Makine Öğrenmesi

5. hafta

- Doğrusal Regresyon
- Çoklu Doğrusal Regresyon
  - En küçük kareler yöntemi

Doğrusal Regresyon

Doğrusal regresyonun temelinde bir özellik vektörüne ait giriş değişkeninin ağırlık kazanmasıyla bir hedef değişkeni belirli bir hata ile ürettiği varsayılır. Bu varsayım aşağıdaki eşitlik ile temsil edilir.

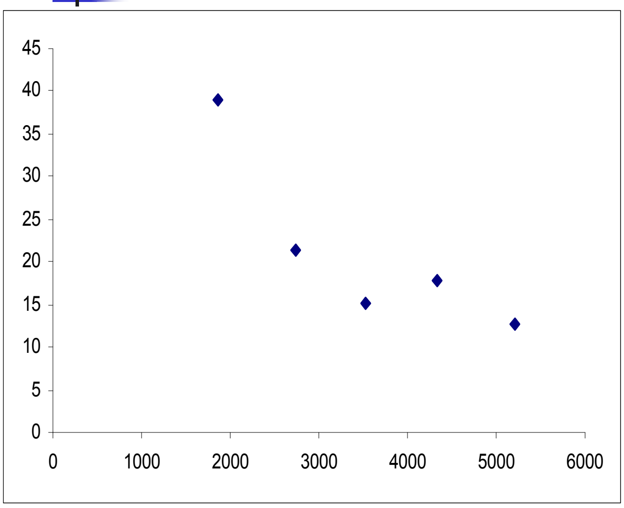
$$Y = W_0 + W_1 X + E$$

W ağırlıkları, X verinin özellik vektörünü, E hatayı ve Y değişkeni de hedef değişkeni temsil etmektedir.

Doğrusal Regresyon

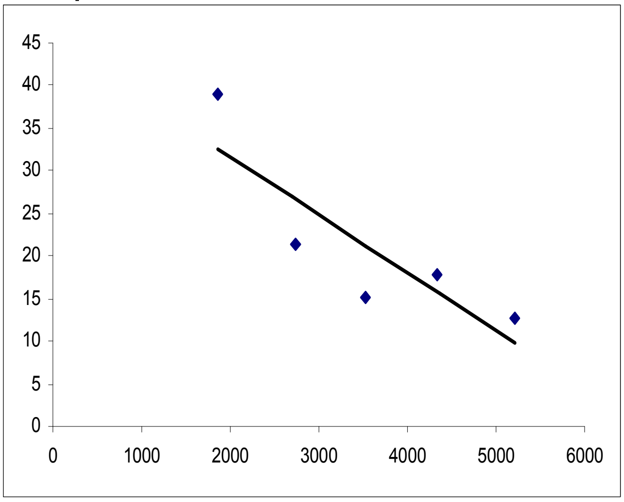
Her veri örneği aslında bir doğru denklemine *E* hatası oranında uzaktır. Bu hata değerleri çok küçük olduğu varsayımı ile çözüm sırasında ihmal edilirler. *X* ve *Y* vektörlerini içeren veri kümesi kullanılarak *W* ağırlık vektörünün tespiti regresyon analizinin temel amacını oluşturur.

Örnek



X	Y
2743	21,4
3518	15,2
1855	38,9
5214	12,7
4341	17,8

Örnek



X	Y
2743	21,4
3518	15,2
1855	38,9
5214	12,7
4341	17,8

Y = 45.11- 0.007 X

MSE = 24.07

Çoklu Doğrusal Regresyon

Doğrusal regresyonun *M* boyutlu bir veri kümesine uygulanabilen şeklidir. Tüm giriş değişkenlerinin ağırlıklı toplamının hedef değişkeni belirli bir hata ile ürettiği varsayılır.

$$Y = W_0 + \sum_{j=1}^{M} W_j X_j + E$$

W_j ağırlıkları, X_j verinin özellik vektörlerini, E hatayı ve Y değişkeni de hedef değişkeni temsil etmektedir.

Çoklu Doğrusal Regresyon

Veri örnekleri *M*+1 boyutlu bir hiper-düzlem denklemine, ihmal edilebilecek kadar küçük *E* hatası oranında uzaktır. Amaç, doğrusal regresyondaki gibi *W* ağırlık vektörünün tespit edilmesine dayanır.

Hem basit doğrusal hem de çoklu doğrusal regresyonun en temel çözümü en küçük kareler yöntemine (least squares) dayanır.

En Küçük Kareler (Least Squares)

En küçük kareler yöntemiyle çözülebilen bu denklemlerin matematiksel çözümü aşağıdaki gibi ifade edilebilir.

$$W_j = \frac{X_j D}{X_j^2} = \frac{\sum_{i=1}^{N} x_{ij} d_i}{\sum_{i=1}^{N} x_{ij}^2} \qquad \mathbf{W} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D}$$

N değeri verideki toplam örnek sayısını, *i* indisi her bir örneğin verideki sırasını ve *j* indisi de verinin boyutlarını temsil eder.

En Küçük Kareler (Least Squares)

En küçük kareler yönteminin temel prensibi aşağıdaki gibi ifade edilen ortalama karesel hatanın (MSE) minimize edilmesidir.

$$\frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(d_i - \sum_{j=1}^{M} w_j x_{ij}\right)^2$$

Bu eşitlikle ifade edilen e_i hataları olabilecek en küçük değerlere sahiptir ve doğrusal bir çözüm ile tahmin edilemez kabul edilirler.

Eşikleme ve Yarışmalı Sınıflandırma

Tahmin ve kestirim yöntemlerinin başarıları MSE ölçütü yardımıyla karşılaştırılır. Regresyon gibi tahminlemede kullanılan bir çözümleyicinin sınıflandırma yapabilmesi için ise hesaplanan sonuçlar üzerinde bir eşikleme yapılmalıdır. İki sınıflı sistemlerde etiketler için 0 ve 1 değerleri seçilirse eşik değeri de 0.5 olmalıdır. Çok sınıflı sistemlerde ise yarışma usulü sonuç tayin edilir.

Sınıflandırma Örneği

Kümeleme ve sınıflandırma için sık kullanılan İRİS veri kümesinden bir kesit alınmıştır. Veri, giriş için 4 özellik vektörüne sahip X değişkenini, hedef içinse 3 sınıf değeri taşıyan Y değişkenini bulundurur. Sınıflandırma yapılacağı için Y değikeni üç ayrı lojik değişkene dönüşmelidir.

X_1	X_2	X_3	X_4	Y
5.1	3.5	1.4	0.2	A
4.9	3	1.4	0.2	A
7	3.2	4.7	1.4	B
6.4	3.2	4.5	1.5	B
6.3	3.3	6	2.5	C
5.8	2.7	5.1	1.9	C

Sınıflandırma Örneği

Sınıflandırma sistemlerinin en kolay öğrendiği hedefler 0 ve 1 olmak üzere ikili sınıf bilgileridir. Tahmin sistemi gerçel sayılar üretir. Bu yüzden tek çıkışlı sistemde 0.5 eşikleme kullanılarak yuvarlama yapılır ve çok çıkışlı sistemlerde ise değişkenlerden hangisi büyükse o kazandı denilir.

Y	Y_1	Y_2	Y_3
A	1	0	0
A	1	0	0
B	0	1	0
B	0	1	0
C	0	0	1
C	0	0	1

Sınıflandırma Örneği

Çoklu doğrusal regresyon çözümü ile aşağıdaki 3 denklem bulunur;

$$Y_1 = 1.2 + 0.29X_1 - 0.16X_2 - 1.02X_3 + 1.43X_4$$
$$Y_2 = -3.4 + 0.82X_1 - 0.2X_2 + 0.04X_3 - 0.52X_4$$
$$Y_3 = 3.2 - 1.11X_1 + 0.35X_2 + 0.98X_3 - 0.92X_4$$

Bulunan ortalama karesel hatalar;

Y_1 için MSE=0.030, Y_2 için MSE=0.301,

Y_3 için MSE=0.014

Sınıflandırma Örneği

Burada $\bar{Y}$ değikenleri regresyon denklemleriyle hesaplanan değerleri gösterir. İlk iki örnekte en büyük değer $\bar{Y}_1$, sonraki iki örnekte $\bar{Y}_2$ ve son iki örnekte $\bar{Y}3$ değişkeni kazanan sınıfı temsil eder. Y değişkenleri sırasıyla A, B ve C sınıflarını temsil eder. Buna göre 6 örneğin tümü de doğru sınıflandırılmıştır.

$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$
0.98	0.06	-0.04
1.00	-0.01	0
-0.06	1.19	-0.13
0.11	0.63	0.25
-0.02	0.07	0.95
-0.02	0.05	0.96

Doğrusal Olmayan Regresyon

Doğrusal olmayan regresyon modelleri de aynı doğrusal modeller gibi basit bir denklemle gösterilebilirler. Doğrusal olmayan regresyon modellerinde bu denklemdeki parametre sayısı verideki değişken sayısıyla doğrudan ilişkili olmayabilir. Doğrusal olmayan regresyon modellerinin parametre tahminleri için önerilen birçok yöntem vardır. Bunlardan en çok bilinenleri en küçük kareler, en çok olabilirlik (maximum likelihood) ve gauss-newton yöntemleridir.

MATLAB Uygulaması

```
>edit Regresyon_ornek.m
```

Hazırlanmış olan farklı datasetler yüklenerek En Küçük Kareler algoritması deneyi yapılmaktadır. Matlab komutları ve `regress` komutu üzerinde değişiklikler yapılarak kodlar irdelenmelidir.

ÖDEV

Doğrusal olmayan regresyon modellerinin parametre tahminleri için kullanılan aşağıdaki iki yöntemi araştırınız.

- Maximum Likelihood
- Gauss-Newton