Giriş

(İktisatçılar İçin) Makine Öğrenmesi (TEK-ES 2020)

Hüseyin Taştan
Yıldız Teknik Üniversitesi

Plan

- Dersin tanıtımı
- Öğrenme ve makine öğrenmesi
- Yapay Zeka, Veri Bilimi, Veri Madenciliği
- Gözetimli ve gözetimsiz öğrenme, pekiştirmeli öğrenme, derin öğrenme
- Ekonometri ve makine öğrenmesi: farklar ve benzerlikler
- Kestirim politikası problemleri

Dersin tanıtımı

- Türkiye Ekonomi Kurumu 2020 online yaz seminerlerinin bir parçası olarak tasarlanan bu derste temel makine öğrenmesi algoritmalarının giriş düzeyinde incelenmesi amaçlanmaktadır. Ders ağırlıklı olarak gözetimli öğrenme algoritmalarını kapsamaktadır.
- Ders kitabı: James, Gareth, D. Witten, T. Hastie, R. Tibshirani (2017), An Introduction to Statistical Learning with Applications in R, Springer. Kitabın elektronik versiyonuna aşağıdaki linklerden ulaşabilirsiniz: http://faculty.marshall.usc.edu/gareth-james/ISL/
- Ders malzemeleri için link: https://github.com/htastan/Makine-Ogrenmesi
- Uygulamalar R programında yapılacaktır.

Günlük ders programı

1. gün: Giriş, Temel Kavrammlar, R'a Giriş, R ile Tidy veri analizi

2. gün: Gözetimli öğrenme, Regresyon problemleri, R uygulamaları

3. gün: Sınıflandırma problemleri, R uygulamaları

4. gün: Geçerleme (Validation) ve Çapraz geçerleme (Cross validation) yaklaşımı; Düzenlileştirme (Regularization)

5. gün: Ağaç bazlı yöntemler (regresyon ve sınıflandırma ağaçları, bagging, rassal ormanlar, boosting)

6. gün: Gözetimsiz öğrenme yöntemleri: Boyut küçültme (temel bileşenler analizi (PCA)), kümeleme (K-ortalamalar ve hiyerarşik kümeleme)

Öğrenme Nedir?

Öğrenme (learning), öğrenenin ya da öğretenin bakış açısından çeşitli biçimlerde tanımlanabilir. Örneğin:

- ‘‘Çalışarak, tecrübe ederek, ya da düşünerek bilgi (knowledge) sahibi olmak’’
- ‘‘Gözlem yaparak ya da haber alarak (information) farkında olmak’’
- ‘‘Hafızaya almak’’
- ‘‘Bilgilendirilmek’’
- ‘‘Komut almak’’

Öğrenmenin gerçekleştiğini nasıl anlarız? Klasik eğitim sisteminde öğretmen öğrencileri düzenli olarak çeşitli araçlarla sınayabilir.

Bilgisayarın (makinenin) öğrendiğini nasıl anlarız?

- “Hafızaya almak”, “Bilgilendirilmek” ve “Komut almak” bilgisayarlar için kolaydır. Ancak bunlar tam olarak öğrenme sayılmazlar.
- Öğrenmeyi bilgi (knowledge) ile değil de belirli bir göreve ait performans ile ilişkilendirdiğimizde öğrenmenin gerçekleşip gerçekleşmediğini anlayabiliriz.
- Yani, bilgisayarlar bir görevi yerine getirirken geçmişteki davranışlarıyla kıyaslandığında performanslarında iyileşme gösteriyorlarsa «öğrenmiş» sayılırlar.
- Training: içinde (öğrenen açısından) düşünme barındırmayan öğrenme süreci. Örneğin, bir evcil hayvanın tuvalet eğitimi, vs.
- Bunu bir öğrencinin düşünerek içinde yer aldığı öğrenme süreci ile kıyaslayınız.

Makine Öğrenmesi Nedir?

> “Makine öğrenmesi bilgisayarlara açıkça programlama olmadan öğrenme yeteneği kazandıran bir çalışma alanıdır.” - Arthur Samuel

- Yapay zeka alanının öncülerinden olan Arthur Samuel, makine öğrenmesi kavramının popülerleşmesine önemli katkılar yapmıştır.
- Geliştirdiği dama programında bilgisayarlar oyunu kendilerine karşı çok sayıda oynayarak hangi hamlelerin daha başarılı olduğunu öğrenebilmektedir. Arthur Samuel (1959).

Öğrenme Problemi

- Tom Mitchell Machine Learning başlıklı ders kitabında makine öğrenmesi kavramını *iyi tanımlı öğrenme problemi* çerçevesinde incelemektedir:

> A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
>
> "Dama oynamayı öğrenen bir bilgisayar, bu oyunu defalarca oynayarak elde ettiği tecrübeyle kazanma yüzdesi ile ifade edilen performansını geliştirebilir" (Mitchell, 1997, p. 2).

- Öyleyse bir makine bir görevi yerine getirirken deneyimlerden (verilerden) hareketle performansını geliştirerek "öğrenir" diyebiliriz.

Dama oyunu için öğrenme problemi

> “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E”.

- Görev (**T**): dama oynama
- Eğitim ya da tecrübe (**E**): bilgisayarın kendi kendine çok sayıda dama oynaması
- Performans ölçütü (**P**): rakiplere karşı kazanılan oyun yüzdesi.

Öğrenme problemi: Eposta örneği

Şöyle bir Eposta yazılımı düşünelim: bu yazılım sizin kararlarınızı takip ediyor ve hangi Epostalara cevap yazdığınızı hangilerini çöp kutusuna düzenli olarak attığınızı ya da gereksiz (spam) olarak etiketlediğinizi kaydediyor olsun. Belirli bir süreden sonra bu yazılım Epostalarınızı spam ya da önemli olarak sınıflandırma becerisine sahip olabilir.

Bu problemde:

- **T**ask: Epostaların spam ya da önemli olarak sınıflandırılması
- **E**xperience: Kullanıcının hangi postaları spam olarak sınıflandırdığının gözlenmesi
- **P**erformance: Doğru bir şekilde spam olarak sınıflandırılan Epostaların oranı

Makine Öğrenmesi Nedir?

> "Yapay öğrenme, bilgisayarların örnek veri ya da geçmiş deneyimi kullanarak başarımlarını artıracak biçimde programlanmasıdır". Alpaydın (2018, s.3)

- Yapay öğrenme ya da makine öğrenmesi, betimleyici veya kestirimsel (predictive) modellerin geçmiş deneyim ve verilerle birleştirilerek tahmin edilmesinde istatistik teorisini kullanır.
- Örnek (Alpaydın, 2018): Bir süpermarket zinciri için sepet çözümlemesi: Eğer X malını alan müşteriler sıklıkla Y malını da alıyorsa ve bir müşteri X malını almışsa, o Y malını almaya aday bir müşteridir (çapraz satış)
- Müşterinin geçmişte satın aldığı ürün ya da ürünleri içeren X kümesine koşullu olarak Y malı için $Pr(Y|X)$ olasılığını öğrenmeyi amaçlıyoruz. Böylece o müşteri grubuna yönelik pazarlama stratejileri geliştirilebilir.

Makine Öğrenmesi Nedir?

> "...Makine öğrenmesi, temel amacı verilerden hareketle kestirim (regresyon), sınıflandırma, ve kümeleme veya gruplama problemlerinin çözümüne yönelik algoritmaların geliştirilmesidir" Susan Athey (2018).

- Athey'nin tanımı makine öğrenmesi çalışma alanına yoğunlaşmaktadır. Ancak diğer tanımlarla ortak odak noktası kestirim (prediction)'dir.

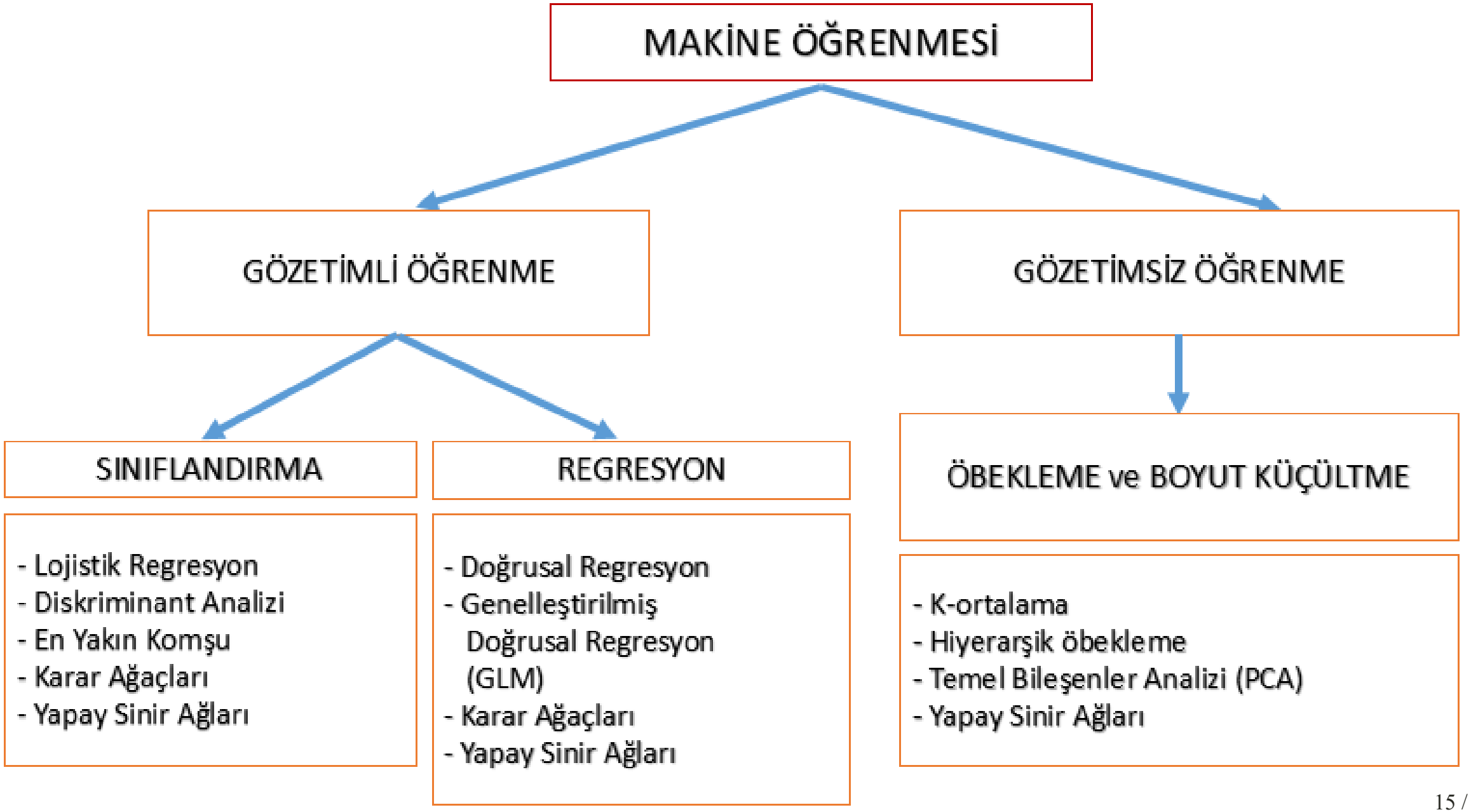
İstatistiksel Öğrenme

> "İstatistiksel öğrenme (statistical learning), karmaşık veri kümelerinin modellenmesi ve anlaşılması için bir dizi araç geliştirmekle uğraşır. İstatistik bilimi içinde yakın zamanda geliştirilen İstatistiksel Öğrenme alanı, bilgisayar bilimi ve özellikle makine öğrenimindeki paralel gelişmelerle yakından ilişkilidir." James, Witten, Hastie ve Tibshirani (2017).

- Bu tanım S. Athey'nin tanımına yakındır.
- İstatistiksel Öğrenme (SL) = Makine Öğrenmesi (ML)?
- Her ikisinin de odak noktası verilerden «öğrenme»
- Verilerin rassallığı (randomness) dikkate alındığında makine öğrenmesi algoritmalarının «istatistiksel» olması kaçınılmazdır.

Makine Öğrenmesi ile ilişkili bazı moda terimler

- **Yapay Zeka** (Artificial Intelligence): bilgisayarların (makinelerin) insanlardan bağımsız olarak akıllıca davranması için araçlar geliştiren disiplin.
  - Günümüzde insan gibi davranan genel yapay zekadan çok, belirli bir görevi başarıyla yapabilen daha dar tanımlı yapay zeka uygulamaları yaygınlaşmaya başlamıştır.
  - Uygulama örnekleri: otonom araçlar, Siri ve benzeri sesli yardımcılar, Google Search, Hastalıkların teşhis edilmesi, vs.
- **Veri Bilimi** (Data Science): verilerde saklı bilgiyi ortaya çıkarmak için yöntem ve algoritmalar geliştiren; istatistik, bilgisayar bilimi, matematik ve ilgili diğer bilim dallarının kesişiminde disiplinlerarası bir çalışma alanı.
- **Veri Madenciliği** (Data Mining): özellikle büyük verilerdeki daha önce bilinmeyen faydalı örüntü ve kalıpların ortaya çıkarılması.


MAKİNE ÖĞRENMESİ
GÖZETİMLİ ÖĞRENME
GÖZETİMSİZ ÖĞRENME
SINIFLANDIRMA
REGRESYON
ÖBEKLEME ve BOYUT KÜÇÜLTME
- Lojistik Regresyon
- Diskriminant Analizi
- En Yakın Komşu
- Karar Ağaçları
- Yapay Sinir Ağları
- Doğrusal Regresyon
- Genelleştirilmiş Doğrusal Regresyon (GLM)
- Karar Ağaçları
- Yapay Sinir Ağları
- K-ortalama
- Hiyerarşik öbekleme
- Temel Bileşenler Analizi (PCA)
- Yapay Sinir Ağları

Gözetimli Öğrenme

- Makine öğrenmesi algoritmaları ikiye ayrılabilir: gözetimli ya da güdümlü (supervised) ve gözetimsiz (unsupervised) öğrenme.
- Gözetimli öğrenmede girdi değişkenleri (özellikler ya da öznitelikler) ile çıktı değişkeni gözlemlenebilir. Her gözleme ait bir çıktı değeri ya da kategorisi (etiketi) mevcuttur. Amaç çıktıyı kestirmekte en başarılı modeli bulmaktır.
- Örnek: bir kredi başvurusundan hareketle ödeyememe riskini (olasılığını) öngörmek
    - Bir banka geçmişteki kredi başvurularından hareketle bir ödeyememe modeli kurabilir. Krediye başvuran bireyin özellikleri (features) kurulan modelde değerlendirilerek bir kestirim yapılabilir (Örneklem-dışı kestirim-out-of-sample prediction).
- Gözetimli öğrenme türleri: regresyon problemleri, sınıflandırma problemleri.

Gözetimsiz Öğrenme

- Girdi değişkenleri (Kestiriciler ya da öznitelikler) gözlemlense de bir çıktı değişkeni ya da etiket yoktur.
- Yaygın kullanılan problemler: kümeleme ve boyut küçültme
- Kümeleme (clustering): bir özellik setinden hareketle homojen gruplar bulunabilir mi? Örneğin benzer özelliklere sahip tüketici grupları, hasta türleri, benzer davranışa sahip seçmen grupları vb.
- Boyut küçültme (dimensionality reduction): çok sayıda potansiyel kestirim değişkeni arasından en önemlilerinin seçilmesi

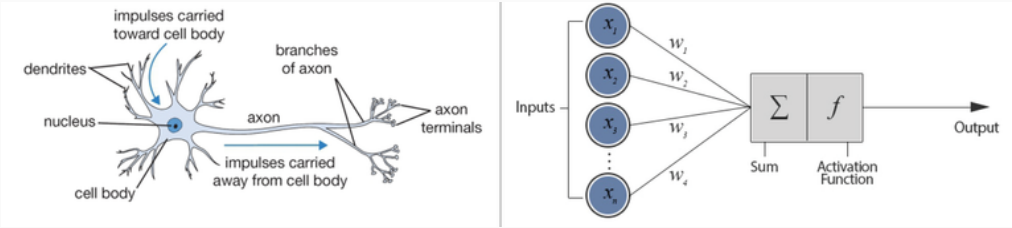
Pekiştirmeli Öğrenme (Reinforcement Learning)

- Bazı durumlarda gözetimli öğrenme pratikte mümkün değildir. Çıktı sadece bir eylem değil, eylemler dizisidir. Örneğin, satranç ve benzeri oyunlar oynamak.
    - Tekil eylemin tek başına önemi yoktur. Önemli olan hedefe ulaşmak için doğru eylem dizisini (politika) uygulamaktır.
- Yapay öğrenme algoritması politikaların başarısını karşılaştırmalı ve geçmiş deneyimlerden hareketle en iyi politikayı önermelidir.
    - Örnek: Satranç oyununda (ve benzeri diğer oyunlarda) tek bir hamlenin çok önemi yoktur. Oyunu kazandıran doğru hamleler dizisi önemlidir.
    - Örnek: bir labirentte yolunu bulmaya çalışan bir robotu düşünelim. Robot çevreyi algılayarak çıkışı bulmak için çeşitli eylemlerde bulunur. Geri besleme olmadığından çıkışı bulup ödül alması ancak bir dizi eylemlerin tamamlanmasına bağlıdır.

Detaylar için bkz. Alpaydın (2018, s. 10 ve s. 383)

Derin Öğrenme (Deep Learning)

- Derin öğrenme beynin işleyiş yapısından hareketle geliştirilmiş yapay sinir ağları (artificial neural network - ANN) ve benzeri algoritmaları kapsayan bir makine öğrenmesi çalışma alanıdır.
- Yapay sinir ağları tıpkı beyinde olduğu gibi nöronlardan ve bunların bağlantılarından oluşur.
- Tipik bir yapay sinir ağı problemi giriş katmanı, ara katmanlar, ve çıkış katmanından oluşur. Kaç katman olacağı ve bu katmanların birbirine nasıl bağlanacağı gibi çok sayıda kararın verilmesi gerekir.
- Kullanım alanları: görüntü tanıma, doğal dil işleme, konuşma tanıma, vb. büyük veri kümelerinin olduğu problemler.



Ekonometri ve Makine Öğrenmesi

- Tipik bir (gözetimli) makine öğrenmesi problemini aşağıdaki gibi yazabiliriz:

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon$$

  Burada y çıktı değişkenini (etiketleri), $\{x_1, x_2, \ldots, x_p\}$ ise özellikleri ifade etmektedir. ϵ rassal hata terimidir.

- Bilinmeyen fonksiyon kalıbı $f(\cdot)$ ile gösterilmiştir.

- Modelin kestirimini şöyle yazalım:

$$\hat{y} = f(x_1, x_2, \ldots, x_p)$$

- Makine öğrenmesi probleminde amaç kestirim hatasını, $y - \hat{y}$, en küçük yapmaktır. Bu çerçeve Ekonometride kullandığımız yaklaşıma çok benzemektedir.

Ekonometri ve Makine Öğrenmesi

- $f(\cdot)$'in (parametrelerde) doğrusal olduğunu varsayarsak:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

  ekonometri uygulamalarında yaygın olarak kullanılan doğrusal regresyon problemini elde ederiz.

- Modeli matris notasyonuyla daha kompak bir biçimde yazabiliriz.

$$\underbrace{\mathbf{y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times(p+1)} \underbrace{\beta}_{(p+1)\times 1} + \underbrace{\mathbf{u}}_{n\times 1}$$

Modelde sabit terim dahil olmak üzere $p+1$ bilinmeyen β parametresi vardır.

Ekonometri ve Makine Öğrenmesi

- **Ekonometri**nin odak noktası modelin bilinmeyen parametre vektörünün, β, sapmasız (mümkün değilse, tutarlı) ve etkin tahminidir.
- İktisat bakımından anlamlı yorumlama yapabilme ön plandadır.
- Modelin tahmini: Sıradan En Küçük Kareler (Ordinary Least Squares - OLS) Gauss-Markov varsayımları altında BLUE özelliklerini sağlar.
- Pratikte Çok üzerinde durulmayan bir nokta: gözlem sayısı, n, $\mathbf{X}$ değişken sayısından (p) çok daha büyük olmalıdır: $n >> p$
- OLS yöntemi $p > n$ durumunda çalışmaz.
- $p = n$ olursa çalışır mı?

Ekonometri ve Makine Öğrenmesi

- Makine Öğrenmesi (ML) ve Ekonometri farkı:

> Makine öğrenmesi çıktı değişkeninin kestirimine, yani $\hat{y}$'ya yoğunlaşır. Ekonometrinin odağında ise $\hat{\beta}_j$, $j = 1, 2, \ldots, p$, vardır. (Mullainathan ve Spiess, 2017)

- Ekonometri: x_j'nin y üzerindeki nedensel (causal) etkisi ne kadardır? (ölçüm problemi).
- Ekonometri: nedensel etkilerin sapmasız/tutarlı ve etkin (en düşük varyanslı) tahmini ön planda. İktisat teorisi yol gösterici.
- **ML**: y'nin kestiriminde en başarılı modellerin verilerden öğrenilmesi (yani tahmin edilmesi) ön planda
- Ekonometri ve ML arasındaki sınırlar çok keskin değil. ML algoritmaları ekonometrik modellemede faydalı olabileceği gibi iktisat teorisi de ML uygulamalarında kullanılabilir. (ekonometri ve makine öğrenmesinin kapsamlı karşılaştırması için bkz. Athey ve Imbens (2019)).

Kestirim Politikası Problemleri

- Ekonometrinin odağında dikkatlice yapılmış nedensel analizler yer alır.
- Çoğu durumda uygulanan bir politikanın bir çıktı değişkenini nasıl ve ne kadar etkilediğini tahmin etmeye çalışırız. Bu analizlerde «eğer politika uygulanmasaydı ne olurdu?» sorusu da önemli bir yere sahiptir (karşıolgusal analiz - counterfactual).
- Ancak bazı politika problemlerinde nedensel çıkarımlar gerekmeyebilir (Kleinberg et al. 2015).
- Bu politika problemleri dikkatlice hazırlanmış yapısal belirlenme şartlarını ve genellikle ciddi varsayımlara yaslanan karşıolgusal analizler yerine sadece başarılı kestirim modellerini gerektirir.
- Başarılı kestirim modellerinin oluşturulmasında makine öğrenmesi yaklaşımları yararlı olabilir.

Kestirim Politikası Problemleri: Örnekler

(Kleinberg et al. 2015)

- **Sağlık**: operasyon öncesinde mevcut verilerden hareketle hangi ameliyatların gereksiz olacağının öngörülmesi.
- **Ceza hukuku**: tutuklunun serbest bırakıldığında bir suç işleme olasılığının kestirimi
- **Eğitim**: hangi öğretmenin en yüksek katma değere sahip olacağının tahmin edilmesi
- **İşgücü piyasası**: işsiz kalınan sürenin uzunluğunu tahmin ederek çalışanların tasarruf oranları ve iş arama stratejileri konularında karar vermelerine yardımcı olmak.
- **Sosyal politikalar**: yüksek risk grubundaki gençlerin tahmin edilerek önleyici müdahalelerin geliştirilmesi
- **Finans**: potansiyel borçluların kredi değerliliğinin belirlenmesi

Kaynakça

Alpaydın, Ethem (2018), *Yapay Öğrenme*, 4. Baskı (Ethem Alpaydın, *Introduction to Machine Learning*, 2. baskıdan çeviri), Boğaziçi Üniversitesi Yayınevi, İstanbul.

Athey, S. (2018), "The Impact of Machine Learning on Economics", Stanford University, unpublished paper. (https://projects.iq.harvard.edu/files/pegroup/files/athey2018.pdf), basılmış versiyon: (https://www.nber.org/chapters/c14009).

Athey, S. ve Imbens, G.W. (2019), "Machine Learning Methods That Economists Should Know About", *Annual Review of Economics*, 11: 685-725.

James, Gareth, Witten D., Hastie T. ve Tibshirani R, (2017), *Introduction to Statistical Learning*, corrected 8th printing, Springer, New York.

Kleinberg, J., Ludwig, J., Mullainathan, J., ve Obermeyer, Z. (2015), "Prediction Policy Problems", *American Economic Review, Papers and Proceedings*, 105(5): 491-495. (http://dx.doi.org/10.1257/aer.p20151023)

Mullainathan, S. ve Spiess, J. (2017) "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31(2), 87-106.

Samuel, A. L. (1959), "Some studies in machine learning using the game of checkers", *IBM Journal*, 3: 210–229.

Varian, H.R. (2014), "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, 28 (2): 3–28.