Sınıflandırma

Fuat Can Beylunioğlu

December 23, 2017

Giriş

- Sınıflandırma makine öğreniminin en önemli konularından biridir.
- Çok geniş bir kapsam alanı vardır.
- Kredi değerleme, hassasiyet (sentiment) analizi, ev fiyatlama vs. gibi her alanda karşımıza çıkar.
- Bu kısımda lineer sınıflandırıcılar, karar ağaçları ve türleri, ve k-NN yöntemlerine odaklanacağız.

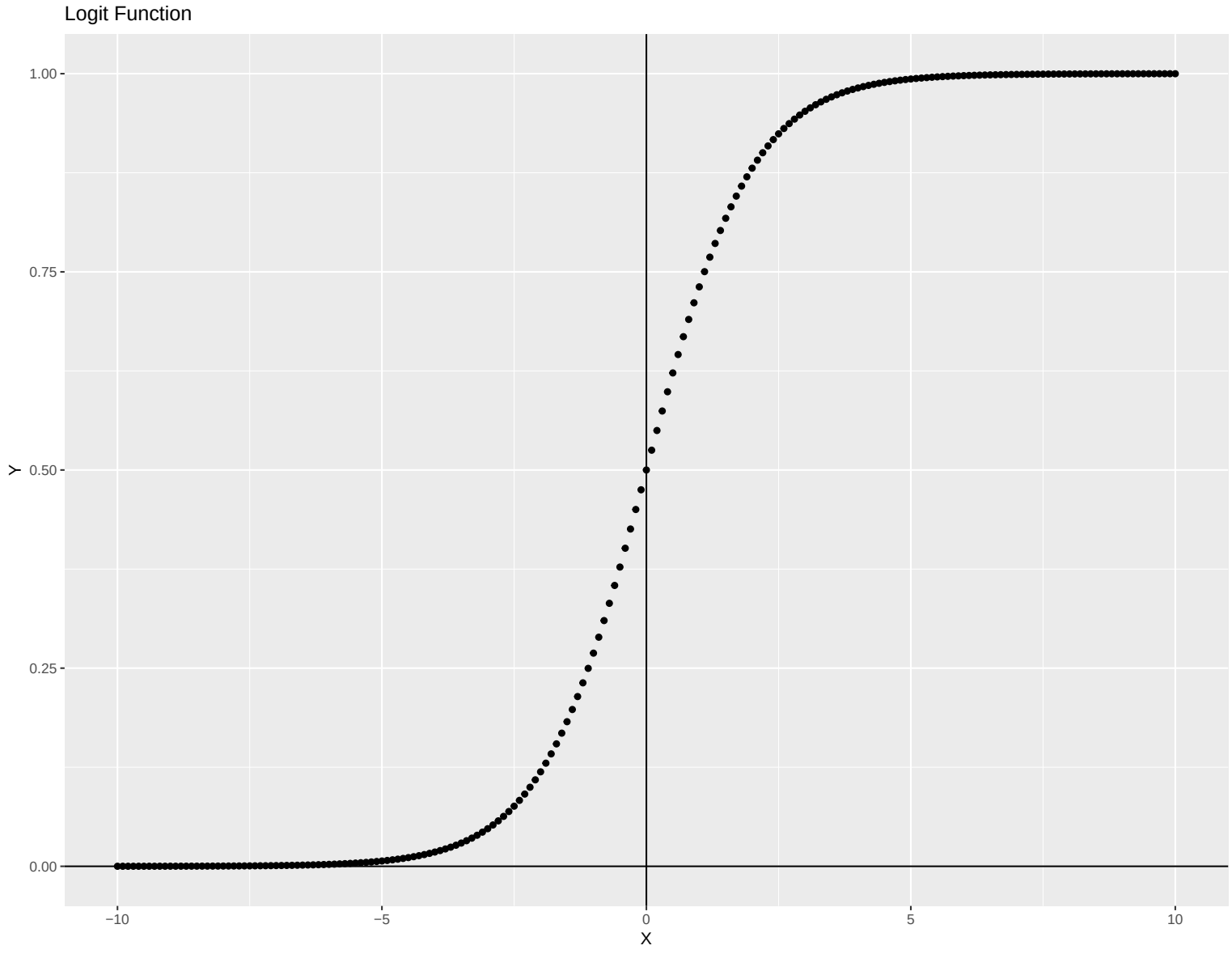
Lineer Sınıflandırıcılar, Lojistik Regresyonlar

- Lineer yöntemler arasındaki iki önemli başlıktan birisidir.
- Lojistik regresyonlar lineer regresyon modeli üzerine kuruludur ve non-lineer uygulamaları yaygın değildir.

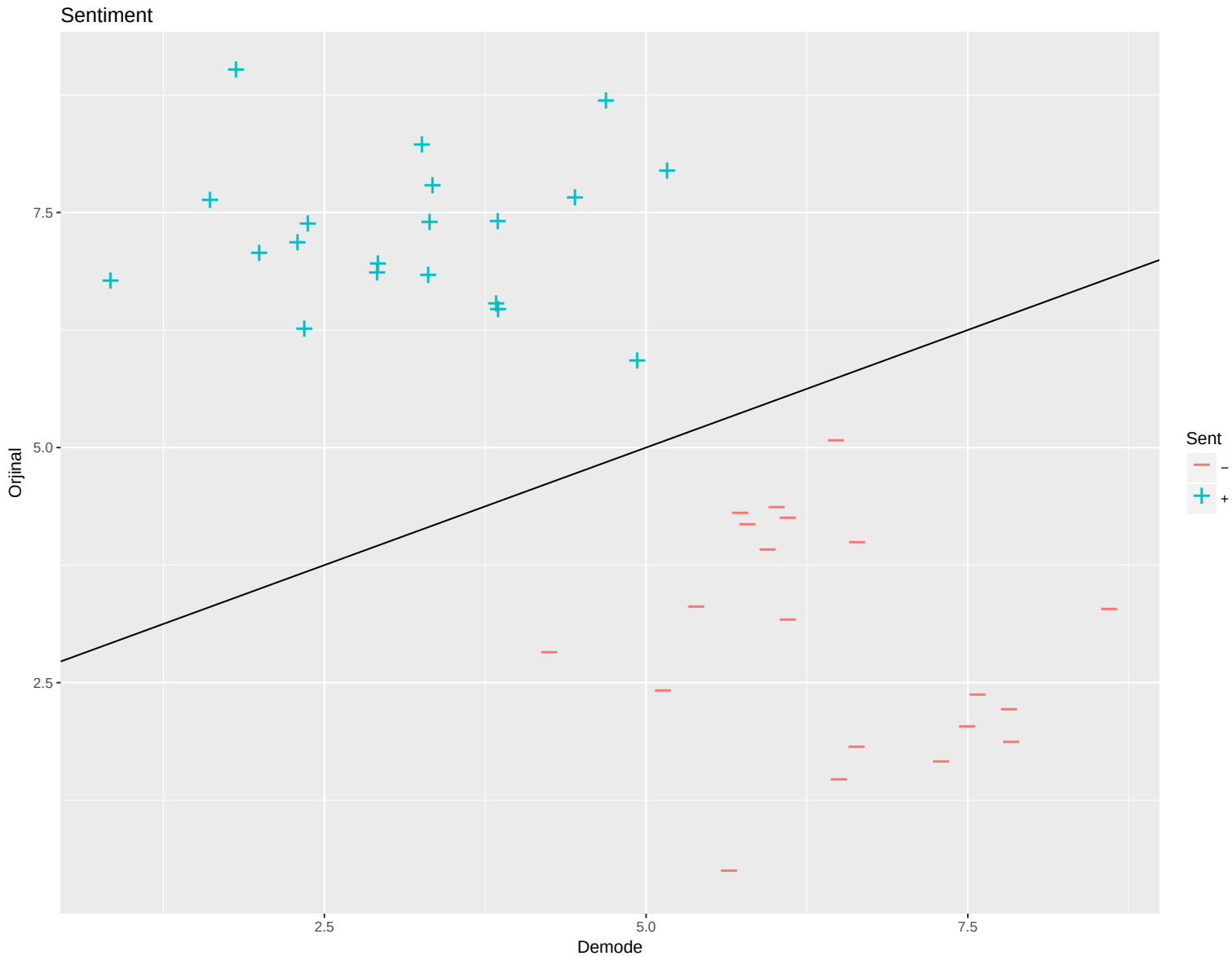
Model denklemi aşağıdaki gibidir:

$$Y_i = \phi(\beta_0 X_{0i} + \beta_1 X_{1i} + \cdots + \beta_n X_{ni}) \tag{1}$$

- Bu denkleme göre X_i'deki herhangi değere karşılık Y_i her zaman 0 ve 1 arası değer üretecektir.
- Bu yüzden Y_i bir olasılık olarak yorumlanır.
- Öte yandan X_i de *score* olarak tanımlanır.


Logit Function
1.00
0.75
0.50
0.25
0.00
Y
−10
−5
0
5
10
X

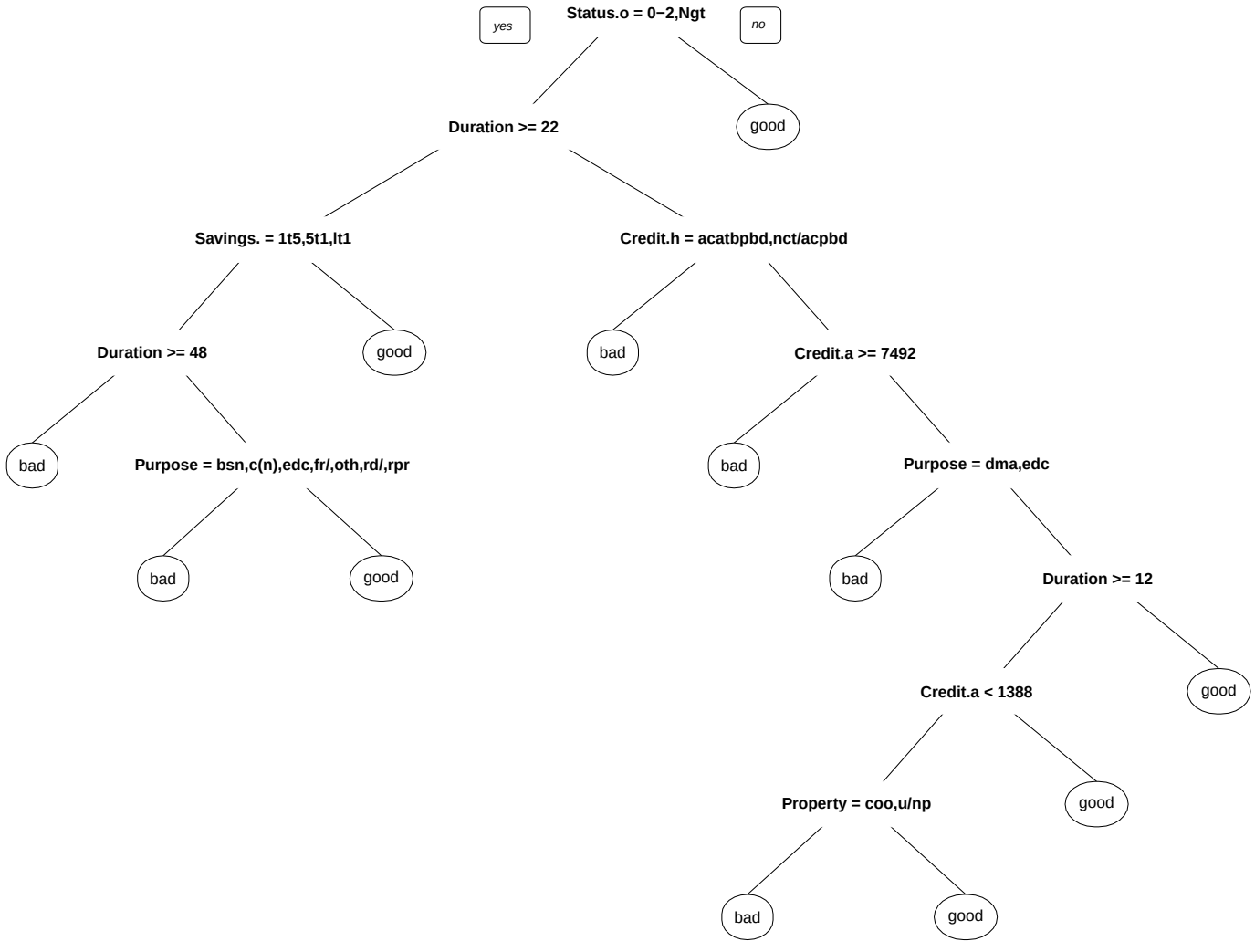
Sentiment Örneği



CART Yöntemleri

- Karar ağaçları (Decision Trees) ve regresyon ağaçları, CART yöntemlerinin en temel iki metotudur.
- Karar ve regresyon ağaçlarının farkı birinin kategorik diğerinin sürekli veriler üzerinde çalışmasıdır.
- Karar ağaçları, lojistik regresyona göre genelde daha düşük ya da benzer performans gösterir.
- Ancak yorumlaması çok daha kolaydır.
- Öte yandan Random Forest ve Boosting gibi daha karmaşık modellerin temelini oluşturur.

Örnek Bir Karar Ağacı



CART Yöntemleri (devam)

- Karar ağaçları veriyi bölmek için GINI katsayısını kullanır:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2}$$

Burada $\hat{p}_{mk} \in \{0, 1\}$'dır ve denklem heterojen sınıflandırılmış gözlemleri ölçer.

- Benzer şekilde regresyon ağaçları da

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T| \tag{3}$$

denklemi üzerinden işler ve sonuçta çıkan her bir grubun RSS'ini minimize etmek üzerine kuruludur.

Karar Ağaçlarının Avantaj/Dezavantajları

+ Karar ağaçları kolay yorumlanabilir
+ Çok boyutlu olmasına karşın grafiğe dökülebilir.
+ Kategorik verilerde oldukça kolay kullanılır
- Ancak başarısı çok yüksek değildir.
- Örneklem kümesi değiştikçe çok farklı karar ağaçları çıkabilir. (Model varyansı)

Bagging

- Bagging, karar ağaçlarının yüksek varyansından kaynaklanan problemleri çözmeyi hedefler.
- Bunun için:
  - Örneklem kümesi genişliği 2/3 ve 1/3 olmak üzere ikiye bölünür.
  - İlk küme bootsrapping ile defalarca çoğaltılır ve karar ağacı uygulanır.
  - Her bir karar ağacı kullanılarak 1/3 genişliğindeki ikinci küme tahmin edilir.
  - Bu tahminlerin ortalaması modelin başarısı olarak kabul edilir.

\+ Metot başarılı sonuçlar vermektedir

\- Ancak birçok karar ağacının ortalaması kullanıldığı için model yorumlanabilir değildir.

Random Forest

- Random Forest ağaç modelleri arasında en popülerlerden biridir.
- Bagging'in bir kümeye bağlı olmasından kaynaklı olarak veri seti bağımlılığını ortadan kaldırmayı hedefler.
- Random forest:
  - Bagging'te olduğu gibi bootsrapping kullanır.
  - Buna ek olarak her bir ayrım noktasında göze alınacak bağımsız değişken arasından rassal $m < n$ tanesini seçer.
  - Kalan $m - n$ bağımsız değişken olarak hesaba katmadan ağacı inşa eder.
  - Bu rassal seçim göz önünde bulunarak defalarca ağaç üretilir ve bunların ortalaması alınır.

Boosting

- Random Forest'e alternatif bir yöntemdir ve başarısı oldukça yüksektir.
- Karar ağacı ve regresyon ağacı yöntemlerinde yapılan hataları açıklamaya ve gerçekten beyaz gürültüye indirgeyene kadar modeli detaylandırmayı hedefler.
- Buna göre:
  - Veriye bir karar ağacı modeli uygulanır
  - Ağacın uç noktalarında çıkan veriye değil, onun hata ölçüsüne (MSE) regresyon ağacı uygular.
  - Bu süreç önceden belirlenmiş *depth* katsayısı kadar devam ettirilir.
  - Böylece hata paylarının verideki bağımsız değişkenlerin tam uygulanmamasından kaynaklanan hata ile beyaz gürültüden kaynaklanan hatayı ayıklar.

R Paketleri

- rpart, tree
- randomForest, gbm
- rattle, rpart.plot,RColorBrewer
- caret, party, partykit
- ROCR

Kaynakça I

James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert,
An introduction to statistical learning,
2013, Springer.