T.C

SAKARYA ÜNİVERSİTESİ
BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BSM 310 – YAPAY ZEKA

Grup üyeleri:
B171210004 YUSUF SİNA YILDIZ
B171210052 MUHAMMED ALİ BALCI
B171210112 AHMET SEFA BOZDEMİR
B171210028 BERKAY KÜÇÜK
B171210106 MERT YAVUZ

Sakarya
2020

K – En Yakın Komşu Algoritması

Giriş

Yapay zekâ, makinelerin karmaşık problemlere insanlar gibi çözümler üretmesini sağlayan bir bilim dalıdır. Bu genellikle insan zekâsının karakteristik özelliklerini alıp, bilgisayarlara algoritmalar biçiminde uygulanarak gerçekleştirilir. Yapay zekâ, genellikle bilgisayar bilimlerinden başka Matematik, Biyoloji, Psikoloji, Felsefe ve diğer farklı bilimler ile de yakından ilgilidir. Tüm bu alanlardaki bilgilerin birleştirilmesi ile eninde sonunda yapay zekâ konusundaki gelişmelere bağlı olacaktır. Yapay zekâ konusu üzerine hiçbirinin doğruluğu ya da yanlışlığı kanıtlanmamış birçok farklı yaklaşım bulunmaktadır.

Makine Öğrenmesi (Machine Learning) ise birçok matematiksel ve istatistiksel yöntemleri kullanarak mevcut verilerden çıkarımlar yapabilen, bu yaptığı çıkarımlarla bilinmeyen tahminlerde bulunan yöntem biçimidir. Makine öğrenmesine sosyal hayatımızdan bazı örnekler verebiliriz: Yüz tanıma, Veri sınıflandırma, Spam tespitleri. KNN algoritması her iki alanda da kullanılmaktadır.

KNN algoritmaları, 1967 yılında T. M. Cover ve P. E. Hart tarafından önerilmiştir. K-NN *(K-Nearest Neighbor)* algoritması en basit ve en çok kullanılan sınıflandırma algoritmasından biridir. K-NN **non-parametric (**parametrik olmayan**)**, **lazy** (tembel) bir öğrenme algoritmasıdır. *lazy* kavramını anlamaya çalışırsak *eager learning* aksine lazy learning'in bir eğitim aşaması yoktur. Eğitim verilerini öğrenmez, bunun yerine eğitim veri kümesini *"ezberler"*. Bir tahmin yapmak istediğimizde, tüm veri setinde en yakın komşuları arar.

Aynı zamanda K-en yakın komşuluk (KNN) algoritması, uygulaması kolay gözetimli öğrenme algoritmalarındandır. Bazı durumlarda çok iyi sonuçlar verebilir ancak maliyet bakımından büyük veriler üzerinde işlemleri süre bazında uzun sürede gerçekleştirdiği için daha çok küçük çaplı öğrenme işlemlerinde tercih edilir. Hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılıyor olmakla birlikte, endüstride çoğunlukla sınıflandırma problemlerinin çözümünde kullanılmaktadır.

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.

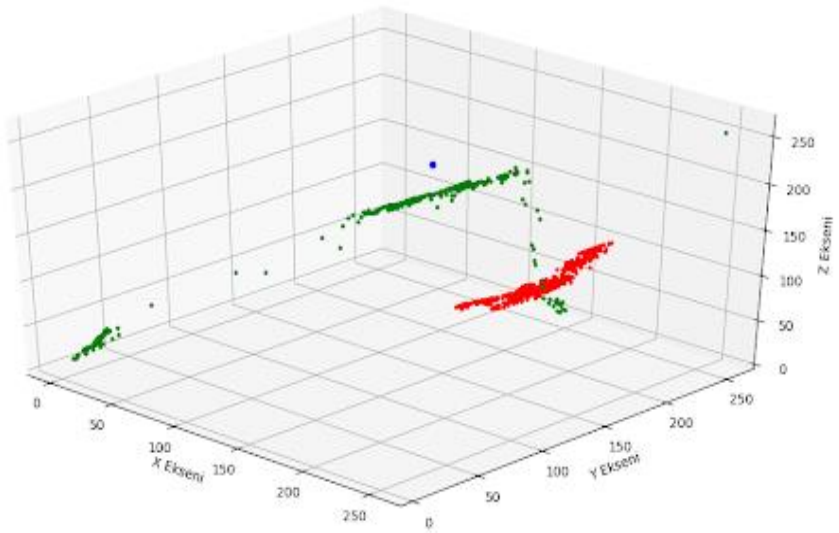
- **K-NN regresyonda** çıktı, cismin özellik değeridir. Bu değer, en yakın komşularının değerlerinin ortalamasıdır.

K-NN, örüntü tabanlı öğrenme türüdür; burada işlev sadece yerel olarak yaklaştırılır ve tüm hesaplama, sınıflandırmaya kadar ertelenir.

Hem sınıflandırma hem de regresyon için, komşuların katkılarına ağırlık koymak, böylece yakın komşuların ortalamaya daha uzak olanlardan daha fazla katkıda bulunmaları yararlı olabilir.

KNN algoritması; (sınıf niteliği belli olan) elemanların meydana getirdiği uzaya yeni bir örnek (sınıf niteliği belli olmayan) eklendiğinde bu örneğin kendisine en yakın olan sınıfa dahil edilmesi gerektiğini kararlaştıran bir algoritmadır. Kendisine en yakın olan sınıfı belirlemek için bir k değişkeni kullanmaktadır. Belirlenen bu k değişkeni örneğe en yakın olan K adet (sınıf nitelikleri belli olan) elemanların sayısını temsil etmektedir.

Örnek verecek olursak uygulamamızın 3 boyutlu uzaydaki görünümü aşağıdaki gibidir. Mavi nokta örneğimiz (sınıfı belli olmayan), kırmızı noktalar cilt(sima) sınıfını, yeşil noktalar ise cilt olmayan sınıfı temsil etmektedir. Grafikte açık bir şekilde göreceğiniz üzere; Örneğimiz, uygulamamız tarafından kendisine en yakın olan sınıfa dahil edilmiştir.



Kırmızı(1)İlCiltİ : 0 -- Yesil(2)İlCilt Değil) : 17 -- ((SONUC : 2. Sınıf(Yesil)))

KNN algoritmasının bir özelliği, verilerin yerel yapısına duyarlı olmasıdır. Algoritma, başka popüler bir makine öğrenme tekniği olan k-means ile karıştırılmamalıdır. Aralarındaki en büyük fark KNN algoritması bir eğitim verisi içerirken k-means algoritması bir eğitim verisi içermez. Yeni bir değer geldiğinde K değerine mesafeler hesaplanır ve yeni değer bir kümeye ilave edilir.

Çalışma Prensibi

1. Veriyi yükle
2. K sayısını belirle ve oluştur.
3. Tüm verileri foreach ile gez.
4. Son eklenen veri ile seçilen verinin mesafesini ölç.
5. Ölçülen verileri mesafeler ile bir diziye ata.
6. Diziyi mesafeye göre küçükten büyüğe sırala.
7. Sıralanmış diziden ilk K adet veriyi seç.
8. Seçilen K adet verinin etiketlerini al.
9. Son eklenen veriyi sayıca üstün olan etiket ile etiketle.

Uzaklık Ölçütleri

1. Minkowski Uzaklığı

Minkowski uzaklığı, Öklid uzayında tanımlı bir dizidir. Sınıflandırma, kümeleme gibi makine öğrenmesi, veri madenciliği uygulamalarında sıklıkla kullanılan Öklid uzaklığı, Manhattan uzaklığı gibi uzaklık ölçütlerinin genelleştirilmiş halidir. Herhangi iki nokta P ve Q arasındaki Minkowski uzaklığı P = (x1, x2, …, xn) ve Q = (y1, y2, …, yn) olmak üzere, aşağıya göre hesaplanır:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

Minkowski uzaklığı, genel bir formül ile ifade edilmekte olup p'nin farklı değerleri için çeşitli uzaklık ölçütlerini tanımlamak amacıyla da kullanılmaktadır. Minkowski ölçütünün p=2 olduğu özel durumu, Öklid uzaklığını, p=1 olduğu özel durumu Manhattan uzaklığını ve $n \to \infty$ olduğu özel durum, Chebyschev uzaklığını vermektedir.

2. Öklid Uzaklığı

Öklid uzaklığı, sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür. Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklık olup herhangi iki nokta, P ve Q arasındaki Öklid uzaklığı P = (x1, x2, …, xn) ve Q = (y1, y2, …, yn) olmak üzere, aşağıya göre hesaplanır:

$$\left(\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \right)$$

Öklid uzaklığı, K-ortalama kümeleme algoritması, temel K-NN algoritması gibi sınıflandırma ve kümeleme algoritmalarında yakınlığın ölçümlenmesi için kullanılan temel uzaklık ölçütüdür.

3. Manhattan Uzaklığı

Manhattan uzaklığı, n boyutlu iki nokta arasındaki farkların mutlak değerlerinin toplamıdır. Herhangi iki nokta, P ve Q arasındaki Manhattan uzaklığı P = (x1, x2, …, xn) ve Q = (y1, y2, …, yn) olmak üzere aşağıya göre hesaplanır:

$$\left(\sum_{i=1}^{n} |x_i - y_i| \right)$$

4. Chebyschev Uzaklığı

Chebyschev uzaklığı (maksimum değer uzaklığı), Minkowski uzaklığının, n→∞ olduğu özel durum olup, iki nokta arasındaki farkların mutlak değerlerinin maksimumu olarak tanımlanmaktadır. Herhangi iki nokta, P ve Q arasındaki Chebyschev uzaklığı P = (x1, x2, …, xn) ve Q = (y1, y2, …, yn) olmak üzere, aşağıya göre hesaplanır.

$$\lim_{p\to\infty}\left(\sum_{i=1}^{n}\left|x_i-y_i\right|^p\right)^{1/p}=\max_{i=1}^{n}\left|x_i-y_i\right|$$

5. Dilca Uzaklığı

Dilca (Distance Learning in Categorical Attribute) uzaklığı, kategorik öznitelik değerleri arasındaki uzaklığı ölçümlemek için kullanılan iki aşamalı bir ölçüttür. Bu ölçütte öncelikle, simetrik belirsizlik katsayısı yöntemi kullanılarak öznitelik seçimi işlemi gerçekleştirilerek eş-oluşum tablosu oluşturulmaktadır. Ardından, eş-oluşum tablosu üzerinde koşullu olasılık ve Öklid uzaklığına dayalı hesaplama gerçekleştirilerek uzaklık ölçümlenmektedir. Bilgi kazancı, yüksek değer içeren özniteliklere karşı taraflıdır. Simetrik belirsizlik katsayısı (symmetrical uncertainty) (SU), bilgi kazancının (information gain) (IG) bu problemini ortadan kaldırmak için, bilgi kazancının X ve Y özniteliklerinin entropi değerleri toplamına bölünmesi ile belirler. Simetrik belirsizlik katsayısı aşağıdaki formüle göre hesaplanır:

$$SU=2\times\left[\frac{IG}{H(Y)+H(X)}\right]$$

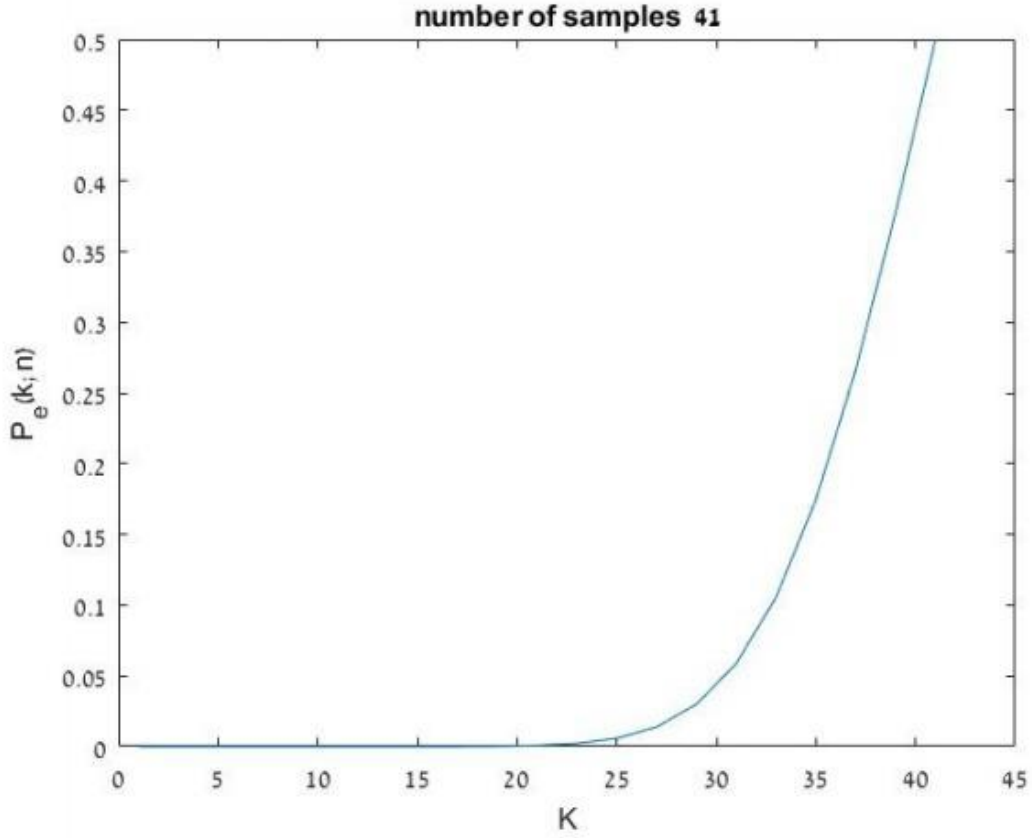
Öznitelik seçiminin ardından, uzaklık hesaplaması aşağıda belirtilen formüle göre gerçekleştirilir:

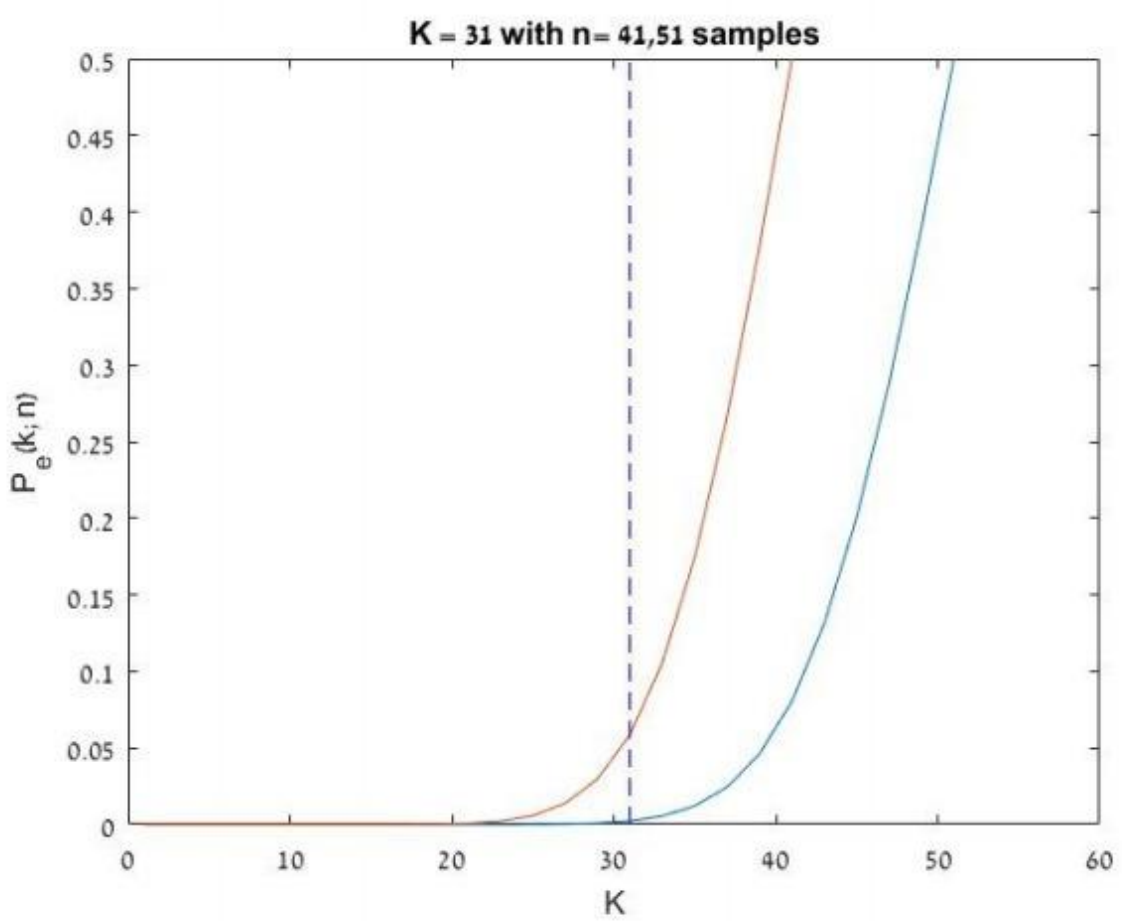
$$d(x_i,x_j)=\sqrt{\sum_{Y\in baglam(X)}\sum_{y_k\in Y}\left(P(x_i|y_k)-P(x_j|y_k)\right)^2}$$

Burada, xi ve xk incelenmekte olan özniteliğin aldığı değer çiftleri ve xi, xj∈X dir. Her bir Y bağlam özniteliği için xi ve xk değerlerine dayalı koşullu olasılık hesaplanıp ardından Öklid uzaklığı alınmaktadır. Her bir öznitelik için bağlam öznitelikleri, simetrik belirsizlik katsayısına dayalı olarak belirli bir sezgisel değerlendirme ölçütü aracılığıyla yapılmaktadır.

K Değerinin Algoritmaya Etkisi

En yakın komşu algoritmasını (1-NN) temel alarak sınıflandırma yapmak yerine çok sayıda örneğimiz varsa, k 'nın komşularının çoğunluğuna bağlı olarak sınıflandırma yapabiliriz. Eğer biz K'yi çok büyük alırsak ne gibi sorun oluşabilir? Ya da eğer biz K'yi çok küçük alırsak ne gibi sorun oluşabilir? Yüksek değerlerde k yi seçmek, mesafe faktörünün etkisini azaltacaktır. Eğer biz K'yi örnek sayısına eşit alırsak, yeni gelen her örnek etiketli örneklerin sayısına bağlı olarak etiketlenmiş olacaktır. Çok küçük bir K seçmek saçma değerlerin etiket sisteminin etkilemesine izin verecektir.





Ağırlıklandırma

Komşular için ağırlık değerleri atanması ile sınıflandırılmakta olan örneğe daha yakın olan komşu örneklerin, çoğunluk oylamasına daha fazla katkı koyması amaçlanır. En çok kullanılan ağırlık değeri atama yöntemleri, her bir komşunun ağırlığının, d, komşular arası uzaklık olmak üzere, 1/d ya da 1/d 2 şeklinde alınmasıdır.

Veri Setleri

Bu çalışmada, K-NN algoritmasının farklı parametre türlerine ve değerlerine ilişkin sınıflandırma performansının incelenmesi için, makine öğrenmesi alanında ve sınıflandırma problemlerinde yaygın olarak kullanılan altı farklı veri setinden yararlanılmıştır. Bu kapsamda, UCI Machine Learning Repository'de yer alan, "Breast Cancer Wisconsin", "Cardiotocography", "Ionosphere", "Leaf", "Parkinsons" ve "Thoracic Surgery" veri setleri kullanılmıştır [20]. Veri setlerine ilişkin temel özellikler, Tablo 1'de özetlenmiştir.

Tablo 1: Veri setlerine ilişkin temel özellikler

Veri Seti	Örnek Sayısı	Öznitelik Sayısı	Sınıf Sayısı
Breast Cancer Wisconsin	699	10	2
Cardiotocography	2126	23	3
Ionosphere	351	34	2
Leaf	340	16	30
Parkinsons	197	23	2
Thoracic Surgery	470	17	2

KNN'nin Avantajları

1. Eğitim Süresi Yok: KNN'ye Tembel Öğrenen (Örnek tabanlı öğrenme) denir. Eğitim döneminde hiçbir şey öğrenmez. Eğitim verilerinden herhangi bir ayrımcı işlev türetmez. Başka bir deyişle, bunun için bir eğitim süresi yoktur. Eğitim veri kümesini saklar ve ondan yalnızca gerçek zamanlı tahminler yaparken öğrenir. Bu, KNN algoritmasını SVM, Doğrusal Regresyon vb. gibi eğitim gerektiren diğer algoritmalardan çok daha hızlı hale getirir.

2. Güncellenebilirlik: KNN algoritması tahminler yapmadan önce eğitim gerektirmediğinden, algoritmanın doğruluğunu etkilemeyecek yeni veriler sorunsuz bir şekilde eklenebilir.

3.Kolaylık: Algoritmanın en önemli avantajlarından birisi uygulamada ve teoride kolay bir algoritma olmasıdır. Algoritmanın gerçeklenebilmesi için sadece iki parametre yeterli olacaktır.
Bu parametreler K değeri ve mesafe fonksiyonlarıdır. (örneğin Öklid veya Manhattan uzaklıkları gibi).

4.Uyarlanabilirlik: Algoritmanın önemli özelliklerinden birisi ise uygulamalara uyarlanabilir olmasıdır.

5.İzlenebilirlik: Algoritma çoğu algoritmalardan izlenebilirlik ve takip edilebilirlik açısından ayrılmaktadır. Veri setinin ve örneklerin seçim kriterleri göz önünde bulundurulduğunda izlenebilirlik açısından başarılı algoritmalara dahildir.

Dezavantajları

- Algoritmanın performansının istenilen seviyede sonuçlar vermesi için gerekli parametrelerin ve örneklerin doğru sayıda alınması gerekmektedir. Buda algoritmayı kullananlara ek bir yük getirir.

- Uzaklık bazlı öğrenme algoritması, en iyi sonuçları elde etmek için, hangi uzaklık tipinin ve hangi niteliğin kullanılacağı konusunda açık değildir. Hangi uzaklık ölçütünün ne zaman doğru işleyeceği belli değildir.Bu da algoritmanın doğruluk oranını değiştiren kriterlerden birisidir.

- Hesaplama maliyeti gerçekten çok yüksektir çünkü her bir sorgu örneğinin tüm eğitim örneklerine olan uzaklığını hesaplamak gerekmektedir. Bazı indeksleme metodları ile (örneğin K-D ağacı), bu maliyet azaltılabilir.

- En yakın komşuluk prensibine dayanır. Tüm dokümanlar vektörel olarak temsil edilir. Sorgu dokümanı ile diğer dokümanlar arasındaki cosinüs benzerliği hesaplanır. Similarty oranı 1'e en yakın olan n tane vektörün kategorisinden çok olanı dokümana atanır.

Algoritmayla İlgili Sorular

1) Aşağıdakilerden hangisi KNN algoritmasının avantajları arasında yer almaz?

A) Eğitim süresinin olmaması
B) Gürültülü verilere karşı dayanıklı olması
C) Yerel bilgilere uyarlanabilir olması
D) Analitik olarak izlenebilir olması
E) Düşük bellek alanına gerek duyması

2) Herhangi bir X konumunda bulunan bir cisim 5 birim aşağı ve 12 birim sağa hareket ettirilirse,
öklit uzaklığına göre toplam kaç birim hareket etmiş olur?
A) 8
B) 10
C) 15
D) 13
E) 17

3) KNN gibi sınıflandırma ve kümeleme algoritmalarında aşağıdaki uzaklık ölçme algoritmalarından hangisi en çok kulllanır?
A) Minkowski Uzaklığı
B) Öklit Uzaklığı
C) Manhattan Uzaklığı
D) ChebyschevUzaklığı
E) Dilca Uzaklığı

4) Toplamda 7 örnek olarak seçilen bir KNN algoritması örneğinde k= 3 olarak belirlenmiştir.Etiketlenecek olan verinin diğer sınıf elemanlarına olan uzaklığı aşağıdaki gibidir. Buna göre yeni veri hangi sınıfa dahil olmalıdır?

E1 = 4 birim (A sınıfı)
E2= 8 birim (B sınıfı)
E3= 3 birim (C sınıfı)
E4= 13 birim (A sınıfı)
E5= 7 birim (D sınıfı)
E6= 7 birim (B sınıfı)
E7= 4.9 birim (A sınıfı)

A) a
B) b
C) c
D) d
E) Hiçbiri

5) Aşağıda KNN algoritmasının çalışma adımları yer almaktadır. Adımların sırayla işlediğini düşündüğümüzde hangi adımda hata yapılmıştır?

1. K değeri belirlenir.
2. Diğer nesnelerden hedef nesneye olan uzaklıklar hesaplanır.
3. En yakın komşu kategorileri toplanır.
4. Uzaklıklar sıralanır ve en minimum uzaklığa bağlı olarak en yakın komşular bulunur.
5. En uygun komşu kategorisi seçilir.

A) 1
B) 2
C) 3
D) 4
E) 5

Kaynaklar:

- T. M. Cover and P. E. Hart, Nearest Neighbor Pattern Classification, May 15, 2018
- Naresh Kumar, Advantages and Disadvantages of KNN Algorithm in Machine Learning
- Xu, G., Zong, Y. and Yang, Z., “Applied Data Mining”, CRC Press, New York, (2013).
- Duda, R.O., Hart, P.E. and Stork, D.G., “Pattern Classification”, John Wiley & Sons, New Jersey
- Han, J. and Kamber, M., “Data mining: concepts and techniques”, Morgan Kaufmann Publishers, Burlington, (2006).

- Doad, P.K. and Bartere, M.M., “A Review: Study of Various Clustering Techniques”, International Journal of Engineering Research & Technology,

- Erdal Taşçı, Aytuğ Onan, K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi