Örüntü Tanıma

Dr. Öğr. Üyesi Mehmet Zahid YILDIRIM

e-mail: m.zahidyildirim@karabuk.edu.tr

Özellik(Öznitelik) Oluşumu ve Özellik (Öznitelik) Seçimi

Özellik Oluşumu ve Özellik Seçimi

Özellik;

•Bir banka müşterisinin **churn tahm**ini için alınan veri setindeki müşterinin üyelik senesi, vade hesap sayısı, yaş, cinsiyet, meslek vb.

•Alışveriş sitesinde kullanıcılara **ürün öneri modeli** için ise gezilen ürün kategorileri, satın alma sıklığı vb. kolonlar olabilir.

Özellik seçimi , veri seti içerisinde en yararlı özellikleri seçme ve bulma sürecidir. Bu işlem Makine Öğrenmesi modelinin performansını çok fazla etkilemektedir.

Özellik Oluşumu ve Özellik Seçimi

Gereksiz öznitelikler;

- Modelinizin Training süresini arttırabilmektedir.

- Modelimizin basit ve açıklanabilir olmasını isteriz. Fazla sayıdaki öznitelik modelinizin yorumlanabilirliğini azaltabilmektedir.

- Model başarısının training veri setinde **overfitting** nedeniyle yüksek ancak test veri setinde ise düşük olmasına sebep olabilmektedir. Test veri setinde gelecek olan kayıtlar training veri setindeki kayıtlar ile benzerlik göstermediği durumda modelde hata oranı yüksek olacaktır.

Özellik Oluşumu ve Özellik Seçimi

Overfitting

Eğer modelimiz, eğitim için kullandığımız veri setimiz üzerinde gereğinden fazla çalışıp ezber yapmaya başlamışsa ya da eğitim setimiz tek düze ise **overfitting** olma riski büyük demektir.

Eğitim setinde yüksek bir skor aldığımız bu modele, test verimizi gösterdiğimizde muhtemelen çok düşük bir skor elde edeceğiz. Çünkü model eğitim setindeki durumları ezberlemiştir ve test veri setinde bu durumları aramaktadır. En ufak bir değişiklikte ezberlenen durumlar bulunamayacağı için test veri setinde çok kötü tahmin skorları elde edebilirsiniz.

Özellik Oluşumu ve Özellik Seçimi

Overfitting

Overfitting problemi aşağıdaki yöntemler uygulanarak çözülebilmektedir;

•**Öz nitelik sayısını azaltmak:** Birbirleriyle yüksek korelasyonlu olan kolonlar silinebilir ya da faktör analizi gibi yöntemlerle bu değişkenlerden tek bir değişken oluşturulabilir.

•**Daha fazla veri eklemek :** Eğer eğitim seti tek düze ise daha fazla veri ekleyerek veri çeşitliliği arttırılır.

•**Regularization (Düzenleme) :** Düzenleme, modelin karmaşıklığını azaltmak için bir kullanılan tekniktir. Yani modelde ağırlığı yüksek olan değişkenlerin ağırlığını azaltarak bu değişkenlerin etki oranını azaltır. Bu yöntem, aşırı öğrenme probleminin çözülmesine yardımcı olur.

Özellik Oluşumu ve Özellik Seçimi

Underfitting

Aşırı öğrenmenin aksine, bir model yetersiz öğrenmeye sahipse, modelin eğitim verilerine uymadığı ve bu nedenle verilerdeki trendleri kaçırdığı anlamına gelir.

Underfitting sorunu olan modellerde hem eğitim hem de test veri setinde hata oranı yüksektir.

Özellik Oluşumu ve Özellik Seçimi

Örüntü tanımada en önemli ve ilk adım mutlaka **veri temizleme** ve **özellik seçimi** olmalıdır.

Hem özellik oluşturma kısmı hem de o özelliklerden hangilerinin modelde kullanılması, aynı problemde farklı sonuçlar elde edilmesini sağlar.

Özellik Oluşumu ve Özellik Seçimi

En baştan doğru değişkenler oluşturmak ve seçmek bu işin en önemli kısmıdır.

Başarının ana kriteri doğru özellikleri bulup temizlenmiş şekilde kullanarak modeli kurmaktır.

5 elemanlı bir kümenin kaç tane altkümesi vardır?

Özellik Oluşumu ve Özellik Seçimi

Neden değişkenleri seçmeye ihtiyacımız var ?

- Çok fazla değişken kullanmak **modelin performansı**nı düşürebilir.
- Daha **kolay anlaşılır model** elde etmek için ihtiyacımız var.
- Daha **hızlı çalışan model** elde etmek için ihtiyacımız var.
- **Aşırı öğrenme riski**ni azaltır.

Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇMEYE NEREDEN BAŞLANMALI?

- Veri toplanmadan önce özellikler belirlenmelidir.
- Özellikleri doğru belirlemek için problemin çözümü noktasında bilgi sahibi olunmalı ve plan hazırlanmalıdır.

Örneğin; banka müşterisinin çıkış ihtimali yani churn durumu hesaplanmak istensin. Müşteriyi bu çıkışa götüren sebepleri ve adımları iyi bilinmesi gerekir. Ham veriden değişken oluşturacak şekilde bir adım atılarak işe başlanmalı.

Demografik özellikleri, hesap hareketleri, çağrı merkezi kayıtları değişkenler olabilir. Model performansına yani doğru tahmin etmeye yardım edecek her değişkeni kullanmak ya da kullanmamak önceden düşünülmeli

Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

1) Temel Temizlik

i) Kayıp Değerleri olan değişkeni çıkartmak

İlgili değişkendeki kayıp değerleri belli bir seviyenin üstünde ise, örneğin %80'i boş ise o değişkeni modele girdi olarak vermek anlamsız olur.

ii) Tüm değerleri aynı olan kategorik değişkeni çıkartmak

iii) Tüm değerleri farklı olan kategorik değişkeni çıkartmak

Eğer bir değişkende tüm gözlemler farklı ise model burada da bir bağıntı yakalayamaz.

iv) Düşük varyanslı değişkeni çıkartmak

Bütün değerlerin ortalamadan uzaklıklarının karelerinin ortalaması şeklinde bulunması varyanstır. Eğer ki bütün değerler aynı ise varyans = 0 olur ve o özelliği kullanmak anlamsız olur.

Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

2) Hedef Değişken ile Özelliklerin İlişkileri

Girdi değişkenlerinin hedef değişken ile olan ilişkisini istatiksel yöntemlerle tespit ederek en güçlü ilişkileri olan değişkenler seçilebilir.

Yani aslında özelliklerden sadece o özellik elimizde olsaydı bu özelliğin değişimi hedef değişkeni ne kadar değiştirir sorusunun cevabına bakılır.

*Değişkenler arasında ilişkileri incelerken **sayısal** ve **kategorik** değişkenler için farklı yöntemler uygularız.

Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

2) Hedef Değişken ile Özelliklerin İlişkileri

i) Pearson Korelasyonu

Korelasyon, iki değişkenin arasındaki doğrusal ilişkinin, değişiminin bir ölçüsüdür. Korelasyon katsayısı matematiksel olarak -1 ile +1 arasında değerler alır. Pearson Korelasyonunu kullanarak özellikler ile hedef değişken arasında bir ilişkiye bakılır. Sayısal değişkenler için kullanılır.

ii) Ki kare Testi

Ki kare testi kategorik değişkenler arasındaki ilişkiyi ölçmek için kullanılır. Ki-kare testini kullanarak, kategorik özellik ile hedef değişken arasında ilişkiyi ölçtükten sonra en iyi k kadar özellik seçilir

Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

2) Hedef Değişken ile Özelliklerin İlişkileri

iii) Anova Testi

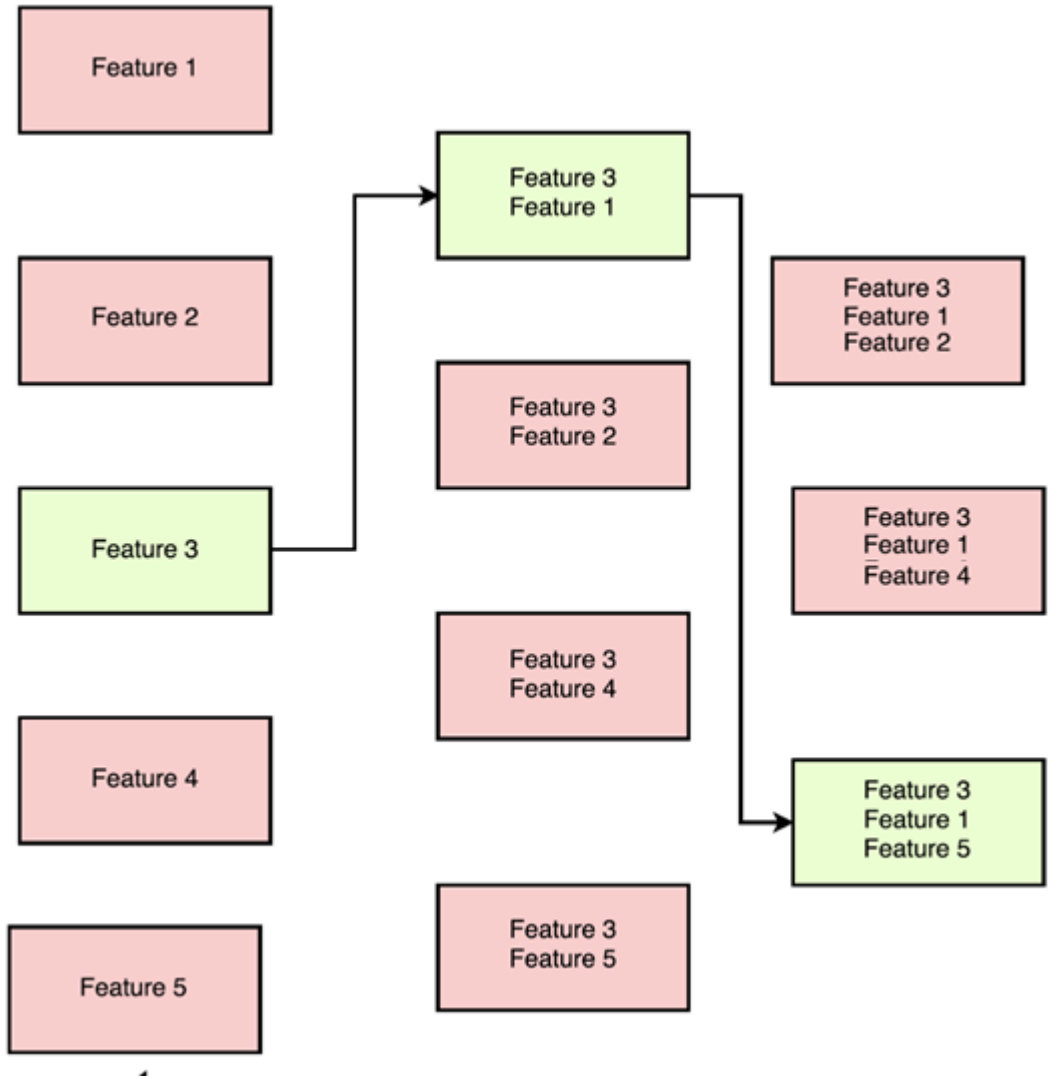
Anova testi, 3 ya da daha çok grup arasında, belirli bir değişkene dayalı olarak farklılık olup olmadığını belirlemek amacıyla kullanılır. Yani kategorik bir değişken ile sayısal bir değişken arasındaki ilişkiyi ölçmek için kullanırız.

Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

- ***Sarmalayıcı Yöntemler (Wrapper — Based)*** bir alt özellik grubuna sahip modeller oluşturur ve model performanslarını ölçer. Bunun için iki tip arama stratejisi vardır ;

Forward/İleri Arama: Boş bir öznitelik kümesi ile başlayarak her seferinde gruptaki öznitelikleri tek tek ekleyerek özniteliklerin kalitesi test edilir.

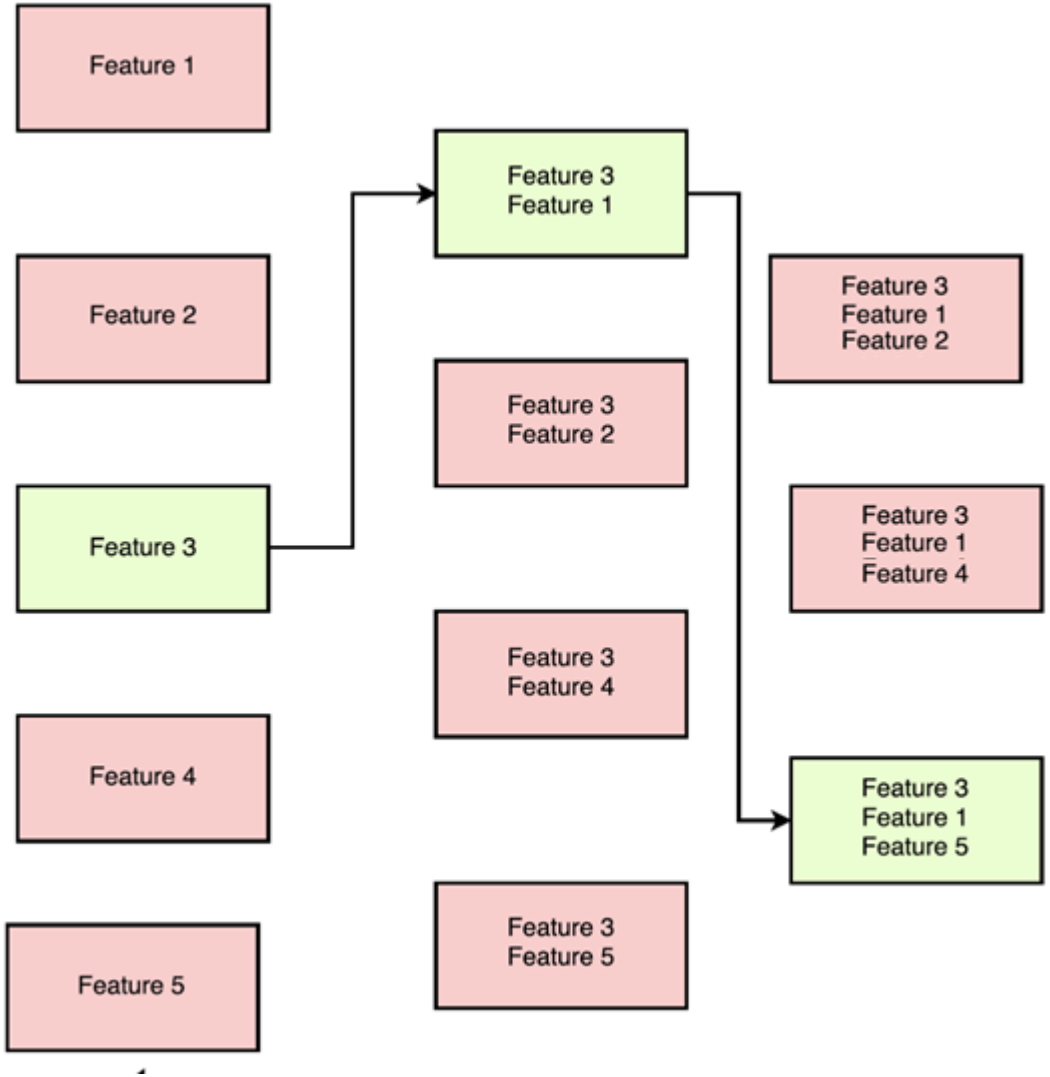


Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

n öznitelikli bir veri setinde;

- İlk adımda en iyi tahminleme yapan tek bir öznitelik seçilir.
- İkinci adımda ilk seçilen öznitelik ile beraber en iyi tahminleme yapan 2. öznitelik belirlenir . Böylece en iyi tahminleme yapan 2'li alt grup oluşturulmuş olur.
- Bu işlemler en iyi tahminleme yapan “m” adet öznitelik kombinasyonu bulunana kadar devam eder.



Özellik Oluşumu ve Özellik Seçimi

ÖZELLİK SEÇME ELEME YÖNTEMLERİ

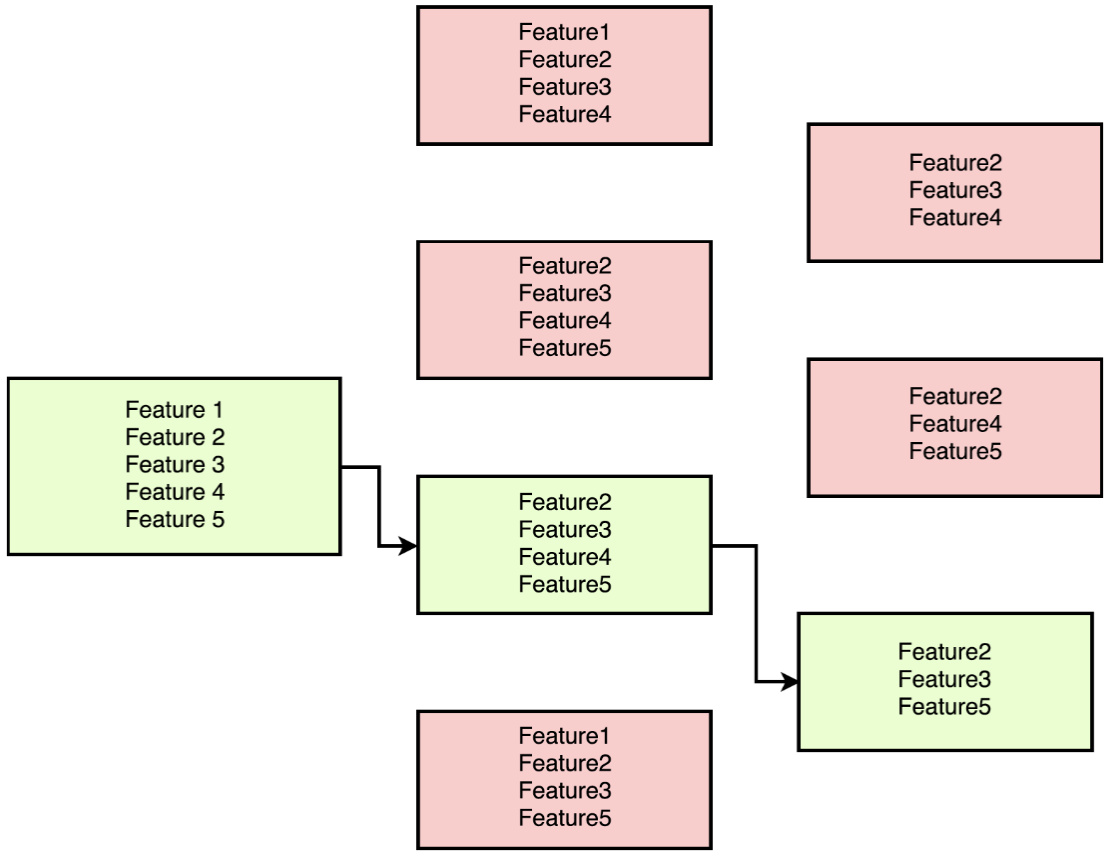
Backward/Geri Arama(Recursive Feature Elimination) : Adından da anlaşılacağı gibi, bu yöntem , en iyi öznitelik alt kümesi bulunana kadar tüm öznitelik kümesinden başlayarak adım adım en kötü performans gösteren öznitelikler elenir.

“n” adet öznitelikli bir veri setinde;

- Tüm veri setindeki öznitelikler alınır ve en az performans gösteren öznitelik kaldırılır.

- İkinci adımda ilk adımda belirlenen alt gruptaki yine en az performans gösteren öznitelik kaldırılır.

- Bu işlemler en iyi tahminleme yapan “m” adet öznitelik kombinasyonu bulunana kadar devam eder.



Kaynaklar

Sargur Srihari (CEDAR),
Jason Corso (SUNY at Buffalo),
Armando Vieira (Closer),
Luis Gustavo Martins (Catolica),
Selim Aksoy (Bilkent)