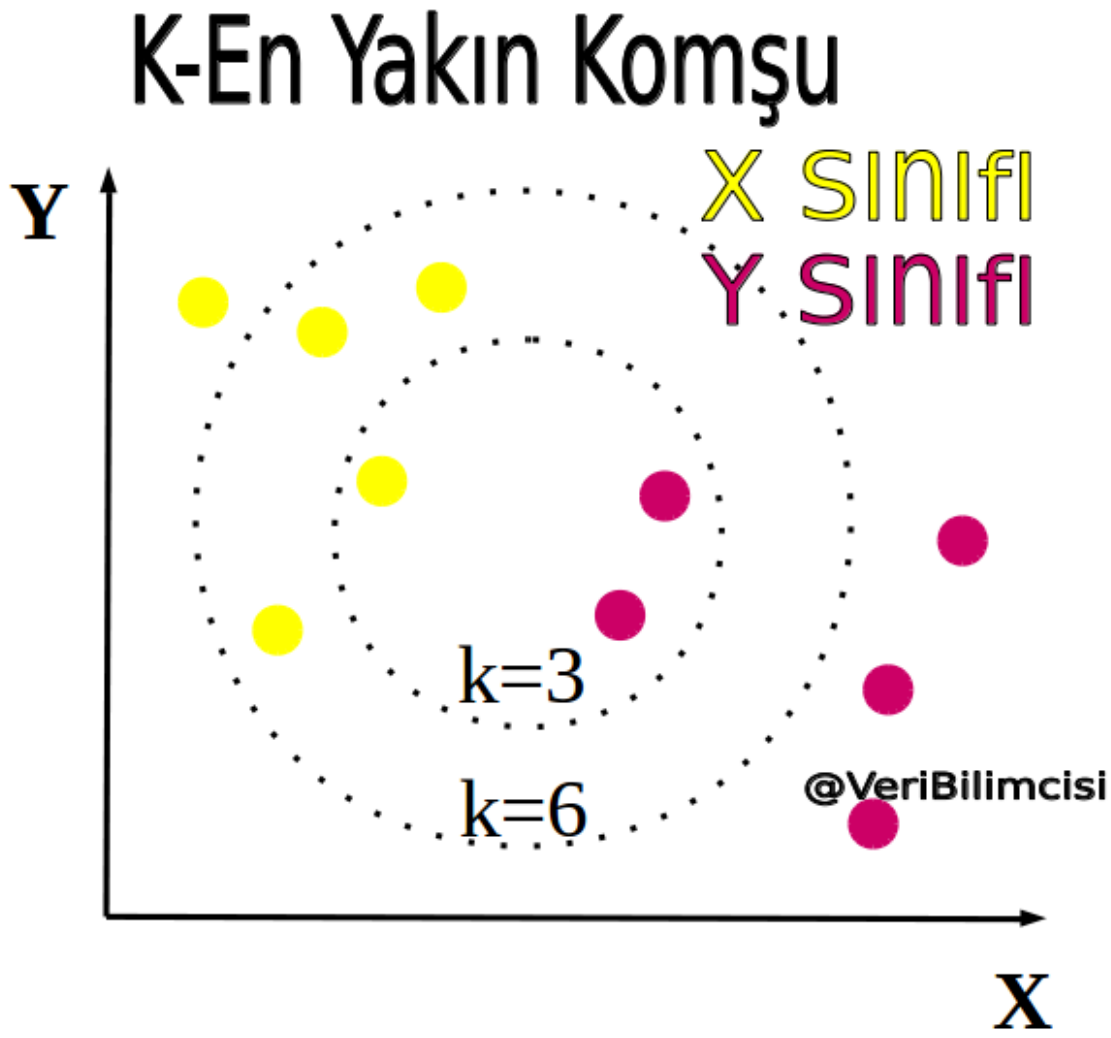
K-En Yakın Komşu (K-Nearest Neighbors(KNN))

veribilimcisi.com/2017/07/20/k-en-yakin-komsu-k-nearest-neighborsknn

20 Temmuz 2017



KNN, Denetimli Öğrenmede sınıflandırma ve regresyon için kullanılan algoritmalardan biridir. En basit makine öğrenmesi algoritması olarak kabul edilir.

Diğer Denetimli Öğrenme algoritmalarının aksine, eğitim aşamasına sahip değildir. Eğitim ve test hemen hemen aynı şeydir. Tembel bir öğrenme türüdür. Bu nedenle, kNN, geniş veri setini işlemek için gereken algoritma olarak ideal bir aday değildir.

KNN ile temelde yeni noktaya en yakın noktalar aranır. K, bilinmeyen noktanın en yakın komşularının miktarını temsil eder. Sonuçları tahmin etmek için algoritmanın k miktarını (genellikle bir tek sayı) seçeriz.

Model tanımada , en yakın komşu algoritması (kNN), sınıflandırma ve regresyon için kullanılan parametrik olmayan bir yöntemdir. Her iki durumda da, girdi, özellik alanında k en yakın eğitim örneklerinden oluşur. Çıktı, kNN'nin sınıflandırma veya regresyon için kullanılıp kullanılmayacağına bağlıdır:

- K-NN sınıflandırmasında , çıktı sınıf üyeliğidir. Bir nesne, komşularının çoğunluk oyuyla sınıflandırılır; nesne, en yakın komşuları arasında en yaygın olan sınıfa verilir (k , küçük bir pozitif bir tam sayı). Eğer k = 1 ise, nesne basitçe o en yakın komşunun sınıfına atanır.
- K-NN regresyonda çıktı, cismin özellik değeridir. Bu değer, en yakın komşularının değerlerinin ortalamasıdır.

K -NN, örüntü tabanlı öğrenme veya tembel öğrenme türüdür; burada işlev sadece yerel olarak yaklaştırılır ve tüm hesaplama, sınıflandırmaya kadar ertelenir. K- N algoritması, tüm makine öğrenmesi algoritmalarının en basitleri arasındadır.

Hem sınıflandırma hem de regresyon için, komşuların katkılarına ağırlık koymak, böylece yakın komşuların ortalamaya daha uzak olanlardan daha fazla katkıda bulunmaları yararlı olabilir. Örneğin, ortak bir ağırlıklandırma şeması, her komşuya 1 / d ağırlığının verilmesini içerir; burada d komşuya olan uzaklıktır.

Komşular, sınıfın (kNN sınıflaması için) veya nesne mülk değerinin (kNN regresyonu için) bilindiği bir takım nesnelerden alınır. Bu, algoritma için ayarlanmış eğitim olarak düşünülebilir, ancak açık bir eğitim basamağı gerekmemektedir.

KNN algoritmasının bir özelliği, verilerin yerel yapısına duyarlı olmasıdır. **Algoritma, başka popüler bir makine öğrenme tekniği olan k- means ile karıştırılmamalıdır.**

K komşuları, tüm mevcut vakaları depolayan ve bir benzerlik ölçüsüne (ör. Mesafe fonksiyonları) dayalı yeni vakaları sınıflandıran basit bir algoritmadır. KNN, 1970'lerin başında halihazırda parametrik olmayan bir teknik olarak istatistiksel tahmin ve örüntü tanımada kullanılmıştır.

Algoritma

Bir örnek, komşularının çoğunluk oyuyla sınıflandırılır; bu olay, bir mesafe fonksiyonuyla ölçülen en yakın komşuları arasında en yakın olan sınıfa atanır. K = 1 ise, örnek yalnızca en yakın komşusunun sınıfına atanır.

Ayrıca, üç mesafenin de yalnızca sürekli değişkenler için geçerli olduğuna dikkat edilmelidir. Kategorik değişkenler söz konusu olduğunda, Hamming mesafesi kullanılmalıdır. Ayrıca, veri kümesinde sayısal ve kategorik değişkenlerin bir karışımı olduğunda 0 ile 1 arasındaki sayısal değişkenlerin standardizasyonu meselesini ortaya çıkarmaktadır.

K için en uygun değeri seçmek için önce verileri incelemek gerekir. Genel olarak, büyük bir K değeri, genel gürültüyü düşürdüğü için daha hassasdır, ancak garantisi yoktur. Çapraz doğrulama, K değerini doğrulamak için bağımsız bir veri kümesi kullanarak, geriye dönük olarak iyi bir K değerini belirlemenin başka bir yoludur. Tarihsel olarak, çoğu veri kümesi için optimal K, 3-10 arasında olmuştur. Bu, 1NN'den çok daha iyi sonuçlar üretir.

Örnek:
Kredi ile ilgili aşağıdaki verileri göz önünde bulundurun. Yaş(age) ve kredi(loan) iki sayısal değişken(tahmini) ve borç ödememek(default) de hedeftir(sonuç – yanıt).

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

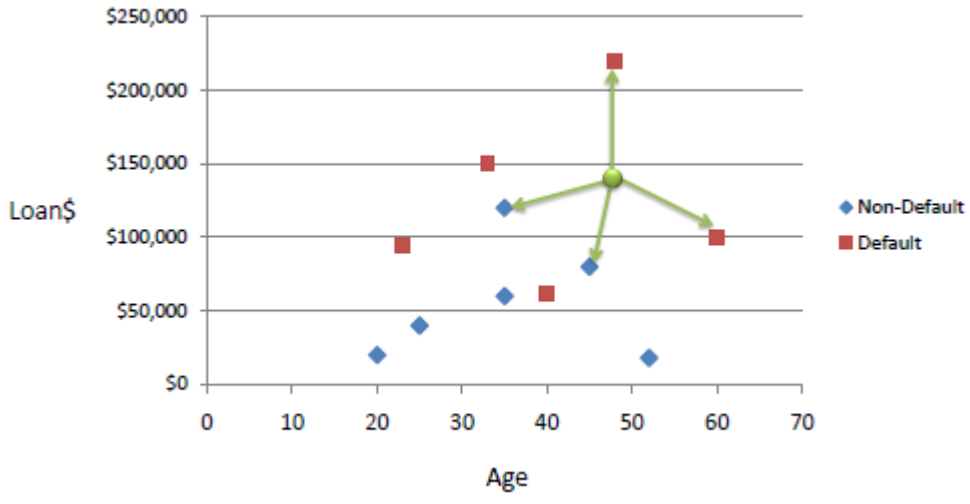
Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1



Artık eğitim setini, Öklid uzaklığını kullanarak bilinmeyen bir durumu (Yaş = 48 ve Kredi = 142.000 ABD Doları) sınıflandırmak için kullanabiliriz. K = 1 ise, en yakın komşu, Varsayılan = Y ile eğitim setindeki son durumdur.

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Standart Mesafe

Doğrudan eğitim setinden mesafe ölçümlerinin hesaplanmasında bir büyük dezavantaj, değişkenlerin farklı ölçüm ölçeklerine sahip olması ya da sayısal ve kategorik değişkenlerin bir karışımı olması durumunda ortaya çıkmaktadır. Örneğin, bir değişken dolar cinsinden yıllık geliri temel alırken diğeri yıllara dayanıyorsa, gelirin hesaplanan mesafe üzerinde çok daha fazla etkisi olacaktır. Bir çözüm, aşağıda gösterildiği gibi eğitim setini standartlaştırmaktır.

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

Aynı eğitim setindeki standart mesafeyi kullanarak, bilinmeyen verilerde şimdi farklı bir komşu bulmuş oldu.

Standartlaştırma Hakkında Daha Fazla Bilgi İçin:

Özellik Ölçekleme ve Normalleştirme (Feature Scaling and Normalization)

Referans:

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

http://www.saedsayad.com/k_nearest_neighbors.htm