

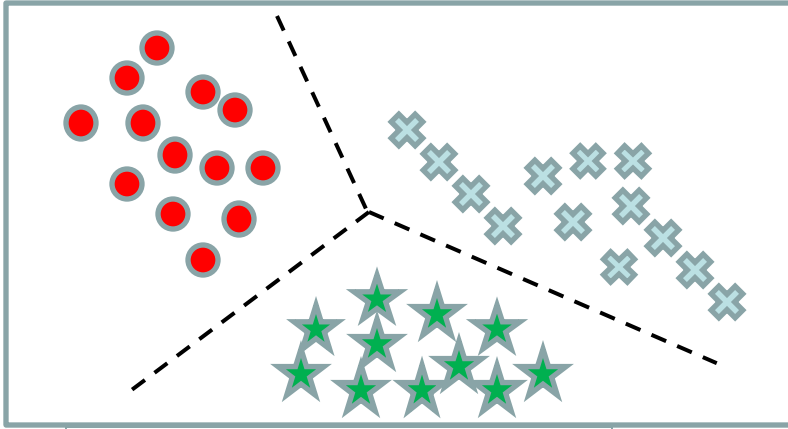
Regresyon Analizi

Prof. Dr. M. Elif Karslıgil

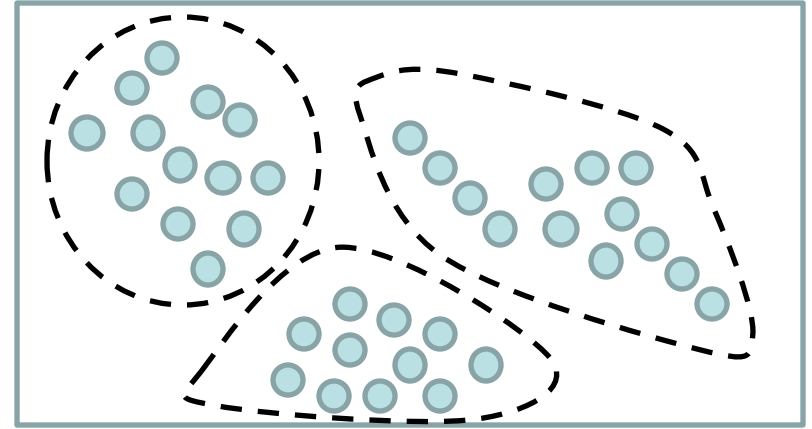
Yıldız Teknik Üniversitesi - Bilgisayar Mühendisliği Bölümü

Akıllı Sistemler Laboratuvarı

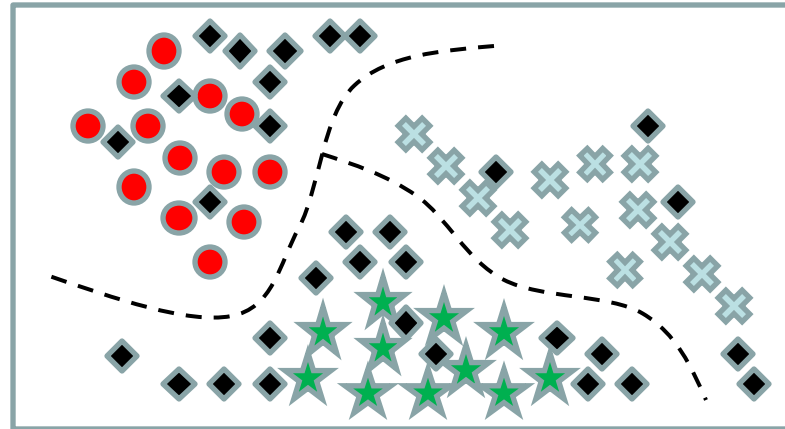
Öğrenme Yöntemleri



Eğitmenli Öğrenme



Eğitmensiz Öğrenme



Yarı-eğitmenli Öğrenme

Eğitmenli Öğrenme

- Regresyon (Regression) :

Sürekli veya ayrık giriş \rightarrow sürekli çıkış

(yağış miktarı \rightarrow mısır hasılatı)

(son 1 ay + hasta sayısı \rightarrow haftaya + kaç hasta olacak)

- Sınıflandırma(Classification) :

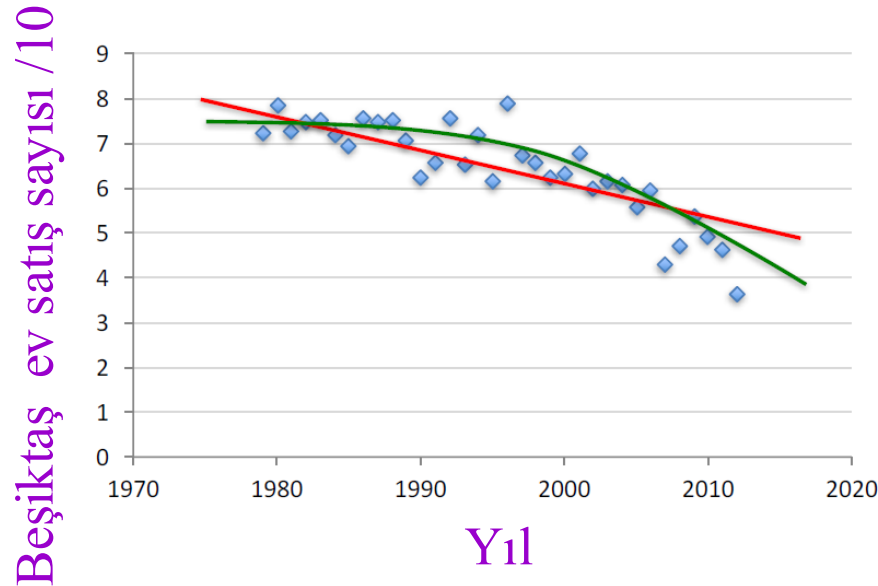
Sürekli veya ayrık giriş \rightarrow ayrık çıkış

(kredi kartı harcamaları \rightarrow sahtekarlık, normal kullanım)

(ses verisi \rightarrow söylenen kelime)

Regresyon

- Verilen : $\langle (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \rangle$
- İstenen : Verilen x girişleri ve y çıkışları ile $y = f(x)$ fonksiyonunu öğren. Bilinmeyen bir x örneği için y çıkışını tahmin et.



Lineer Regresyon

- Verilen x girişı için y çıkışı ne olur?
- Lineer regresyon :

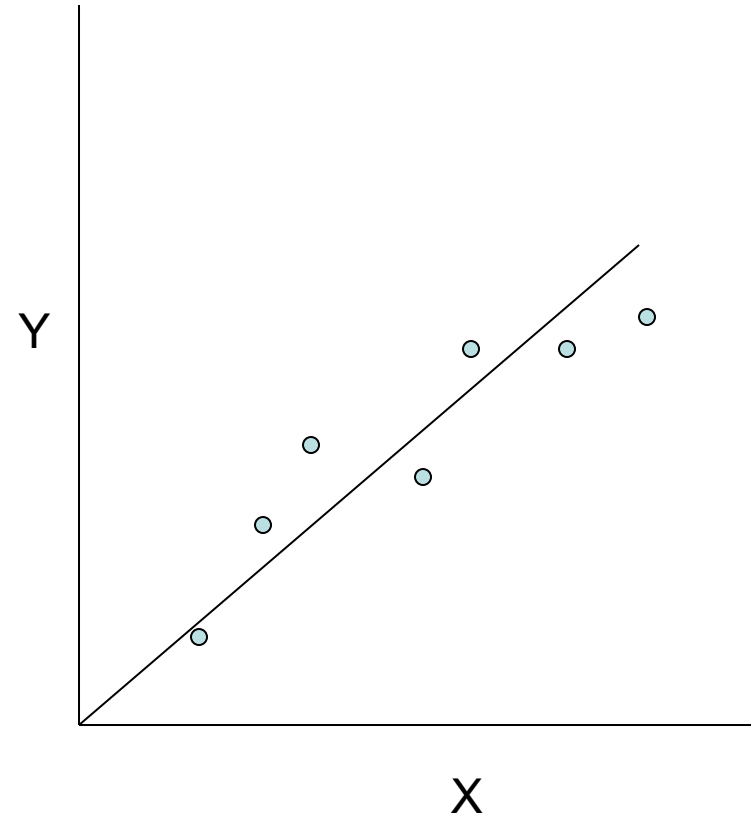
Bilinmeyen \swarrow $y = w^*x + \varepsilon$ \searrow Gözlenen

w : ağırlık değeri

ε : hata

Hedef hatayı minimum yapmak:

$$\arg \min_w \sum_i (y_i - wx_i)^2$$

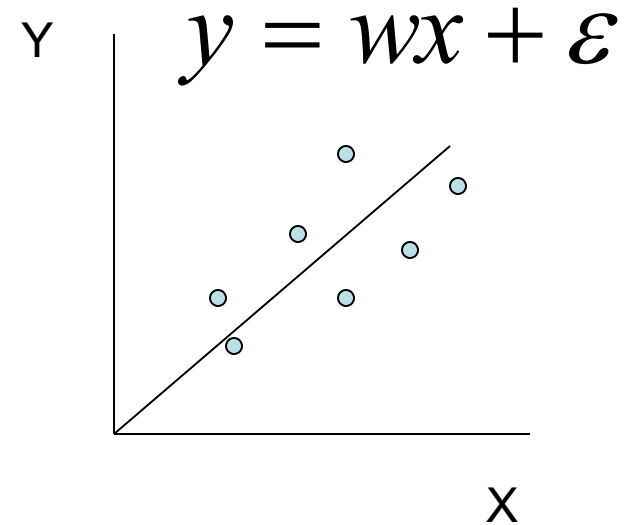


Lineer Regresyon

- Verilen $\langle \mathbf{x}_i, y_i \rangle$ çiftleri için w ağırlık değerleri hesaplanır.
- Hedef, hatanın minimum olması:

$$\arg \min_w \sum_i (y_i - wx_i)^2$$

↑ ↑
istenen çıkış hesaplanan çıkış



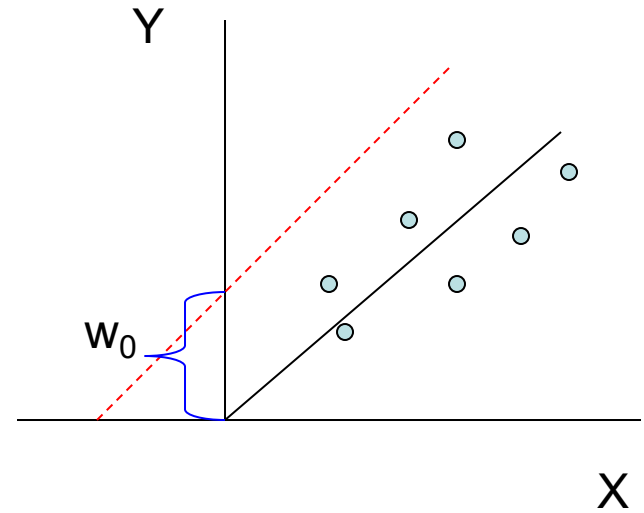
Bias terimi

- Doğru orjinden geçmiyorsa bias terimi (w_0) eklenir :

$$y = w_0 + w_1x + \varepsilon$$

$$w_0 = \frac{\sum_i y_i - w_1 x_i}{n}$$

$$w_1 = \frac{\sum_i x_i (y_i - w_0)}{\sum_i x_i^2}$$



Çok Değişkenli Lineer Regresyon

- Birden fazla özellik varsa :

$$y = w_0 + w_1x_1 + \dots + w_kx_k + \varepsilon$$

Google Hisse Seneti

Yahoo Hisse Seneti

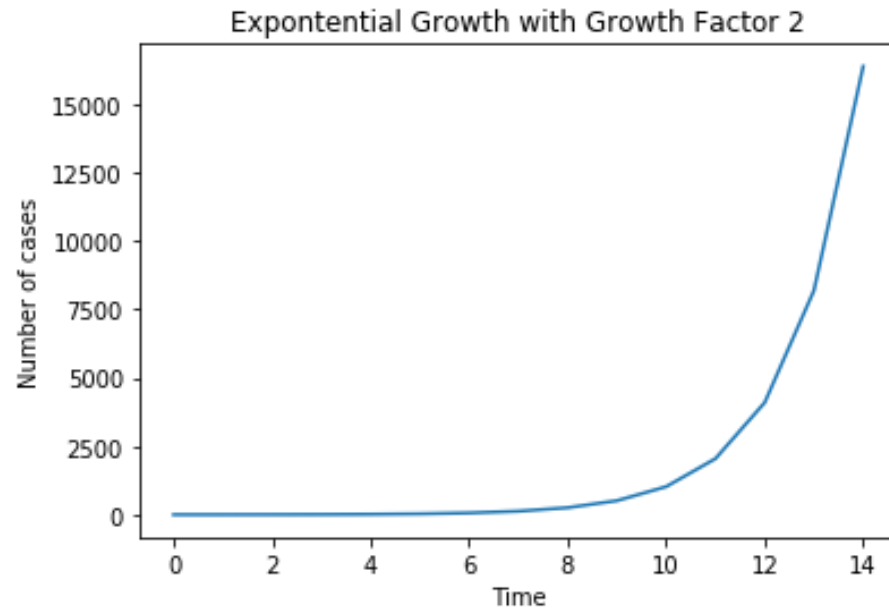
Microsoft Hisse Seneti

Öğrenme : En uygun ağırlıkların belirlenmesi

Lineer Regresyon - Örnek

- Amerika’da verilen tarihlerde Corona virüsü vaka sayısı:
giriş bilgisi: zamana göre değişen toplam vaka sayısı
çıkış bilgisi: k. günde kaç vaka olacağı

location	United States
Row Labels	Sum of total_cases
24/01/2020	1
27/01/2020	5
30/01/2020	5
02/02/2020	8
05/02/2020	11
08/02/2020	12
11/02/2020	13
14/02/2020	15
17/02/2020	15
20/02/2020	15
23/02/2020	35
26/02/2020	53
29/02/2020	62
03/03/2020	64
06/03/2020	148
09/03/2020	213
12/03/2020	987
15/03/2020	1678
16/03/2020	1678



Linear Regresyon

Online-linear-regression (D , *number of iterations*)

Initialize weights $\mathbf{w} = (w_0, w_1, w_2 \dots w_d)$

for $i=1:1: \text{number of iterations}$

do **select** a data point $D_i = (\mathbf{x}_i, y_i)$ from D

set learning rate $\alpha(i)$

update weight vector

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha(i)(y_i - f(\mathbf{x}_i, \mathbf{w}))\mathbf{x}_i$$

end for

return weights \mathbf{w}

Linear Regresyon

Online-linear-regression (D , *number of iterations*)

Initialize weights $\mathbf{w} = (w_0, w_1, w_2 \dots w_d)$

for $i=1:1: \textit{number of iterations}$

do **select** a data point $D_i = (\mathbf{x}_i, y_i)$ from D

set learning rate $\alpha(i)$

update weight vector

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha(i)(y_i - f(\mathbf{x}_i, \mathbf{w}))\mathbf{x}_i$$

end for

return weights \mathbf{w}

Lojistik Regresyon

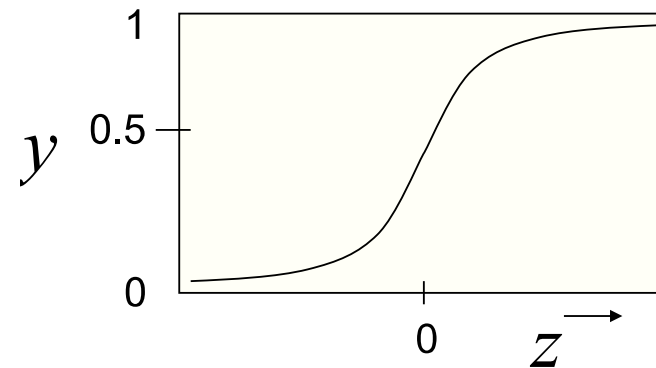
- Sınıflandırma yöntemidir.
- Verilen giriş değerleri için bilinmeyen örneğin hangi sınıfa ait olduğunun **olasılığını** verir.
- Giriş değerleri nümerik olmalıdır.
- Genellikle **iki sınıf** veriyi sınıflandırmak için kullanılır.
- Sınıf sayısı ikiden fazla ise, sınıflar ikili ikili değerlendirilip **en yüksek olasılıklı sonuç** alınır.

Lojistik Regresyon

- Lojistik Regresyon için aşağıdaki fonksiyon kullanılır.

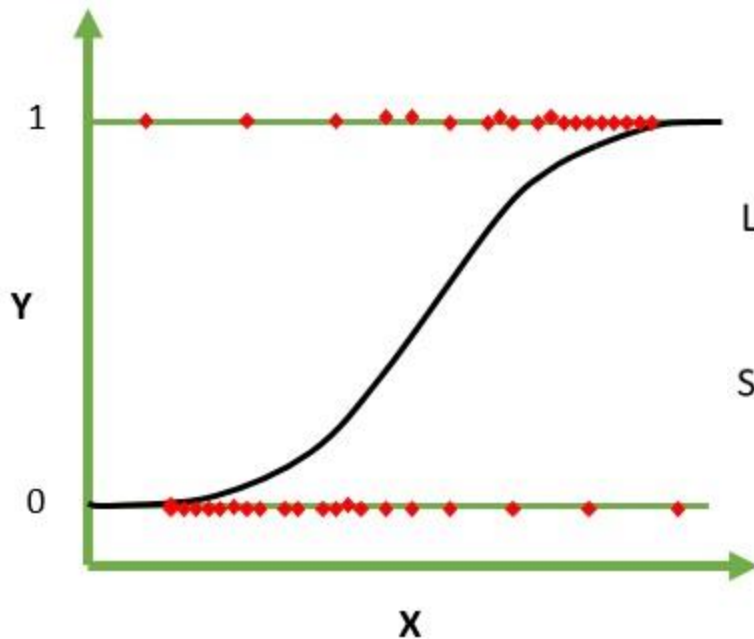
$$y(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0) \quad \sigma(z) = \frac{1}{1 + \exp(-z)}$$

- Y çıkış değeri $[0,1]$ aralığındadır. Bilinmeyen örneğin her sınıf için olasılık değerini verir.



Lojistik Regresyon

Verilen değerler için bu kişinin hasta olma olasılığı nedir? Sonuç 0.5'den büyük ise hasta, küçük ise sağlam



Linear regression

$$Y = b_0 + b_1 \times X_1 + b_2 \times X_2 + \dots + b_K \times X_K$$

Sigmoid Function

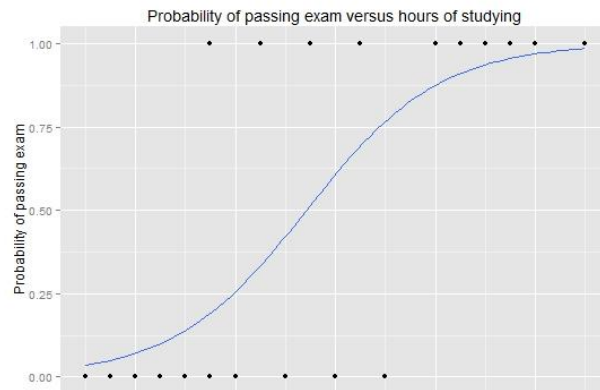
$$P = \frac{1}{1 + e^{-Y}}$$

Lojistik Regresyon - Örnek

- Problem: Given the number of hours a student spent learning, will (s)he pass the exam?
- Training data (top row: $x^{(i)}$, bottom row: $t^{(i)}$)

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

- Learn w for our model, i.e., logistic regression (coming up)
- Make predictions:



Hours of study	Probability of passing exam
1	0.07
2	0.26
3	0.61
4	0.87
5	0.97

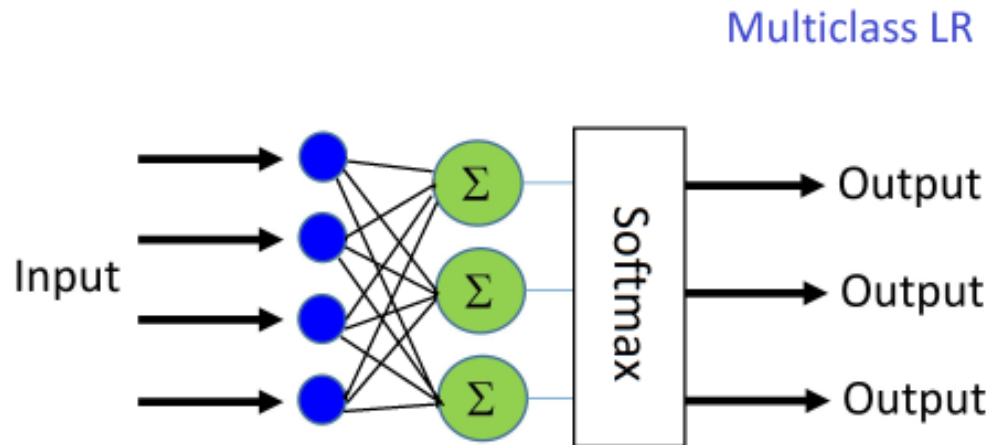
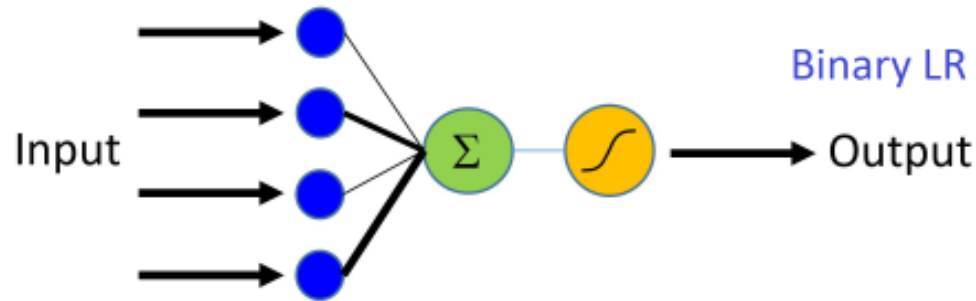
Lojistik Regresyon

Online logistic regression algorithm

```
Online-logistic-regression (stopping_criterion)  
  initialize weights  $\mathbf{w} = (w_0, w_1, w_2 \dots w_d)$   
  while stopping_criterion = FALSE  
    do      select next data point  $D_i = \langle \mathbf{x}_i, y_i \rangle$   
             set  $\alpha(i)$   
             update weights (in parallel)  
                 $\mathbf{w} \leftarrow \mathbf{w} + \alpha(i)[y_i - f(\mathbf{w}, \mathbf{x}_i)]\mathbf{x}_i$   
    end  
  return weights  $\mathbf{w}$ 
```

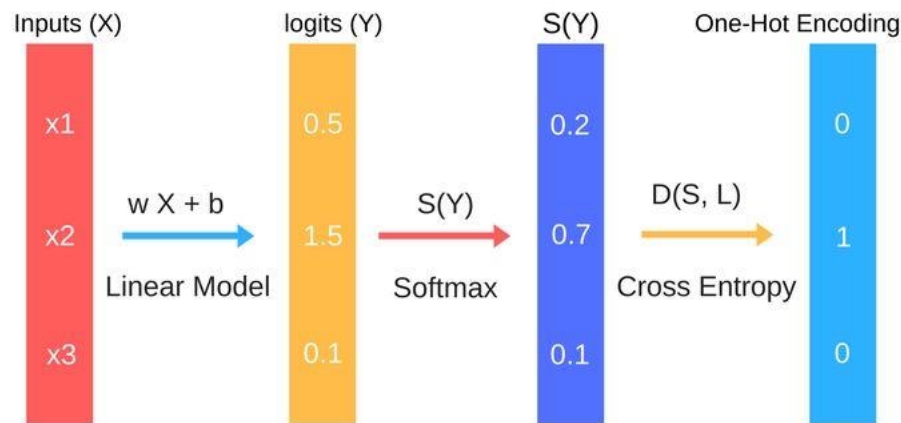

Multinomial Lojistik Regresyon

Sınıf sayısı ikiden fazla ise kullanılır.



Softmax Regresyon

- Verilen X bilinmeyen örneği için softmax regresyon her sınıf için ayrı ayrı sonuç hesaplar.
- Bilinmeyen X örneği en yüksek olasılıklı sınıfa aittir.



Multinomial Logistic Classifier

@ dataaspirant.com

Multinomial Lojistik Regresyon

Her sınıfın kendi ağırlık vektörü vardır:

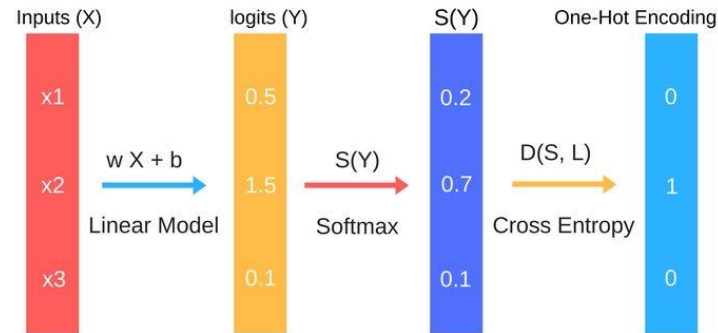
$$w_1, w_2, \dots, w_K$$

Bir örnek x için her sınıfın **skorunu (logit)** hesaplarız:

$$z_k = w_k^T x$$

Daha sonra bu skorları **softmax** fonksiyonundan geçiririz:

$$P(y = k|x) = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}}$$



Multinomial Logistic Classifier

@ dataaspirant.com


Softmax

$$P(y = k|x) = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}}$$

Softmax, tüm skorları (z değerleri) **pozitif** hale getirip **normalize eder**, böylece toplamı **1** olur → geçerli bir olasılık dağılımı elde ederiz.

Örnek:

Sınıf	Skor z	Softmax olasılığı
1	2.0	0.71
2	1.0	0.26
3	0.0	0.03

Toplam = 1 

Bu olasılıklar "bu örneğin hangi sınıfa ait olma olasılığı"dır.

Model, en yüksek olasılığa sahip sınıfı seçer:

$$\hat{y} = \arg \max_k P(y = k|x)$$