Veri Ön İşleme

by Sefa Isci

VBO

Veri Ön İşleme

- Veri Ön İşleme Genel Bakış
- Aykırı Gözlem Analizi
- Eksik Gözlem Analizi
- Standartlaştırma
- Değişken Dönüşümleri

Veri mi Model mi?

DATA

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

APRIL 02, 2018

SUMMARY SAVE SHARE COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

Veri Ön İşleme Genel Bakış

- Veri Temizleme (data cleaning / cleasing)
  - Gürültülü Veri
  - Eksik Veri Analizi
  - Aykırı Gözlem Analizi
- Veri Standardizasyonu
  - 0-1 Dönüşümü
  - z-skoruna Dönüştürme
  - Logaritmik Dönüşüm
- Veri İndirgeme
  - Gözlem Sayısının Azaltılması
  - Değişken Sayısının Azaltılması
- Değişken Dönüşümleri
  - Sürekli değişkenlerde dönüşümler
  - Kategorik değişkenlerde dönüşümler
- Değişken Mühendisliği

Gürültülü Veri

- **Veri Kaynağına Bağlı Hatalar**
  (anketler, veri tabanları, ara unsurlar)

- **Tutarsızlık**
  (Cinsiyet = Erken, Gebelik Durumu= 1,
  Kategori = Biberon, Fiyat = 900 BİN TL,
  Vasıta Türü = Otomobil, Motor Gücü = 25HP)

- **Kayıtlarda Çoklama**

Aykırı Gözlem Analizi

Veride genel eğilimin oldukça dışına çıkan ya da diğer gözlemlerden oldukça farklı olan gözlemlere aykırı gözlem denir.

Aykırı Değer Nedir?

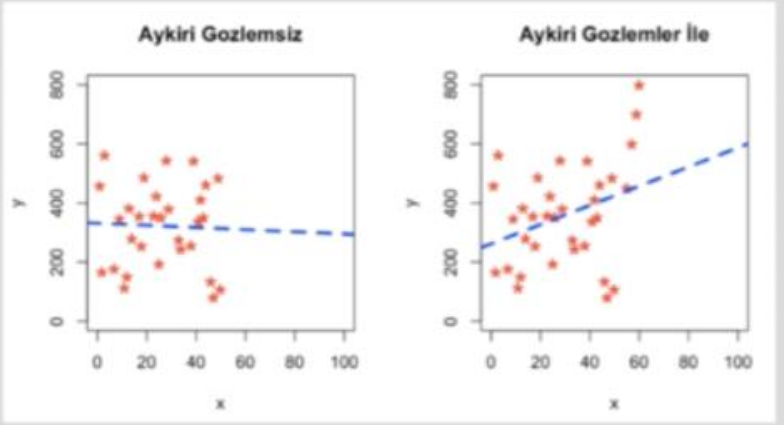
Aykırılığı ifade eden nümerik değere **aykırı değer** denir.

Aykırı Gözlem Nedir?

Aykırı değeri barındıran gözlem birimine **aykırı gözlem** denir.

Aykırı Değer Neye Sebep Olur?

Genellenebilirlik kaygısı ile oluşturulan **kural setlerini** ya da **fonksiyonları** yanıltır. **Yanlılığa** sebep olur.



Neye Göre Kime Göre Aykırı Gözlem?

«Veride genel eğilimin oldukça dışına çıkan gözlemler.»

Peki veri setinin genel eğiliminin dışına çıkmayı nasıl tanımlarız?

Neye Göre Kime Göre Aykırı Gözlem?

1. Sektör Bilgisi
Örneğin bir ev fiyat tahmin modelinde 1000 metrekarelik evleri modellemeye almamak.

2. Standart Sapma Yaklaşımı
Bir değişkenin ortalamasının üzerine aynı değişkenin standart sapması hesaplanarak eklenir. 1,2 ya da 3 standart sapma değeri ortalama üzerine eklenerek ortaya çıkan bu değer eşik değer olarak düşünülür ve bu değerden yukarıda ya da aşağıda olan değerler aykırı değer olarak tanımlanır.

3. Z-Skoru Yaklaşımı
Standart sapma yöntemine benzer şekilde çalışır. Değişken standart normal dağılıma uyarlanır, yani standartlaştırılır. Sonrasında -örneğin- dağılımın sağından ve solundan -+ 2,5 değerine göre bir eşik değer konulur ve bu değerin üzerinde ya da altında olan değerler aykırı değer olarak işaretlenir.

4. Boxplot(interquartile range - IQR) Yöntemi
En sık kullanılan yöntemlerden birisidir. Değişkenin değerleri küçükten büyüğe sıralanır. Çeyrekliklerine (yüzdekliklerine) yani Q1,Q3 değerlerine karşılık değeler üzerinden bir eşik değer hesaplanır ve bu eşik değere göre aykırı değer tanımı yapılır.

Neye Göre Kime Göre Aykırı Gözlem?

- **Tek Değişkenli**
  - Box-Plot
  - Histogram
  - Standart Sapma
  - Standart Normal Dağılım
- **Çok Değişkenli - İstatistiksel Yöntemler**
  - Kümeleme yöntemi
  - İkişerli saçılım grafiği ve kontur grafikleri (yüzde 90)
  - Kare Mahalanobis uzaklığı hesaplamak
  - Genelleştirilmiş varyans oranı
- **Çok Değişkenli - Diğer Yöntemler**
  - Derinlik Temelli Yaklaşımlar
  - Sapma Temelli Yaklaşımlar
  - Uzaklık Temelli Yaklaşımlar
  - Yoğunluk Temelli Yaklaşımlar
  - Yüksek Boyutlu Yaklaşımlar

File Edit View Run Kernel Tabs Settings Help

pydsml_dersler › veri_on_isleme

Name	Last Modified
veri_on_islem...	seconds ago
lof_intuition.p...	2 months ago

LOF Teorisi ve Yapay Veri Seti Oluşturma

veri_on_isleme.ipynb

Code

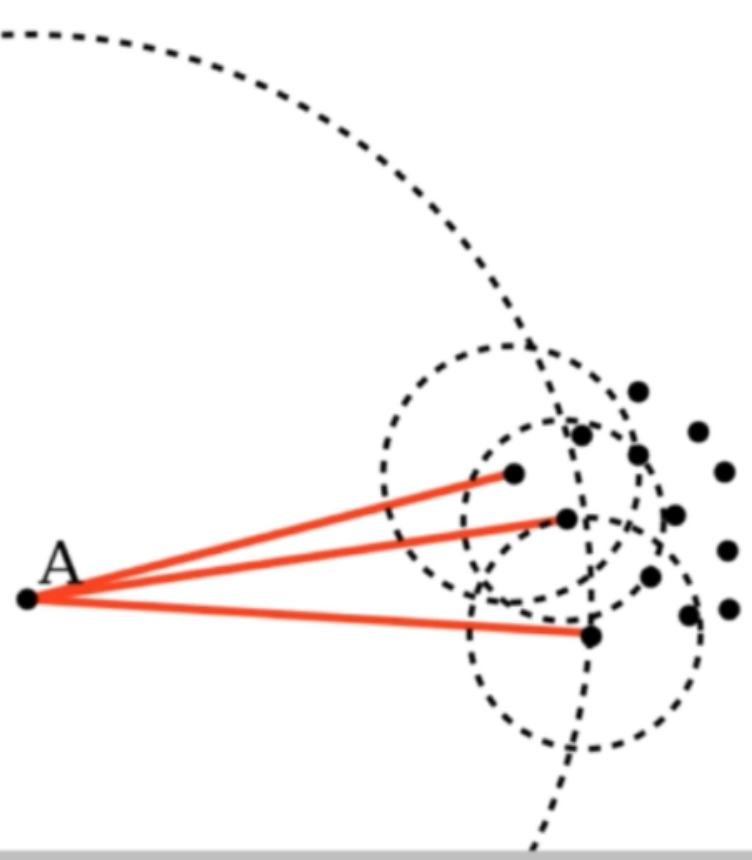
Python 3

Çok Değişkenli Aykırı Gözlem Analizi

Local Outlier Factor

[67]:
```
from IPython.display import Image
Image(filename =  "lof_intuition.png" , width=400, height=400)
```

[67]:



Eksik Veri Analizi

İncelenen veri setindeki gözlemlerde eksiklik olması durumunu ifade etmektedir.

Eksik Veri Adımları

- Eksik verinin belirlenmesi
- Yapısının görsel teknikler ile incelenmesi
- Eksikliğin rassallığının test edilmesi
- Uygun yöntemler ile doldurulması

Eksik Veriyi Direk Silmenin Zararları

Eksik değere sahip gözlemlerin veri setinden direk çıkarılması ve rassallığının incelenmemesi yapılacak istatistiksel çıkarımların, modelleme çalışmalarının güvenilirliğini düşürecektir. (Alpar, 2011)

Eksik Veriyi Direk Silmenin Zararları

Eksik gözlemlerin veri setinden direk çıkarılabilmesi için veri setindeki eksikliğin bazı durumlarda kısmen bazı durumlarda tamamen rastlantısal olarak oluşmuş olması gerekmektedir. Eğer eksiklikler değişkenler ile ilişkili olarak ortaya çıkan yapısal problemler ile meydana gelmiş ise bu durumda yapılacak silme işlemleri ciddi yanlılıklara sebep olabilecektir. (Tabachnick ve Fidell, 1996)

Eksik Veriyi Direk Silmenin Zararları

- 1. Veri setindeki eksikliğin yapısal bir eksilik olup olmadığının bilinmesi gerekir!
- 2. NA her zaman eksiklik anlamına gelmez!
- 3. Bilgi kaybı!

Eksik Veri Türleri Nelerdir?

- Tümüyle Raslantısal Kayıp: Diğer değişkenlerden ya da yapısal bir problemden kaynaklanmayan tamamen rastgele oluşan gözlemler.
- Raslantısal Kayıp: Diğer değişkenlere bağlı olarak oluşabilen eksiklik türü.
- Raslantısal Olmayan Kayıp: Göz ardı edilemeyecek olan ve yapısal problemler ile ortaya çıkan eksiklik türü.

Eksik Veri Rassallığının Testi

- Bağımsız iki örneklem t testi
- Korelasyon testi
- Little'nin MCAR testi

Eksik Veri Problemi Nasıl Giderilir?

"The idea of imputation is both seductive and dangerous" (R.J.A Little & D.B. Rubin)

Eksik Veri Problemi Nasıl Giderilir?

- **Silme Yöntemleri**
  - Gözlem ya da değişken silme yöntemi
  - Liste bazında silme yöntemi (Listwise Method)
  - Çiftler bazında silme yöntemi (Pairwise Method)
- **Değer Atama Yöntemleri**
  - Ortanca, ortalama, medyan
  - En Benzer Birime Atama (hot deck)
  - Dış Kaynaklı Atama
- **Tahmine Dayalı Yöntemler**
  - Makine Öğrenmesi
  - EM
  - Çoklu Atama Yöntemi