Introduction to Statistical Learning with Applications in R

Chapter 6 – Linear Model Selection and Regularization – 6.4 Considerations in High Dimensions

Elif Kurt

Yüksek Boyutlarda Dikkat Edilmesi Gerekenler

Giriş

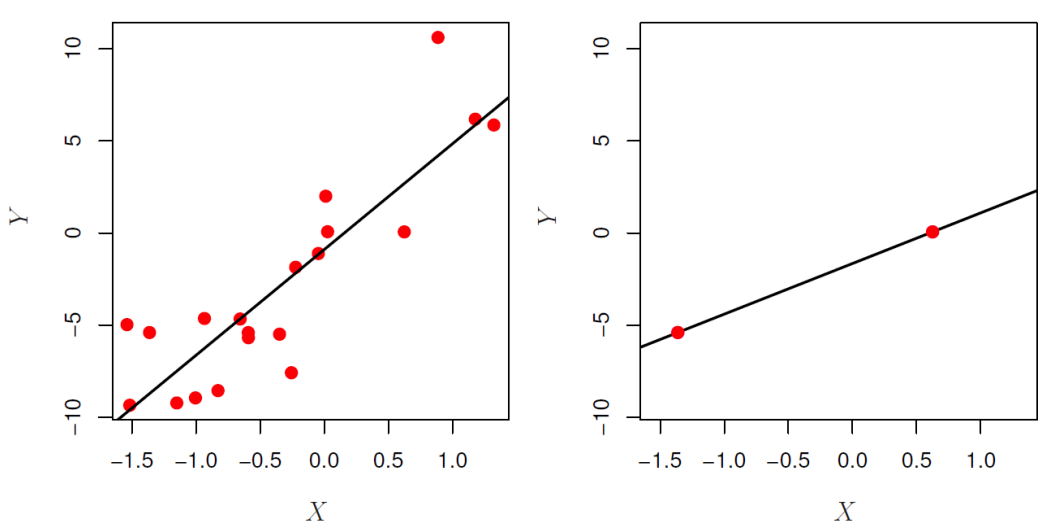
- Elimizdeki veri setlerinde genelde gözlem sayısı n, tahmin edici sayısı p'den daha büyüktür n>>p . Örneğin bir hastanın kan basıncını tahmin etmek için verilen feature'lar *yaşı, cinsiyeti ve vücut kitle indeksi* şeklinde olur.

- Fakat artık günümüzde veri toplamadaki kolaylıktan dolayı çok daha fazla feature'a/tahmin ediciye ulaşabilir haldeyiz. Öyle ki artık p sayısı n'den daha büyük olabilmektedir. p>>n

Bu şekilde gözlemlerden daha fazla feature içeren data setleri **yüksek boyutlu (high-dimensional)** olarak isimlendirilir.

- Bu tarz yüksek boyutlu veri setlerinde least squares lineer regresyonu gibi klasik yaklaşımlar uygun değildir. Normalde de n>p olsa dahi least squares uyguladığımızda bias-varyans trade offu ve train setinde overfitting olması tehlikesi vardı. Burada bu konu daha fazla önem taşır hale geliyor.
- Burada p>n olduğunda veri setini yüksek boyutlu olarak tanımladık ama bahsedeceğimiz mevzular p, n'den biraz küçük olsa dahi geçerli.

Yüksek Boyutlardaki Problem Ne?



Buradaki grafikte 1 feature'ımız varken; sol taraftaki panel 20 gözlem olduğundaki durumu, sağ taraftaki panel 2 gözlem olduğundaki durumu ifade ediyor.

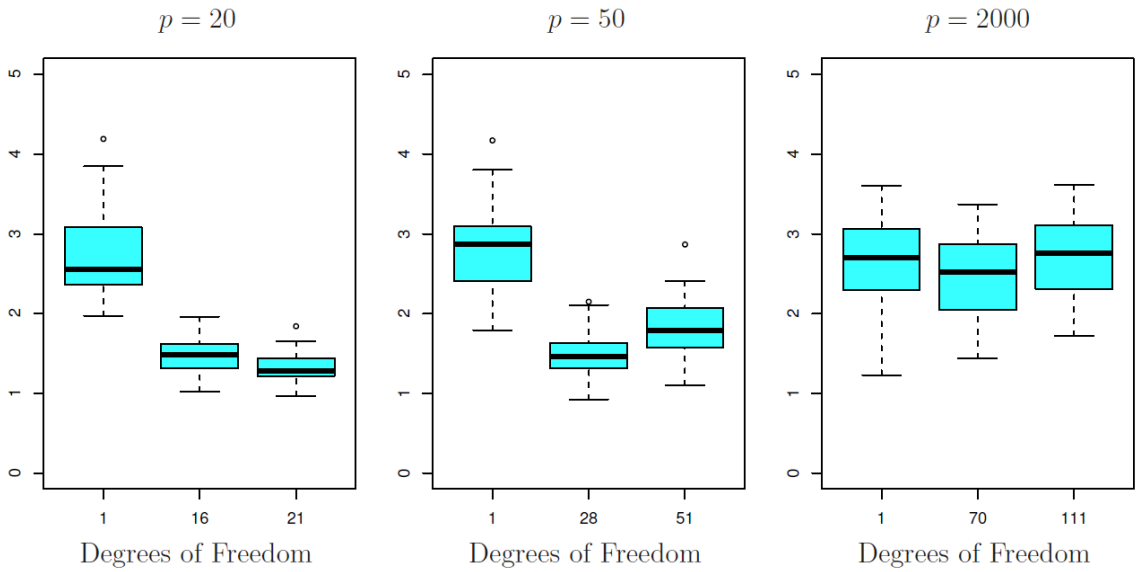
- 20 gözlem olduğunda, (n>p) ve least squares verilere tam olarak fit olmaz; bunun yerine regresyon çizgisi mümkün olduğu kadar 20 gözleme yaklaşmaya çalışır.

- Öte yandan, sadece 2 gözlem olduğunda (yani p, n'den çok az küçük olduğunda) bu gözlemlerden bağımsız olarak regresyon çizgisi verilere tam uyacaktır. Bu problemlidir; çünkü mükemmel uyum verilerin overfittingine yol açacaktır. Bu yüzden bu modeli farklı bir test setinde denediğimizde düşük performans gösterecektir.

Dolayısıyla **p>n veya p≈n olduğunda veya p, n'den çok az küçük olduğunda** least squares regresyon eğrisi çok esnektir ve bu nedenle overfitting olur.

Tahmin Edici Sayısının Artmasının Modelin Performansına Etkisi

Yüksek boyutlu data setlerinde MSE, R^2, Cp, AIC ve BIC yaklaşımlarını kullanamıyoruz. Bunun yerine verileri fit etmek için forward stepwise selection, ridge regression, lasso regression, principal components regression gibi yöntemleri kullanıyoruz. Bu yaklaşımlar, least squares'ten daha az esnek bir fit etme yaklaşımı kullanarak overfittingi önlüyor.



Bu grafikte 100 training gözleminde 20'si gerçekten yanıtla ilişkili olan bir modelin lasso gerçekleştirildiğinde tahmin edici sayısı arttıkça performansının nasıl değiştiğini gözlemliyoruz. Degrees of freedom (serbestlik derecesi) burada lasso'nun sıfır olmayan katsayılarını ifade ediyor.

p=20 olduğunda yüksek serbestlik derecesi olduğunda en düşük test MSE elde ediliyor. p=50 olarak arttığında daha fazla katsayıya ihtiyaç duyuyoruz test MSE'sini küçültümek için. p=2000 olduğunda 2.000 özelliğin yalnızca 20'sinin gerçekten sonuçla ilişkili olması nedeniyle, katsayı miktarına bakılmaksızın lasso kötü performans gösteriyor.

Burada feature sayısı arttıkça test seti hatasının arttığını görüyoruz.

Bu, **curse of dimensionality (boyut laneti)** olarak bilinen yüksek boyutlu verilerin analizinde anahtar bir ilkedir. Bir modele fit edilmesi için kullanılan featureların sayısı artıkça, fit edilen modelin kalitesinin de artacağı düşünülebilir. Ancak yanıtla ilişkili olmayan gürültü özelliklerinin eklenmesi, fit edilen modelde bir bozulmaya ve sonuç olarak artan bir test hatasına yol açacaktır.

Yüksek Boyutlarda Sonuçları Değerlendirmek

3. chapter'da 3 veya daha fazla değişken arasında korelasyon olduğunda bunu multicollinearity olarak tanımlamıştık. Yüksek boyutlu durumlarda regresyon uyguladığımızda bu, çok daha fazla dikkate edilmesi gereken bir problem haline geliyor. Çünkü modeldeki herhangi bir değişken, modeldeki diğer tüm değişkenlerin lineer bir kombinasyonu olarak yazılabilir.

50 tahmin edicili bir veri setimiz var diyelim. 17'si yanıtı tahmin etmek için iyi sonuç veriyor. Farklı bir veri setinde farklı 17 değişken daha iyi sonuç verebilir. Bu yüzden çalışan modelimizin birçok olası modelden sadece biri olduğunu bilmeli ve bunu farklı veri setlerinde doğrulamaya dikkat etmeliyiz.