Makine Öğrenmesinde Regresyon

—

Derleyen: Ahmet Kaşif
Bursa, 2018

İçerik

- Regresyon
  - Doğrusal Regresyon
    - En Düşük Karesel Hata (MSE)
    - Çok Sınıflı Doğrusal Regresyon
  - Lojistik Regresyon
    - Çok Sınıflı Lojistik Regresyon
    - Örnek Python Uygulaması

Regresyon

Regresyon, bağımlı ve bağımsız değişkenler üzerinden önceden etiketlenmiş örnekleri kullanarak, yeni bir örneğin dahil olduğu sınıfı tahmin etmede kullanılan bir tekniktir. Aslen bir istatistik kavramı olup, makine öğrenmesi alanında da çokça kullanılmaktadır. Makine öğrenmesinde, destekli öğrenme sınıfından bir tekniktir.

Eğer tanımlayıcı değişken sayısı bir tane ise basit regresyon, birden fazla sayıda ise çok değişkenli regresyon olarak ifade edilir.

Tahmin etmeye çalıştığımız değişken (bağımlı değişken) eğer sürekli bir yapıda ise lineer regresyon, sabit bir kategoriyi temsil eden kesikli bir yapıda ise lojistik regresyon kullanılır.

Örnek : Evlerimizin metrekare cinsinden büyüklüğü ve fiyatının tutulduğu bir veri setimiz olduğunda, ev büyüklüğü değiştikçe fiyatın ne olacağını tahmin etmede doğrusal regresyon, 200 bin ₺'den yüksek fiyatlı bir evin satın alınması ihtimalinin ise lojistik regresyon modeliyle çözülmesi önerilir.

Doğrusal Regresyon

Doğrusal regresyon modeli, veri setindeki (2|n) boyut için veriyi temsil eden en iyi (doğru|hiperdüzlem) in çizilebilmesini hedefler. Temelde bir matris problemidir ve lineer cebir kullanılarak çözülebilir.

Bağımlı ve bağımsız değişkenler birbirleriyle doğrusal ilişkidedir. Birinin artması veya azalması, diğerinin de belli oranda artması veya azalması sonuçlanır.

$$Y = \hat{Y} + \varepsilon = (B_0 + B_1.X_1 + B_2.X_2 + ... + B_N.X_N) + \varepsilon$$

Y: Hedef Gerçek Değeri (Bağımlı, tahmin edilen değişken, çıktı)

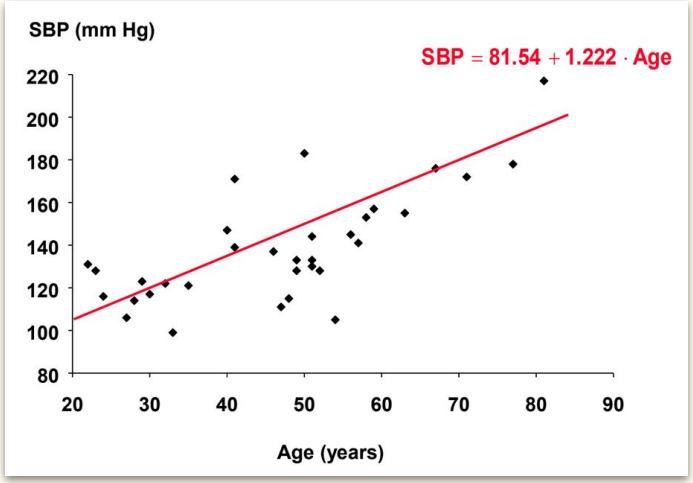
$\hat{Y}$: Tahmin Edilen Hedef Değeri

e: Sapma (Hata)

B_0: Kesim Noktası (Bias), $B_{1\text{-}N}$: Eğim (Bağımsız, belirleyici değişkenler)

Doğrusal Regresyon

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217



Doğrusal Regresyon - Parametrelerin Bulunması

Doğrusal regresyon modelinde parametrelerin bulunmasında en düşük karesel hata yöntemi kullanılır (MSE).

Bu yöntemde, veri seti ile tahmin arasındaki farklılığı azaltmayı hedefleyen bir maliyet fonksiyonu tanımlanır. Bu maliyet fonksiyonunun, parametrelerin aldığı farklı değerlere göre tahmin ile gerçek değer arasındaki farkı en düşük tutmasına çalışılır.

En Düşük Karesel Hata Yöntemi (MSE)

Yöntemin işleyişi şu şekildedir :

1. B_0 (bias) ve $B_{1\text{-}N}$ (bağımsız değişken(ler)) için başlangıç değerleri belirlenir.
2. Fonksiyonun karesel hatası hesaplanır.
3. Fonksiyonun karesel hatası en düşük olana kadar adım 1 ve 2 tekrar edilir.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

En Düşük Karesel Hata Yöntemi

Yöntemde, B0 ve B1 parametreleri 0'dan başlatılır. Şekillerde bu parametrelerin nasıl güncellendikleri görülmektedir.

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

$$\hat{B}_1 = (\sum X_i Y_i - (\sum X_i * \sum Y_i)/n)/(\sum X_i^2 - (\sum X_i)^2/n)$$

$$\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{X}$$

Çok Sınıflı Doğrusal Regresyon

İkiden fazla tanımlayıcı özelliğe sahip verilerin kullanıldığı bu tür regresyon problemlerinin çözümünde kullanılan yöntemlerden bazıları şu şekildedir:

- Multi Response Linear Regression : Karar ağaçları ile birden fazla doğrusal regresyon modelinin birleştirilmesiyle fikri üzerinden işler. Bir karar ağacı oluşturulur, bu ağaçtaki her yaprak bir doğrusal regresyon modelini temsil eder.
- Partial Regression Coefficients : Tüm bağımsız değişkenler için, bu değişkenlerden (X_i) birinin bir birim değiştirilip, diğerleri sabit tutulduğunda tahmin değerinin (Y) değişiminin gözlenmesini hedefler. Örnek: SBP'nin yaş, ağırlık ve boy gibi etkenlere göre değişimi

Lojistik Regresyon

Regresyon tekniğinin çıkış noktası, sınıflandırma problemidir ve verilerin dahil olabileceği sınıfların olasılık bilgileri, sınıfları tahmin etmekten daha etkili bir yöntem olacaktır. Fakat, doğrusal regresyon'da elde edilen veriler, olasılıksal kullanım için uygun değildir, etkili bir sonuç almak için normal dağılım gösteren bir veriye ihtiyaç duyar ve değişkenler süreklidir.

Lojistik regresyon'da amaç, doğrusal regresyonu olasılık çıktısı verecek şekilde güncelleyerek bu açığı kapatmaktır. Verilerin normal dağılım göstermemesi sonucu etkilemez ve değişkenler kategoriktir.

Lojistik Regresyon

Odds: Bir olayın olma olasılığı (p) iken, odds p / (1 - p) ‘ dir. Olayın olma olasılığı, lojistik regresyonda bağımsız değişkenler ve katsayılar ile ifade edildiğine göre, aşağıdaki eşitlik yazılabilir:

$$p = B_0 + B_1X$$

$$\text{logit}(y) = \ln(\text{odds}) = \ln(1/(1\text{-}p)) = B_0 + B_1X$$

$$p = e^{B0 + B1X} \,/\, 1 + e^{B0 + B1X}$$

Lojistik Regresyon

Aynı mantık, birden fazla bağımsız değişken için de geçerlidir:

$$\text{logit}(y) = \ln(1/(1\text{-}p)) = B_0 + B_1X + \ldots + B_KX_K$$

$$p = e^{B0 + B1X + \ldots + BkXk} \,/\, 1 + e^{B0 + B1X + \ldots + BkXk}$$

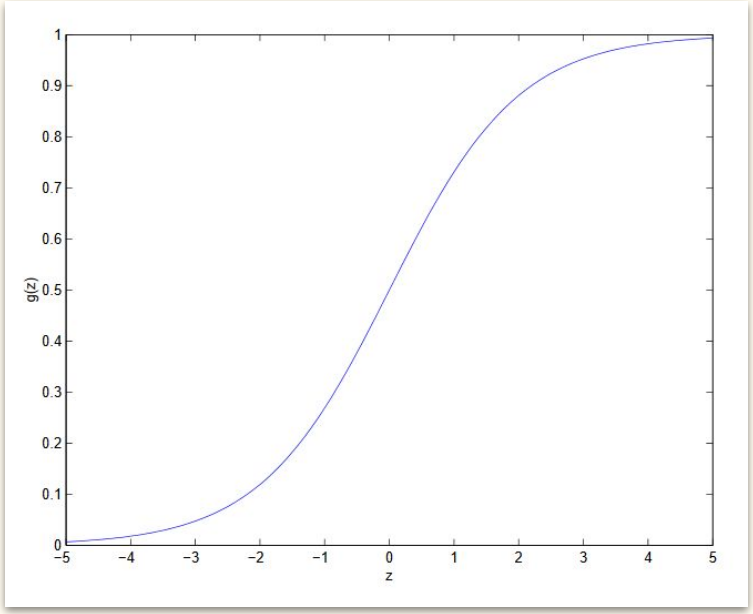
Lojistik (Sigmoid) Fonksiyonu

Lojistik fonksiyonu (diğer adıyla sigmoid - logit), 0 ve 1 arasında kalarak yumuşak geçiş yapan bir s eğrisi fonksiyonudur.

Gerçek değerli çıktı üretir.

İkili sınıflandırma için kullanılır.

$$Logit(x) = \frac{e^x}{1+e^x}$$



Lojistik Regresyon

İkili Sınıflandırma Problemi : Çıktı olarak doğru/yanlış aranan ikili sınıflandırma probleminde doğrusal regresyon, lojistik regresyona göre daha verimli sonuç vermektedir, bu nedenle lojistik regresyonun 2 sınıflı modellerde kullanımı önerilmemektedir.

Çok Değişkenli Sınıflandırma Problemi : İkiden fazla sınıfa sahip olduğumuzda ne yapmalıyız ?

Çok Değişkenli (Sınıflı) Lojistik Regresyon

Çok sınıflı durumlarla karşılaştığımızda, lojistik regresyon kullanılabilmekte olup, uygulamada iki farklı sınıflandırma yöntemi önerilmektedir:

- Veriler aynı anda birden fazla sınıfa dahil olabilir. (Multi-Label)
- Veriler aynı anda tek sınıfa dahil olur ve sınıflar özgündür. (Multinomial)

Çok Değişkenli (Sınıflı) Lojistik Regresyon

Tanımlayıcı bağımsız değişkenlerden diğerleri sabit iken birindeki birim miktar artışın log(odds)'ta meydana getirdiği artışın iteratif bir yöntemle tüm bağımsız değişkenler için uygulanmasını ve olasılıksal değerin en iyileştirilmesini amaçlar.

Lojistik Regresyon - Örnek Python Uygulaması

```
1.  import pandas
2.  from sklearn import model_selection
3.  from sklearn.linear_model import LogisticRegression
4.  url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
5.  names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
6.  dataset = pandas.read_csv(url, names=names)
7.  array = dataset.values
8.  X = array[:,0:4]
9.  Y = array[:,4]
10. validation_size = 0.20
11. seed = 7
12. X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, Y, test_size=validation_size, random_state=seed)
13. print("Logistic Regression")
14. logreg = LogisticRegression()
15. logreg.fit(X_train, Y_train)
16. lrPredictions = logreg.predict(X_validation)
17. print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_validation, Y_validation)))
18. from sklearn import model_selection
19. kfold = model_selection.KFold(n_splits=10, random_state=7)
20. modelCV = LogisticRegression()
21. scoring = 'accuracy'
22. results = model_selection.cross_val_score(modelCV, X_train, Y_train, cv=kfold, scoring=scoring)
23. print("10-fold cross validation average accuracy: %.3f" % (results.mean()))
```

Örnek Python Uygulaması - Çıktılar

Accuracy of logistic regression classifier on test set: 0.80
10-fold cross validation average accuracy: 0.967

Kaynaklar

- https://gerardnico.com/data_mining/start (Son Erişim : 22.03.2018)
- https://datascience.stackexchange.com/questions/10188/why-do-cost-functions-use-the-square-error (Son Erişim 22.03.2018)
- https://machinelearningmastery.com/logistic-regression-for-machine-learning/ (Son Erişim 22.03.2018)
- An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain (Makale)(Son Erişim: 22.03.2018)