En Küçük Kareler Veri Uydurma

T.C. Trakya Üniversitesi
Mühendislik Fakültesi
Elektrik - Elektronik Mühendisliği Bölümü
Kontrol Anabilim Dalı

Dr. Öğr. Üyesi Işık İlber Sırmatel
sirmatel.github.io

Kaynak (source)

Lecture Slides for Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares. Stephen Boyd, Lieven Vandenberghe

Konu listesi

1. En küçük kareler model uydurma
2. Geçerleme
3. Öznitelik mühendisliği

Bölüm 1

En küçük kareler model uydurma

Problem formülasyonu

- skaler y ile n-vektör x'in bağıntılı olduğunu düşünüyoruz

$$y \approx f(x)$$

- x'e bağımsız değişken (*independent variable*) denir
- y'ye amaç değişken (*outcome variable* veya *response variable*) denir
- $f : \mathbb{R}^n \to \mathbb{R}$, x ile y arasındaki bağıntıyı verir
- genellikle, x bir öznitelik (*feature*) vektörüdür, y ise öngörmek (*predict*) istediğimiz bir şey
- x ile y arasındaki doğru ilişkiyi veren f'i bilmiyoruz

Veriler

- elimizde bazı veriler (*data*) bulunuyor:

$$x^{(1)}, x^{(2)}, \ldots, x^{(N)} \qquad y^{(1)}, y^{(2)}, \ldots, y^{(N)}$$

  bunlara gözlemler (*observations*), örnekler (*examples*), örneklemler (*samples*), veya ölçümler (*measurements*) de denir
- $x^{(i)}$, $y^{(i)}$, i. veri çiftidir
- $x_j^{(i)}$, i. veri noktası $x^{(i)}$'nin j. elemanıdır
- N: veri kümesinin (*data set*) büyüklüğü (veri noktası sayısı)

Model

- $f : \mathbb{R}^n \to \mathbb{R}$: x ile y arasındaki doğru ilişki
- f'in ne olduğunu bilmiyoruz
- f'in bir yaklaşıklığı (*approximation*) olarak model $\hat{f} : \mathbb{R}^n \to \mathbb{R}$'i seçelim
- parametrelere göre doğrusal (*linear in the parameters*) model formu:

$$\hat{f}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_p f_p(x)$$

- $f_i : \mathbb{R}^n \to \mathbb{R}$ bizim seçtiğimiz taban fonksiyonlarıdır (*basis function*)
- θ_i bizim seçtiğimiz model parametreleridir
- $\hat{y}^{(i)} = \hat{f}(x^{(i)})$ modelin $y^{(i)}$'ye dair öngörüsüdür
- $\hat{y}^{(i)} \approx y^{(i)}$ olsun isteriz (modelin gözlemlenen verilerle tutarlı (*consistent*) olmasını isteriz)

En küçük kareler veri uydurma

- öngörü hatası veya kalıntı: $r^{(i)} = y^{(i)} - \hat{y}^{(i)}$
- en küçük kareler veri uydurma (*data fitting*) problemi: öngörü hatasının RMS değerini minimize edecek şekilde model parametrelerini (θ_i) seçmek
- amaç fonksiyonu (öngörü hatasının RMS değeri)

$$\sqrt{\frac{(r^{(1)})^2 + (r^{(2)})^2 + \cdots + (r^{(N)})^2}{N}}$$

- bu problem bir en küçük kareler problemi olarak formüle edilebilir ve çözülebilir

En küçük kareler veri uydurma

- $y^{(i)}$, $\hat{y}^{(i)}$ ve $r^{(i)}$'yi N-vektörler olarak yazalım

$$y^{\mathsf{d}} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix} \qquad \hat{y}^{\mathsf{d}} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(N)} \end{bmatrix} \qquad r^{\mathsf{d}} = \begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(N)} \end{bmatrix}$$

- öngörü hatasının RMS değerini $\mathrm{rms}(r^d)$ ile gösterelim
- elemanları $A_{ij} = f_j(x^{(i)})$ olan bir $N \times p$-matris A tanımlayalım. buradan $\hat{y}^d = A\theta$ yazabiliriz

En küçük kareler veri uydurma

- en küçük kareler veri uydurma problemi:

$$\|r^d\|^2 = \|y^d - \hat{y}^d\|^2 = \|y^d - A\theta\|^2 = \|A\theta - y^d\|^2$$

  ifadesini minimize edecek θ'yı seçmek
- çözüm: $\hat{\theta} = (A^T A)^{-1} A^T y$ (A'nın sütunları doğrusal bağımsız ise)
- minimum ortalama karesel hata (*minimum mean square error*, MMSE): $\frac{\|A\hat{\theta} - y\|^2}{N}$

Sabit model uydurma

- olası en basit model: $p = 1$, $f_1(x) = 1$
- model formu: $\hat{f}(x) = \theta_1$ (sabit bir sayı)
- $A = \mathbf{1}$, dolayısıyla

$$\hat{\theta}_1 = (\mathbf{1}^T\mathbf{1})^{-1} = (1/N)\mathbf{1}^T y^d = \mathrm{avg}(y^d)$$

- sonuç olarak: y^d'nin ortalaması sabit bir sayı şeklindeki model için en küçük kareler uydurmasıdır
- MMSE $\mathrm{std}(y_d)^2$, RMS hata $\mathrm{std}(y^d)$
- daha gelişmiş modeller, başarımları sabit modelle karşılaştırılarak sınanabilir

Tek değişkenli fonksiyon uydurma

- tek değişkenli fonksiyon $f : \mathbb{R} \to \mathbb{R}$'nin yaklaşıklığını bulmak istiyoruz
- verileri $((x_i, y_i))$ ve model $\hat{y} = \hat{f}(x)$'i çizdirebiliriz

Düz çizgi uydurma

- $p = 2,\ f_1(x) = 1,\ f_2(x) = x$
- model formu: $\hat{f}(x) = \theta_1 + \theta_2 x$ (düz çizgi)
- A matrisinin formu:

$$A = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(N)} \end{bmatrix}$$

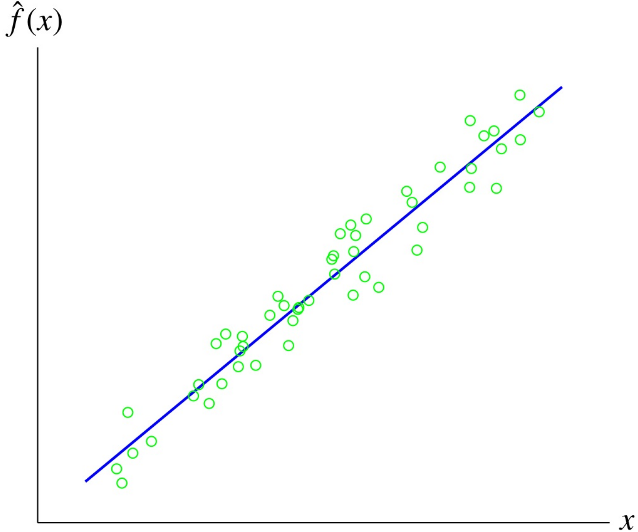
- θ_1 ve θ_2 açık şekilde hesaplanabilir:

$$\hat{f}(x) = \mathrm{avg}(y^d) + \rho \frac{\mathrm{std}(y^d)}{\mathrm{std}(x^d)}(x - \mathrm{avg}(x^d))$$

- burada $x^d = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(N)} \end{bmatrix}^T$
- ρ (x ile y arasındaki korelasyon katsayısı):

$$\rho = \frac{x^T y}{\|x\| \|y\|}$$

Düz çizgi uydurma, örnek



Polinom uydurma

- $f_i = x^{i-1}$, $i = 1,\, 2,\, \ldots,\, p$
- model formu: derecesi p'den düşük bir polinom

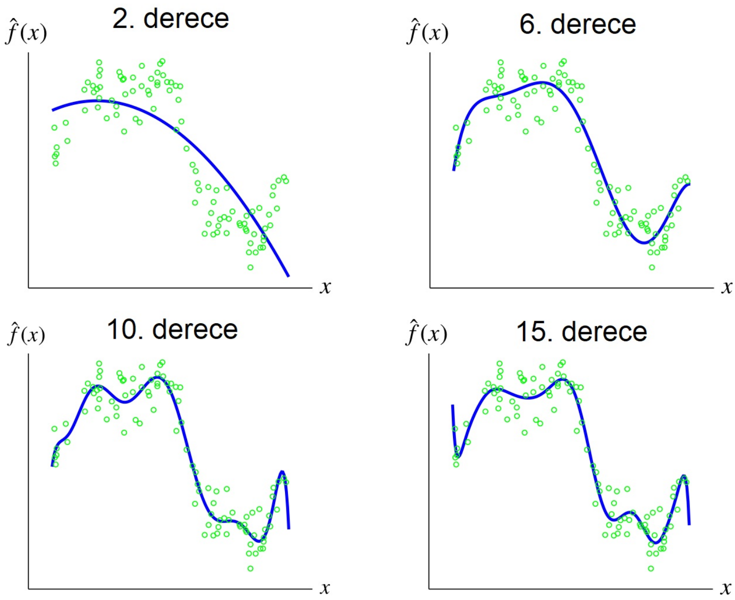
$$\hat{f}(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \cdots + \theta_p x^{p-1}$$

- dikkat: x^i "x üzeri i" demek; $x^{(i)}$ i. veri noktası
- A matrisinin formu:

$$A = \begin{bmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \cdots & (x^{(1)})^{p-1} \\ 1 & x^{(2)} & (x^{(2)})^2 & \cdots & (x^{(2)})^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x^{(N)} & (x^{(N)})^2 & \cdots & (x^{(N)})^{p-1} \end{bmatrix}$$

(bu matrise Vandermonde matrisi denir)

Polinom uydurma, örnek ($N = 100$)



Genel veri uydurma olarak bağlanım

- bağlanım (*regression*) modeli, $\hat{y} = \hat{f}(x) = x^T\beta + \nu$ ile verilen afin fonksiyondur
- $f_1(x) = 1$, $f_i(x) = x_{i-1}$ $\big(i = 2, 3, \ldots, n+1\big)$ şeklindeki taban fonksiyonları ile genel uydurma formunda uydurma yapar. model formu:

$$\hat{y} = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \cdots + \theta_{n+1} x_n = x^T\theta_{2:n+1} + \theta_1$$

- $\hat{y} = x^T\beta + \nu$ formunda yazarsak: $\beta = \theta_{2:n+1}$, $\nu = \theta_1$

Bağlanım olarak genel veri uydurma

- genel uydurma modeli:

$$\hat{f}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_p f_p(x)$$

- olağan varsayım: $f_1(x) = 1$
-
  - $\tilde{x} = \begin{bmatrix} f_2(x) & f_3(x) & \dots & f_p(x) \end{bmatrix}$ şeklinde dönüştürülmüş (*transformed*) öznitelikler
  - $\nu = \theta_1$, $\beta = \theta_{2:p}$

  tanımlarıyla, $\hat{f}(\tilde{x}) = \tilde{x}^T \beta + \nu$ formundaki bağlanım modeliyle aynıdır

Bölüm 2

Geçerleme

Genelleştirme

temel fikir:

- modelin amacı eldeki veriler için amaç değişkenini öngörmek **değildir**
- bunun yerine, modelin amacı yeni, önceden görülmemiş veriler için amaç değişkenini öngörmektir

- yeni, önceden görülmemiş veriler için makul öngörüler yapan bir modelin "genelleştirebilme yeteneği" (*generalization ability*) vardır (veya, "model genelleştirebiliyor" denir)
- yeni, önceden görülmemiş veriler için kötü öngörüler yapan modelde aşırı uyumlama (*over-fit*) sorunu vardır

Geçerleme

geçerleme: modelin genelleştirebilme yeteneğini test ermek için basit ve etkili bir yöntem

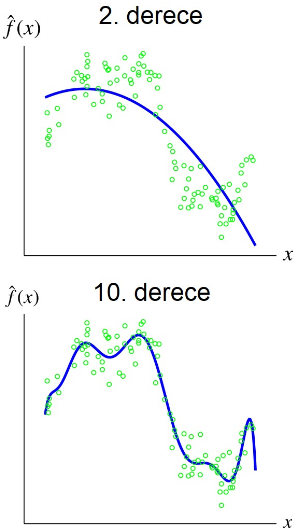
- asıl veri kümesini eğitim kümesi (*training set*) ve test kümesi (*test set*) olarak ayıralım
- sık kullanılan ayırmalar: $80\%/20\%$, $90\%/10\%$
- eğitim kümesi üzerinde modeli kuralım (eğitelim (*train*))
- sonra, model öngörülerini test kümesi üzerinde test edelim
- ayrıca, modelin eğitim ve test kümeleri için öngörü hatasının RMS değerlerini karşılaştırabiliriz
- eğer hatalar benzer ise, modelin genelleştirebileceğini tahmin edebiliriz

Geçerleme

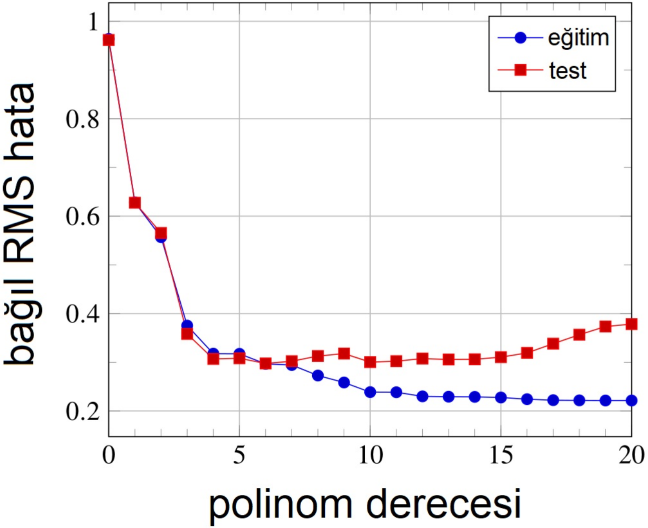
- geçerleme prosedürü, farklı aday modellerin arasından seçim yapmak için kullanılabilir, örneğin:
  - farklı dereceli polinomlar
  - farklı açıklayıcı değişkenler (*regressor*) kümelerine sahip bağlanım modelleri
- farklı modeller arasından (en) düşük test hatasına sahip olanı kullanmak isteriz

Geçerleme, örnek

modeller 100 veri noktası içeren eğitim kümesi ile uyduruldu, grafikler 100 veri noktası içeren test setini gösteriyor



Geçerleme, örnek



grafik 4., 5. veya 6. derecelerin makul seçenekler olduğunu gösteriyor

Çapraz geçerleme

çapraz geçerleme (*cross-validation*) prosedürü:

- veri kümesini k adet veri altkümesine (*fold*) ayır (örneğin: $k = 10$)
- i. altküme hariç bütün altkümeleri kullanarak modeli eğit
- i. altkümedeki veri üzerinde modeli test et
- bu işlemleri $i = 1, 2, \ldots, k$ için tekrarla

(bu yönteme k-kat (*k-fold*) çapraz geçerleme denir)

çapraz geçerleme sonuçlarını yorumlamak:

- test kümesi için RMS hatalar eğitim kümesi için olanlardan çok daha büyük ise modelde aşırı-uyum vardır
- test ve eğitim kümeleri için RMS hatalar benzer ve tutarlı ise, gelecekteki veriler için modelin benzer RMS hatalara sahip olacağını **tahmin** edebiliriz (kesin olarak bilemeyiz)

Çapraz geçerleme, örnek

- ev fiyatı tahmini; bağlanım modeli ($\hat{f}(x) = x^T \beta + \nu$)
- öznitelikler: alan (x_1) ($\times 92.9$ m^2), yatak odası sayısı (x_2)
- veri kümesi: 775 ev satışı verisi; 5 altkümeye ayrılıyor
- çapraz geçerleme ile her bir alt küme için bir bağlanım modeli eğitiliyor (örneğin, 1. modelde eğitim kümesi 2., 3., 4. ve 5. altkümeler, test kümesi 1. alt küme)

	model parametreleri			RMS hata	
kat	ν	β_1	β_2	eğitim	test
1	60.65	143.36	−18.00	74.00	78.44
2	54.00	151.11	−20.30	75.11	73.89
3	49.06	157.75	−21.10	76.22	69.93
4	47.96	142.65	−14.35	71.16	88.35
5	60.24	150.13	−21.11	77.28	64.20

Bölüm 3

Öznitelik mühendisliği

Öznitelik mühendisliği

öznitelik mühendisliği prosedürü:

- temel öznitelik vektörü n-vektör x ile prosedüre başla
- taban fonksiyonlarını ($f_1, f_2, \ldots, f_p$) seçerek

$$\begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_p(x) \end{bmatrix}$$

  şeklindeki eşlenmiş (*mapped*) öznitelikler vektörünü oluştur
- eşlenmiş öznitelikleri olan, parametrelere-göre-doğrusal modeli veriye uydur

$$\hat{y} = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_p f_p(x)$$

- modelin geçerleme analizini yap

Öznitelikleri dönüştürmek

- standartlaştırma: x_i'yi $\frac{x_i - b_i}{a_i}$ ile değiştir
  - $b_i \approx$ özniteliğin veri kümesi için ortalama değeri
  - $a_i \approx$ özniteliğin veri kümesi için standart sapması

  bu şekilde (standartlaştırılmış) yeni özniteliklere "standart normal değişken" (*z-score*) denir
- logaritmik dönüşüm: x_i negatif olmayan sayı ise ve geniş bir değer aralığında yer alıyorsa, $\log(1 + x_i)$ ile değiştir
- yüksek ve alçak öznitelikler: $\max(x_1 - b, 0)$ ve $\min(x_1 - a, 0)$ ile verilen yeni öznitelikler oluştur (bunlara asıl öznitelik x'in yüksek ve alçak versiyonları denir)

Öznitelik mühendisliği, örnek

- ev fiyatı tahmini
- temel öznitelikler ile başlayalım
  - x_1: alan ($\times 92.9$ m^2)
  - x_2: yatak odası sayısı
  - x_3: apartman dairesi ise $x_3 = 1$, müstakil ev ise $x_3 = 0$
  - x_4: adresin posta kodu (62 farklı değer alabilir)
- 8 adet taban fonksiyonu kullanalım:
  - $f_1(x) = 1$, $f_2(x) = x_1$, $f_3(x) = \max(x_1 - 1.5, 0)$
  - $f_4(x) = x_2$, $f_5(x) = x_3$
  - $f_6(x)$, $f_7(x)$, $f_8(x)$: x_4'in Boole fonksiyonları (birbirine yakın posta kodlarından oluşan 4 grubu (yani, mahalleleri) ifade ederler)
- 5-kat model geçerleme yapalım

Öznitelik mühendisliği, örnek

	model parametreleri								RMS hata	
kat	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	eğitim	test
1	122.35	166.87	−39.27	−16.31	−23.97	−100.42	−106.66	−25.98	67.29	72.78
2	100.95	186.65	−55.80	−18.66	−14.81	−99.10	−109.62	−17.94	67.83	70.81
3	133.61	167.15	−23.62	−18.66	−14.71	−109.32	−114.41	−28.46	69.70	63.80
4	108.43	171.21	−41.25	−15.42	−17.68	−94.17	−103.63	−29.83	65.58	78.91
5	114.45	185.69	−52.71	−20.87	−23.26	−102.84	−110.46	−23.43	70.69	58.27