

Derin Öğrenmeye Giriş

HAFTA 4

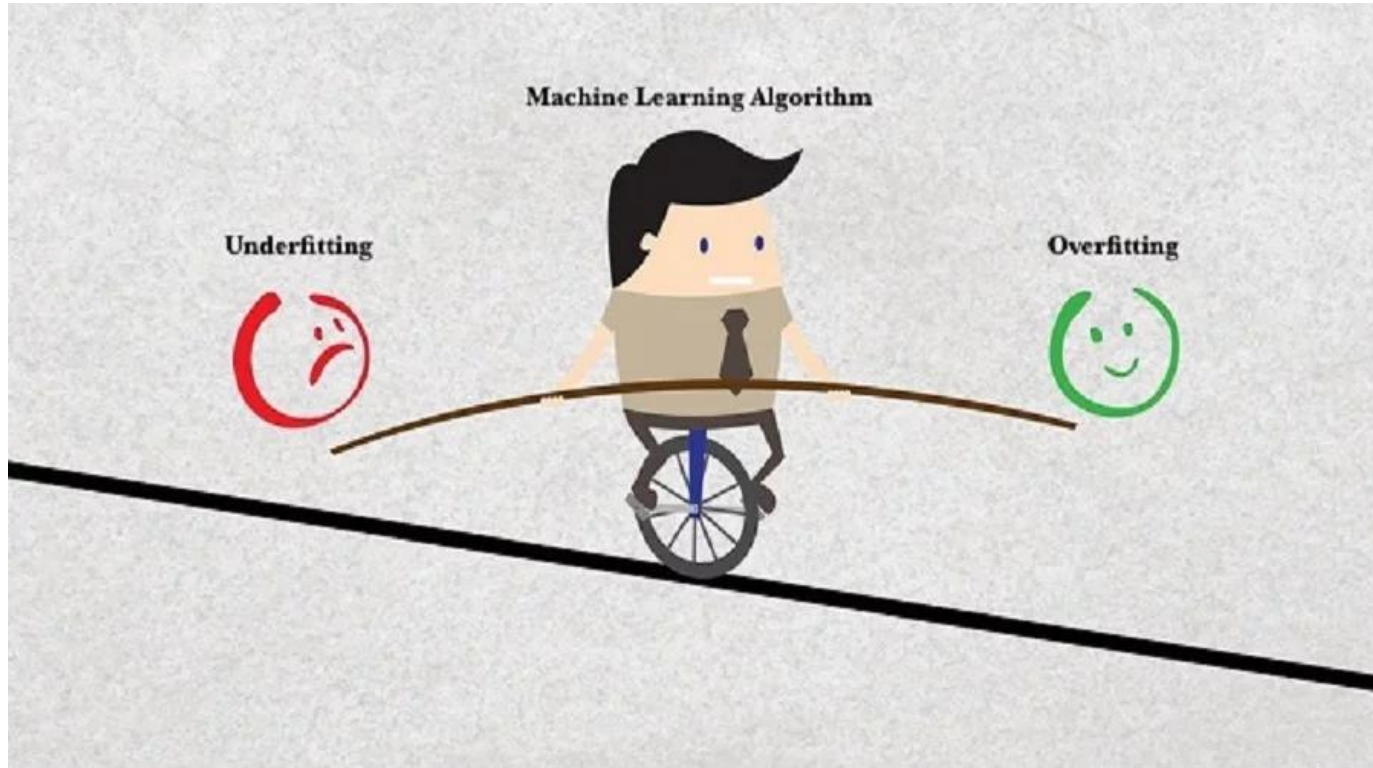
Underfitting (Eksik Öğrenme)

Overfitting (Aşırı Öğrenme)

Dr. Öğretim Üyesi Burcu ÇARKLI YAVUZ

bcarkli@sakarya.edu.tr

Underfitting & Overfitting



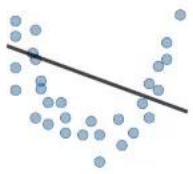


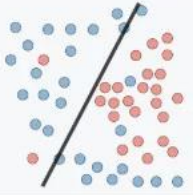
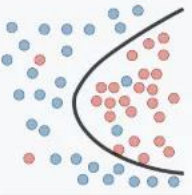
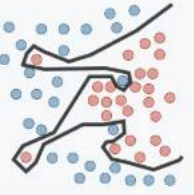
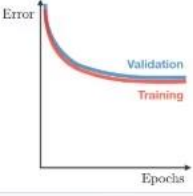
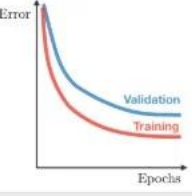
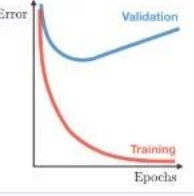
Underfitting & Overfitting

- **Aşırı öğrenme (over fitting)** , algoritmanın eğitim verisi üzerinden en alt kırılima kadar çalışıp, sonuçları ezberlemesi ve sadece o veriler üzerinde başarı elde edebilmesidir.
- Aşırı öğrenme problemi ile karşılaştığımızda, eğitim verisi ile kurduğunuz modeli, test verisi üzerinde çalıştırdığınızda muhtemelen sonuçlar eğitim verisine göre çok düşük çıkacaktır.
- Modelimizin amacı, her şeyi tahmin etmesi değil genel bir doğru elde etmesi yani genel bir örüntü bulmasıdır ve bu genel doğrunun(örüntünün) sonraki verilere de uygulanabiliyor olmasıdır.
- Algoritma çok karmaşık ise veri içindeki örüntüyü bulmak yerine, öğrenme süreci gürültüyü ezberlemek ile sonuçlanabilir.
- Öğrencinin sınava girmeden önce bilgileri çok iyi bir şekilde ezberleyip, sınavda farklı türde sorular ile karşılaştığında sınavda başarısız olması gibi.
- Aşırı öğrenme problemi olan modeller yüksek varyans problemi içerebilir.

Underfitting & Overfitting

- **Az öğrenme (underfitting)**, modelin verilerdeki temel örüntüleri yakalamak için çok basit olması ve bu nedenle kötü performans göstermesi olarak tanımlanır.
- Bu modeller eğitim verilerini çok yakından takip etmek yerine, eğitim verilerinden alınan dersleri yok sayar ve girdiler ile çıktılar arasındaki temel ilişkiyi öğrenemez.
- Az öğrenme durumunda model hem eğitim hem de test verilerinde başarısız bir performans gösterir.
- Öğrencinin sınava hiç çalışmadan girip sınavdan da doğal olarak kötü not alması gibi.

Underfitting & Overfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none">• High training error• Training error close to test error• High bias	<ul style="list-style-type: none">• Training error slightly lower than test error	<ul style="list-style-type: none">• Very low training error• Training error much lower than test error• High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none">• Complexity model• Add more features• Train longer		<ul style="list-style-type: none">• Perform regularization• Get more data

Az öğrenme problemini nasıl çözebiliriz?

- Az öğrenmeye sebep olan başlıca nedenler modelin basit yapısı, değişken sayısının yetersiz olması, yetersiz veri, gürültülü veri olarak sayılabilir.
- Az öğrenme, aşırı öğrenmeye göre önüne geçilmesi daha kolay bir sorundur.
- Eğer sorun modelin basit yapısı ise bu sorunu gidermek için modelimizin kapasitesini artırmamız gerekir. Modelimizin kapasitesini modelin yapısında (örneğin katmanlardaki nöron sayılarını yada katman sayılarını arttırmak) değişikliklere giderek artırabiliriz.
- Eğer problem veri sayısının azlığından ya da özellik sayısının azlığından kaynaklanıyorsa veri sayısını ya da özellik sayısını arttırmak çözüm olacaktır.

Aşırı öğrenme problemini nasıl çözebiliriz?

➤ Değişken Sayısını Azaltmak

- Aşırı öğrenmeyi önlemek için korelasyon, eksik değer ve aykırı değer analizleri büyük önem taşır. Örneğin, veri setinde yüksek korelasyona sahip bağımsız değişkenlerin varlığı, aynı bilgiyi taşımaları nedeniyle hem yanlılığa hem de aşırı öğrenmeye neden olabilir. Bu nedenle veri setinde yüksek korelasyon gösteren değişkenlerin yeniden irdelenmesi gerekebilir.
- Doğru açıklayıcı değişkenleri bularak basit bir model kurmak daha mantıklıdır.

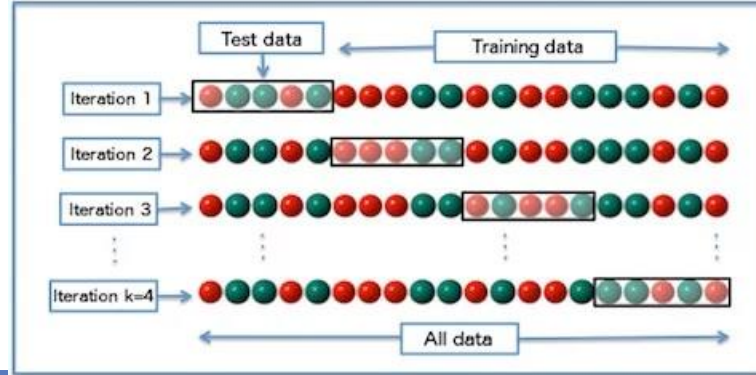
➤ Daha fazla veri eklemek

- Eğer aşırı öğrenme problemi, eğitim verisinde az veri olmasından, dolayısıyla tek tip veri olmasından kaynaklanıyor ise daha fazla çeşitli veri eklemek gerekir.
- Burada engele takılmamak için veri hazırlığını dikkatli yapmak, eğitim verisi ve test verisi ayrımını dikkatli incelemekte fayda var.

Aşırı öğrenme problemini nasıl çözebiliriz?

➤Çapraz doğrulama (cross-validation)

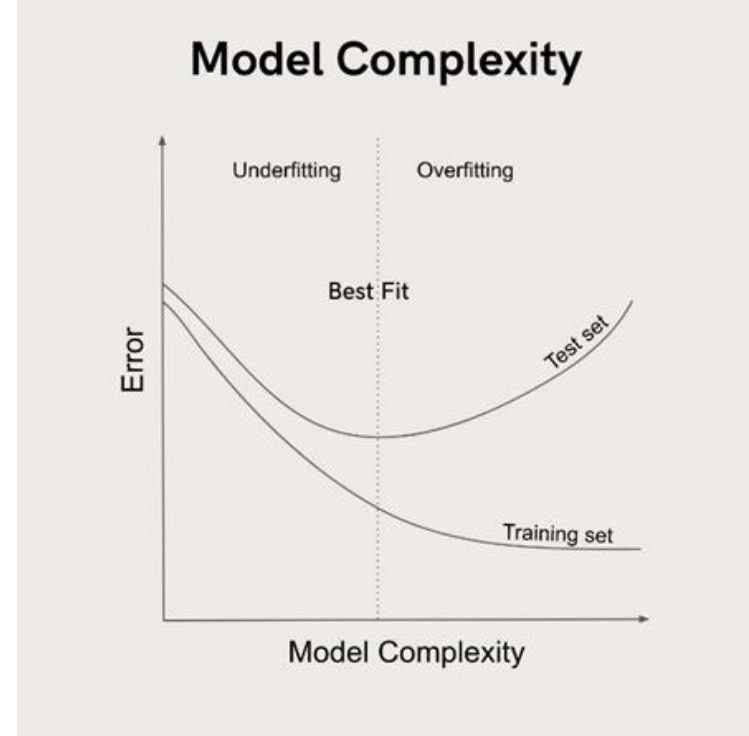
- Sonuçları doğru değerlendirmek için eğitim verisi ve test verisinin benzer özelliklere sahip olması gerekir. K katlamalı çapraz doğrulama ile raslantısallık azaltılarak sonuç metriklerinin tutarlılığı sağlanmaktadır.
- K katlamalı çapraz doğrulama (k-fold cross validation) yöntemi kullanarak verinin tüm parçalarının eğitim ve test verisinde yer almasıyla daha doğru bir öğrenme süreci oluşturmak, performans göstergelerini bu şekilde incelemek aşırı öğrenme hakkında bize daha net bilgi verecektir.



Aşırı öğrenme problemini nasıl çözebiliriz?

➤ Erken Durdurma (Early stopping)

- İki hatanın birbirinden ayrılmaya başladığı nokta (çatallanmanın başladığı nokta) itibariyle aşırı öğrenme başlamış demektir.
- Eğitim verisi ile test verisi hataları arasındaki fark açıklığı belli bir seviyeye geldiğinde eğitimi durdur.
- Bazı algoritmalar bunu sürekli kontrol ederek otomatik olarak yapar.



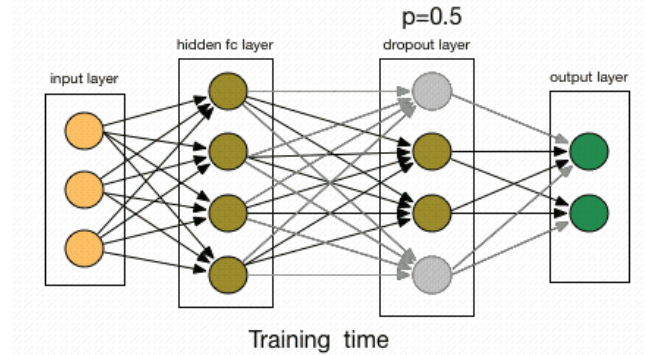
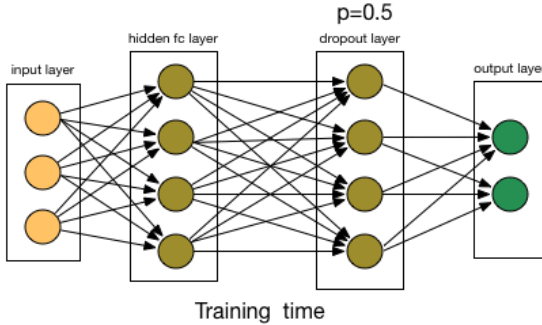
Aşırı öğrenme problemini nasıl çözebiliriz?

➤ Düzenleştirme(Regularization)

- Düzenleme, modelin karmaşıklığını azaltmak için bir kullanılan tekniktir. Bunu kayıp fonksiyonunu cezalandırarak yapar. Yani modelde ağırlığı yüksek olan değişkenlerin ağırlığını azaltarak bu değişkenlerin etki oranını azaltır.
- Ridge (L1) Regresyonu ve Lasso (L2) Regresyonu düzenleştirme çözümlerinin aralarındaki fark cezalandırma derecesidir.
- Ridge Regresyon önemsiz değişkenin kat sayısını azaltır ama yinede tüm değişkenleri kullanır, Lasso ise değişken kat sayısını tamamen sıfır yaptığı için sadece belli değişkenleri seçmiş olur.
 - ❖ Ridge Regresyonu (Önemsiz değişkenlerin kat sayılarını küçült.)
 - ❖ Lasso Regresyonu (Önemsiz değişkenlerin kat sayılarını 0 yap.)

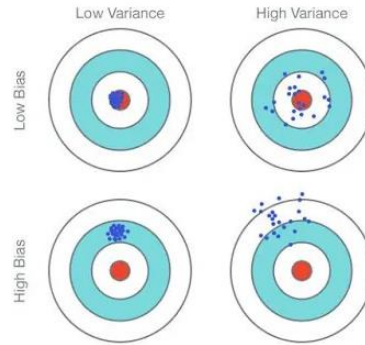
Aşırı öğrenme problemini nasıl çözebiliriz?

- **Seyreltme (Dropout):** Sinir ağı içerisinde yer alan bazı nöronların rastgele olarak ortadan kaldırılmasında seyreltme (*dropout*) katmanı kullanılmaktadır.
- Sinir ağında nöronların yüzde kaçı ortadan kaldırılacağı kullanıcı tarafından belirlenmektedir (0 ile 1 arasında bir değer ile belirlenir). Böylece ağın *overfitting* olması önlenerek performansın artması sağlanmaktadır.
 - Basitliği ve etkili olması nedeniyle, günümüzde çeşitli mimarilerde, genellikle *fully connected* katmanından sonra *dropout* katmanı kullanılmaktadır.



Varyans-Bias Çelişkisi

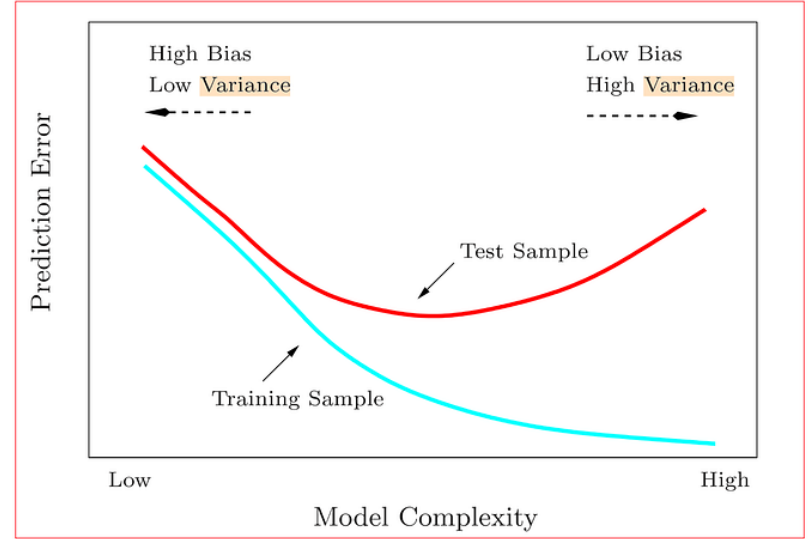
- **Varyans**, model eğitim veri setinde iyi performans gösterdiğinde, ancak bir test veri kümesi veya doğrulama veri kümesi gibi, eğitilmemiş bir veri kümesinde iyi performans göstermediğinde ortaya çıkar.
- Varyans, gerçek değerden tahmin edilen değerlerin ne kadar dağınık olduğunu söyler.
- **Bias**, gerçek değerlerden tahmin edilen değerlerin ne kadar uzak olduğudur. Tahmin edilen değerler gerçek değerlerden uzaksa, bias yüksektir.



Source: <https://towardsdatascience.com/regularization-the-path-to-bias-variance-trade-off-b7a7088b4577>

Varyans-Bias Çelişkisi

- Yüksek bias'a sahip bir modelin çok basit olduğunu söyleyebiliriz.
- Yüksek varyansa sahip bir model, veri noktalarının çoğuna uymaya çalışır ve bu da modeli karmaşık yapar ve modellenmesini zorlaştırır.
- Overfitting problemi olan modellerde yüksek varyans, düşük bias durumu görülmektedir.
- Underfitting sorunu olan modeller düşük varyans ve yüksek bias'a sahiptir.
- Grafikte görüldüğü gibi model karmaşıklığı arttıkça eğitim seti üzerinde hatalı tahmin oranı azaltmakta ancak test veri seti üzerinde tahmin hatası artmaktadır.



Source: Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman

Varyans-Bias Çelişkisi

- **Yüksek Bias Düşük Varyans:** Modeller tutarlıdır, ancak ortalama hata oranı yüksektir.
- **Yüksek Bias Yüksek Varyans:** Modeller hem hatalı hem de tutarsızdır .
- **Düşük Bias Düşük Varyans:** Modeller ortalama olarak doğru ve tutarlıdır. Modellerimizde bu sonucu elde etmek için çabalamaktayız.
- **Düşük Bias Yüksek Varyans:** Modeller bir dereceye kadar doğrudur ancak ortalamada tutarsızdır. Veri setinde ufak bir değişiklik yapıldığında büyük hata oranına neden olmaktadır.

Yüksek bias veya yüksek varyansa sahip olduğumuzu bulmanın yolu nedir?

- Eğer **model yüksek bias'a sahipse** aşağıdaki sonuçlarla karşılaşmamız kaçınılmazdır;
 - Modelin eğitim setinin hata oranı yüksektir.
 - Test / doğrulama veri seti hata oranı eğitim seti ile benzer oranda yüksektir.
- Eğer **model yüksek varyans'a sahipse** aşağıdaki sonuçlarla karşılaşmamız kaçınılmazdır;
 - Modelin eğitim setinin hata oranı düşüktür.
 - Modelin test/doğrulama veri setinin hata oranı yüksektir.
- **Yüksek bias problemini çözmek için** aşağıdaki yöntemleri uygulayabiliriz:
 - **Daha fazla veri eklemek** : Daha fazla veri ekleyerek veri çeşitliliğini arttırmak gereklidir.
 - **Daha fazla değişken eklemek** : Model karmaşıklığının artmasını sağlamaktadır.
 - **Regularization (düzenleme)** : Değişkenlerin ağırlığını arttırmak için regularization değerini azaltın.
- **Yüksek varyans problemini çözmek için** overfitting sorununu çözme yöntemlerini uygulamamız gereklidir.

Kaynaklar

- Prof. Dr. Devrim Akgün, Deep Learning ders notları
- <https://www.veribilimiokulu.com/overfitting/>
- <https://medium.com/@gulcanogundur/overfitting-a%C5%9F%C4%B1r%C4%B1-%C3%B6%C4%9Frenme-underfitting-eksik-%C3%B6%C4%9Frenme-ve-bias-variance-%C3%A7eli%C5%9Fkisi-b92bef2f770d>
- <https://miuul.com/blog/as%C4%B1r%C4%B1-ogrenme-problemleri-nedir-ve-nas%C4%B1l-%C3%A7ozulur>
- <https://medium.com/kodluyoruz/makine-%C3%B6%C4%9Frenmesinde-overfitting-underfitting-best-fitting-kavramlar%C4%B1-9f80a5d42719>