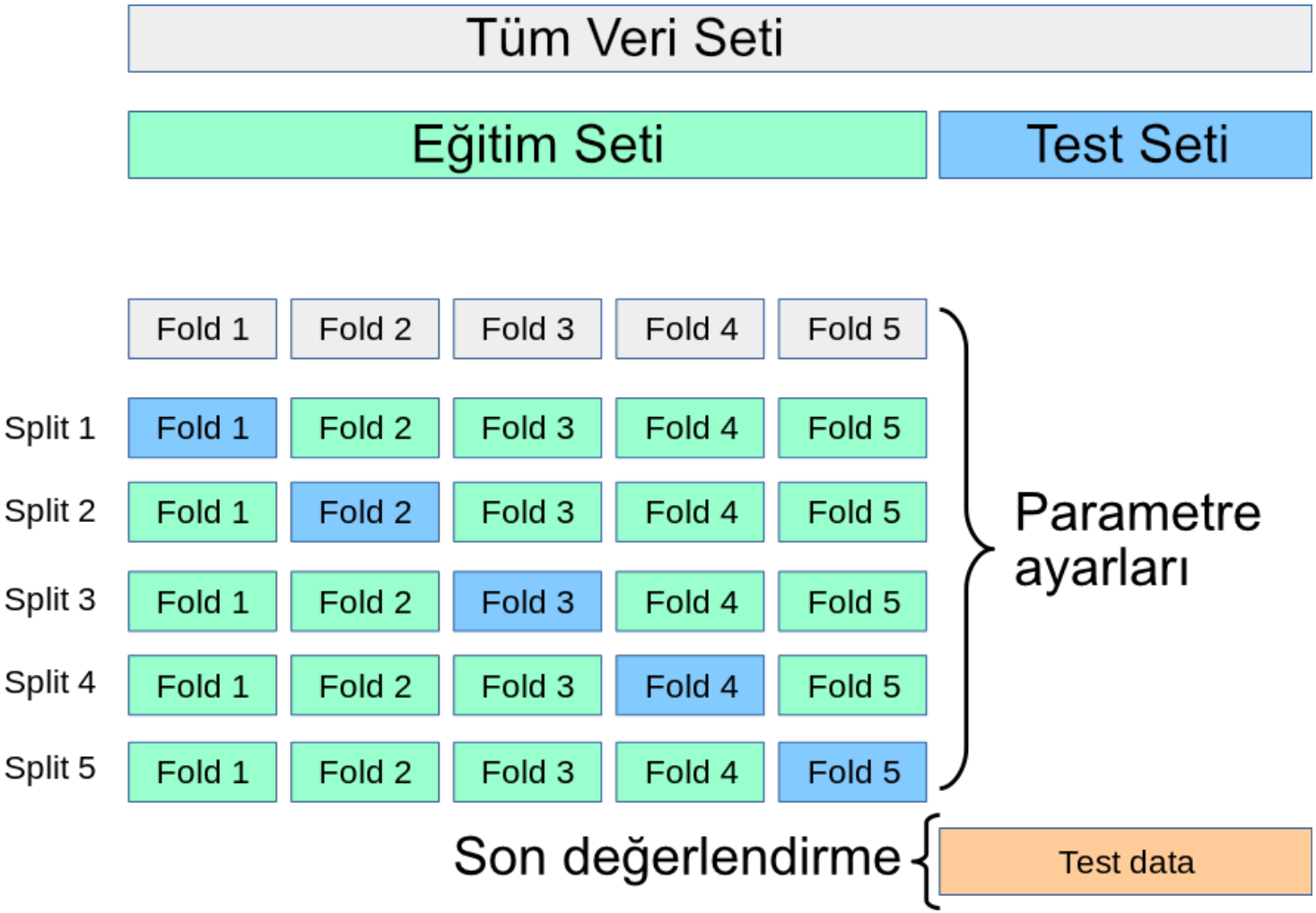
https://scikit-learn.org/stable/modules/cross_validation.html

Çapraz Doğrulama (CV)

- Bu soruna bir çözüm, çapraz doğrulama (kısaca CV) adı verilen bir prosedürdür. Nihai değerlendirme için yine de bir test seti hazırlanmalıdır, ancak CV hazırlanırken artık doğrulama setine ihtiyaç duyulmamaktadır.
- N-Katlamalı (N-Fold) Çapraz Doğrulama olarak adlandırılan temel yaklaşımda, eğitim seti *N* tane daha küçük sete bölünür. Her "katlama" için aşağıdaki prosedür izlenir:
  - Model, eğitim verileri olarak katlamalar kullanılarak eğitilir.
  - Modelin eğitim başarısı katman döngüsündeki başarıların ortalaması olarak alınır.
  - Ortaya çıkan model, verilerin geri kalan kısmı üzerinde doğrulanır.

https://scikit-learn.org/stable/modules/cross_validation.html

Çapraz Doğrulama (CV)



https://scikit-learn.org/stable/modules/cross_validation.html

Bir Modelin En İyi Parametrelerini Aramak

- Hiper-parametreler, kullanıcının Makine Öğrenimi modelini oluştururken belirttiği değişkenlerdir.
- Hiper-parametrelerin değerlerine modeli oluşturan kullanıcı tarafından karar verilir.
- Peki en uygun hiper-parametrelere nasıl ulaşabiliriz?
- Izgara Arama (Grid Search), belirtilen tüm hiper-parametrelerin ve değerlerinin farklı bir kombinasyonunu kullanır ve her bir kombinasyonun başarısını hesaplar ve hiper-parametreler için en iyi değeri seçer. Bunun karşılığında ise hesaplama zamanı ve toplam işlemsel masraf artmış olur.