Model Seçimi ve Arttırımı Yöntemleri

Çok sık gelen ve doğal olarak her veri bilimi uzmanının bilmesi gereken önemli konulardan birisi model seçiminin nasıl yapılacağıdır. Bu bölümde, çok sayıda alternatif modelden nasıl seçim yapılacağını öğrenirken aşağıdaki soruları cevaplamaya çalışacağız.

1. bağlantı / varyans (bias / variance) ikileminde bir model nasıl seçilir ve performansı nasıl değerlendirilir
2. Hiper parametreler için değer seçimi nasıl yapılır (hiper parametre öğrenme sürecinde yer almayan parametrelerdir)
3. iş süreçleri analiz edilen bir problem için en uygun modeli nasıl buluruz?

Yukarıdaki soruların cevaplanabilmesi için aşağıdaki iki yöntemi bu bölümde anlatacağız:

1. k-katlamalı çapraz doğrulama (k-fold Cross validation)
2. Izgara Araması (Grid Search)

Dersin sonunda, Bonus kısmında anlattığımız XGBoost algoritması ile bu derste öğrendiklerimizi birleştireceğiz.

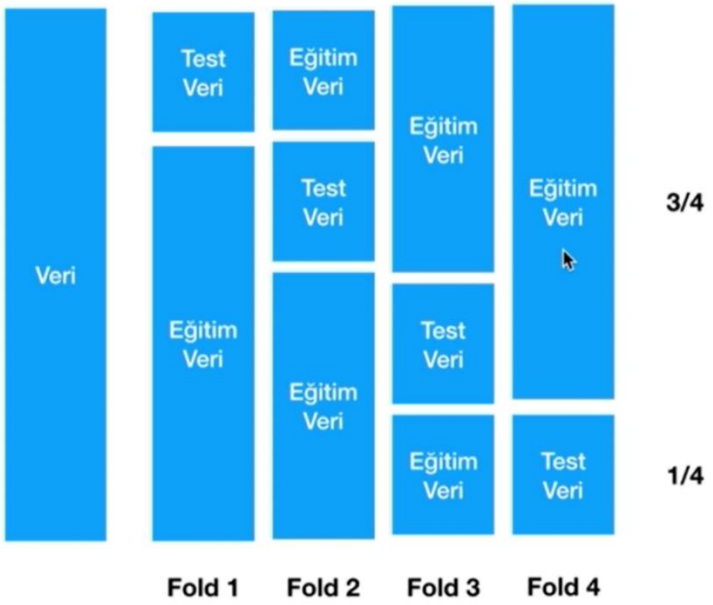
Model Değerlendirmesi: k-fold Cross Validation

Model seçiminden önce dikkat edilmesi gereken ilk nokta, model değerlendirmesidir. (evaluation)

Şimdiye kadar test kümesindeki başarıyı ölçtük.

Tüm kümeyi 2/3 training ve 1/3 test set olarak bölüyorduk. Test set üzerinde başarı ölçüyorduk. Ancak burada sonuç hagi 1/3 lük kısmın alındığına göre değişiyor.

k-fold için ise ¼ test ¾ training alınır. Daha sonra ¼ kaydırılır yani diğer kısım alınır, bu şekilde 4 kez aynı işlem yapılır ve sonuçta tüm data training ve test sınıfları arasında bulunmuş olur:



Sonuç olarak bu 4 farklı test ve eğitim verisi için system eğitilip test ediliyor. Ve daha sonra farklı yöntemlerde evaluation tamamlanıyor, min alınabilir max alınabilir mean alınabilir vb.

Model Seçimi – GridSearchCV

Hangi modeli kullanmamız gerektiğine nasıl karar vereceğiz?

1. Bağımlı değişken var mı? Ya da labeled data mı? (O halde ya classification ya da regression problemi – clustering değil) Labeled data yok ise sonuçta problem unsupervised demektir.
2. Bağımlı değişken yani output var ise ikinci soruya geçelim, continuous mu yoksa kategorik mi? Sürekli ise regression, categoric ise classification.
3. Ancak şunun altını da çizmek lazım çoğu durumda regression problemleri classification algoritmaları ile de kullanılabilir. Ya da vice versa. Mesela müşteri bizi bırakıyor mu bırakmıyor mu? Bırakanlara 1 bırakmayanları 0 dedik. Bu categoric görünüyor ama. Ben 0-1 arası continuous değer alarak da bu problem çözebilirim.
4. Ayrıca problemin doğrusal olup olmadığına da bakılabilir.

Peki algoritmayı seçtikten sonra algoritmanın parametrelerini nasıl optimize ederiz?

Bu işi GridSearchCV ile yapabiliriz.