Yapay Zeka ve Uygulamaları

Prof. Dr. Hamdi Tolga KAHRAMAN

Eğiticili ve Eğiticisiz Sınıflandırma

- Eğiticili (supervised) sınıflandırma: Sınıflandırma

  Sınıf sayısı ve bir grup örneğin hangi sınıfa ait olduğunu bilinir

- Eğiticisiz (unsupervised) sınıflandırma: Kümeleme (Demetleme, öbekleme,…) Hangi nesnenin hangi sınıfa ait olduğu ve grup sayısı belirsizdir.

Sınıflama Tanımı

Sınıflamanın temel kuralları:

- Öğrenme eğiticilidir
- Veri setinde bulunan her örneğin bir dizi niteliği vardır ve bu niteliklerden biri de sınıf bilgisidir.
- Hangi sınıfa ait olduğu bilinen nesneler (öğrenme kümesi- training set) ile bir model oluşturulur
- Oluşturulan model öğrenme kümesinde yer almayan nesneler (deneme kümesi- test set) ile denenerek başarısı ölçülür

Örnek Verikümesi

Örnekler (intances, samples)

Table 3.1 • **The Credit Card Promotion Database**

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19

Örnek Verikümesi

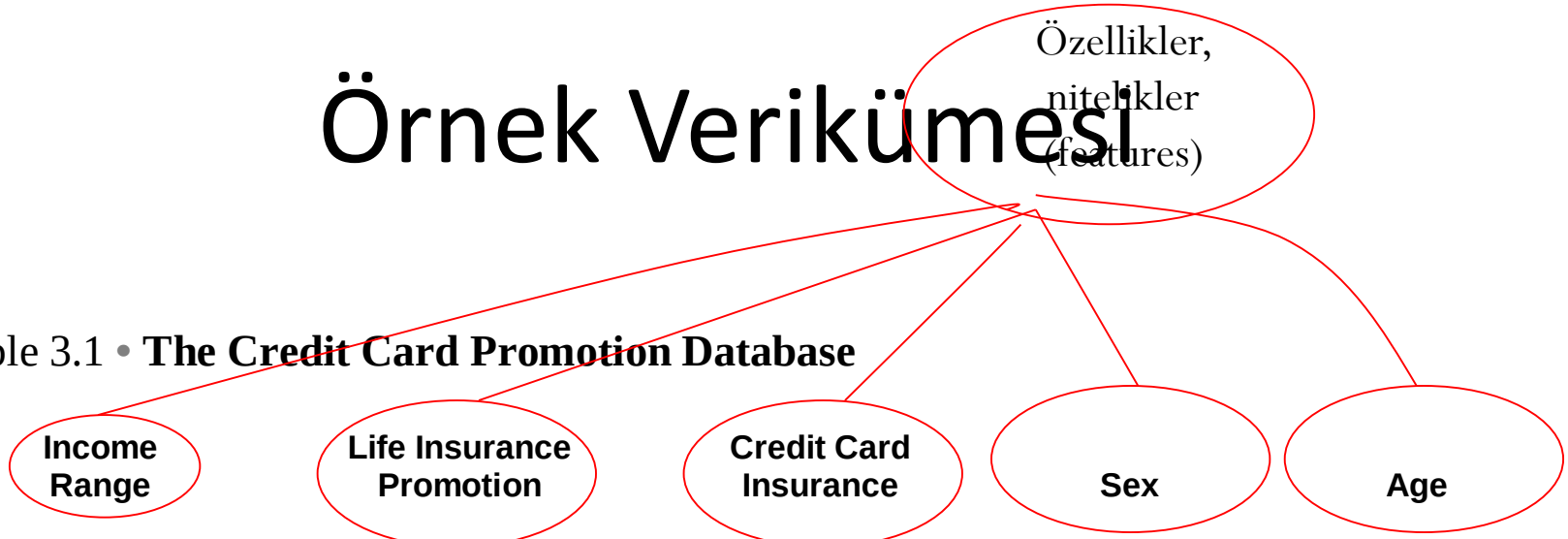


Table 3.1 • **The Credit Card Promotion Database**

Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	No	No	Male	45
30–40K	Yes	No	Female	40
40–50K	No	No	Male	42
30–40K	Yes	Yes	Male	43
50–60K	Yes	No	Female	38
20–30K	No	No	Female	55
30–40K	Yes	Yes	Male	35
20–30K	No	No	Male	27
30–40K	No	No	Male	43
30–40K	Yes	No	Female	41
40–50K	Yes	No	Female	43
20–30K	Yes	No	Male	29
50–60K	Yes	No	Female	39
40–50K	No	No	Male	55
20–30K	Yes	Yes	Female	19

Sınıflandırma Yöntemleri:

- Karar Ağaçları (Decision Trees)
- Örnek Tabanlı Yöntemler: k en-yakın komşu (Instance Based Methods- k nearest neighbor)
- Bayes Sınıflandırıcı (Bayes Classifier)
- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algorithms)

Uzaklığa dayalı sınıflandırma

k-nearest neighbor (k-nn) sınıflayıcı

- Eğitim adımı yok
- Test verileri en yakınlarındaki k adet komşularının sınıf etiketlerine bakılarak sınıflandırılır
- Sınıflandırma yapılırken eldeki verilerin birbirlerine olan uzaklığı veya benzerliği kullanılarak sınıflamanın gerçekleştirildiği bir tekniktir. Veriler (sınıf örnekleri) arasındaki mesafe ölçülürken en çok kullanılan yöntemler Euclidean, Manhattan ve Minkovski uzaklık metrikleridir.

- Mesafeye/uzaklığa dayalı algoritmalardan en bilineni ve yaygın kullanılanı da k-en yakın komşu sınıflandırıcıdır.

Uzaklığa dayalı sınıflandırma

k-nearest neighbor (k-nn) sınıflayıcı

- En yakın komşu algoritmasına “tek bağlantı kümeleme yöntemi” adı da verilmektedir. En yakın komşu algoritmasında öncelikle gözlemler arasındaki uzaklıklar belirlenir. ive j gözlemleri arasındaki uzaklıkların belirlenmesinde herhangi bir uzaklık bağıntısı kullanılabilir.

- k en yakın komşuluk algoritması sorgu vektörünün en yakın k komşuluktaki vektör ile sınıflandırılmasının bir sonucu olan denetlemeli öğrenme algoritmasıdır.

- Sınıflandırma yapılırken veri tabanındaki her bir kaydın diğer kayıtlara olan uzaklığı hesaplanır.

Uzaklığa dayalı sınıflandırma

Oklit uzaklığı

- iki nokta arasındaki uzaklığı hesaplayan bağıntılara gereksinim vardır. Uygulamada en çok kullanılan uzaklık bağıntısı Öklid uzaklık bağıntısıdır. p değişken sayısı olmak üzere, i,j=1,2,...,n ve k=1,2,...,p için Öklid uzaklığı şu şekilde olacaktır:

$$d(i,j)=\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2}$$

Uzaklığa dayalı sınıflandırma

Manhattan uzaklığı

- Diğer bir uzaklık bağıntısı ise Manhattan uzaklığıdır. Bu uzaklık, gözlemler arasındaki mutlak uzaklıkların toplamı alınarak hesaplanır. Söz konusu uzaklık şu şekilde ifade edilir

$$d(i,j) = \sum_{k=1}^{p} \left(\left| x_{ik} - x_{jk} \right| \right)$$

Uzaklığa dayalı sınıflandırma

Minkovski uzaklığı

- Gözlemler arasındaki uzaklıkların hesaplanmasında Minkowski uzaklık bağıntısı da kullanılabilmektedir.

$$d(i,j) = \left[\sum_{k=1}^{p} \left(\left| x_{ik} - x_{jk} \right|^{m} \right) \right]^{\frac{1}{m}}$$

Uzaklığa dayalı sınıflandırma

k-nearest neighbor (k-nn) sınıflayıcı

- Uygulamada k değeri önceden seçilir; değerinin yüksek olması birbirine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyse birbirine benzediği, yani aynı sınıfın noktaları oldukları halde, bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur. Tipik k-değerleri 3, 5, ve 7 dir.

k-nearest neighbor (k-nn) sınıflayıcı

Avantajları

- Uygulanabilirliği basit bir algoritmadır.
- Gürültülü eğitim örneklerine karşı dirençlidir.
- Eğitim örnekleri sayısı fazla ise etkilidir.
- model yaratmaz
  - Bu sebeple eğitim için bir zaman harcamaz. Ama bu durum ay nı zamanda bir dezavantaj yaratmaktadır.
  - Çünkü algoritma, sınama için diğer modellerin aksine daha çok zamana ihtiyaç duyar. Örnek olarak 4000 boyutlu eğitim kümesi düş ünelim. Eğer k-nn algoritması uygulanırsa bir tek sınama örneği için 4000 satırın herbiriyle karşılaştırma yapılır.

k-nearest neighbor (k-nn) sınıflayıcı

Dezavantajları

- k parametresine ihtiyaç duyar.
- Uzaklık bazlı öğrenme algoritması, en iyi sonuçları elde etmek için, hangi uzaklık tipinin ve hangi niteliğin kullanılacağı konusunda açık değildir.
- Hesaplama maliyeti gerçekten çok yüksektir çünkü her bir sorgu örneğinin tüm eğitim örneklerine olan uzaklığını hesaplamak gerekmektedir. Bazı indeksleme metotları ile (örneğin K-D ağacı), bu maliyet azaltılabilir.
- En yakın komşuluk prensibine dayanır. Tüm örnekler vektörel olarak temsil edilir. Sorgu örneği ile diğer örekler arasındaki cosinüs benzerliği hesaplanır.
- Benzerlik oranı 1'e en yakın olan n tane vektörün kategorisinden çok olanı sorgulanan örneğin sınıfı olarak atanır.

k-nearest neighbor (k-nn) sınıflayıcı

Algoritmanın adımları

i. k parametresi belirlenir.
   - Bu parametre verilen bir noktaya (sınıfı aranan sorgu örneği) en yakın komşuların sayısıdır.

ii. Sorgu örneği ile diğer tüm örnekler arasındaki uzaklıklar tek tek hesaplanır.

iii. Yukarıda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.

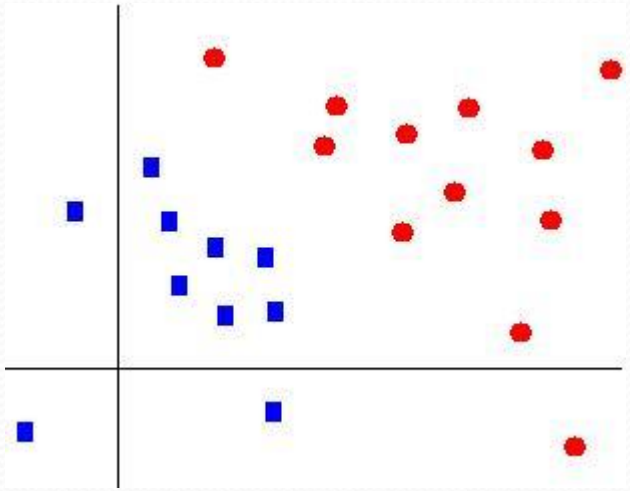
iv. Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrar eden kategori değeri seçilir.

v. Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

k-nearest neighbor (k-nn) sınıflayıcı

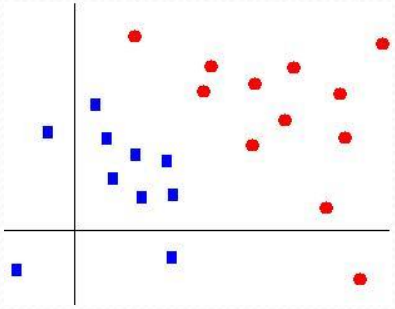
Örnek

i. Örneğin k = 3 için yeni bir eleman sınıflandırılmak istensin. bu durumda eski sınıflandırılmış elemanlardan en yakın 3 tanesi alınır. Bu elamanlar hangi sınıfa dahilse, yeni eleman da o sınıfa dahil edilir. Mesafe hesabından genelde öklit mesafesi (euclid distance) kullanılabilir.

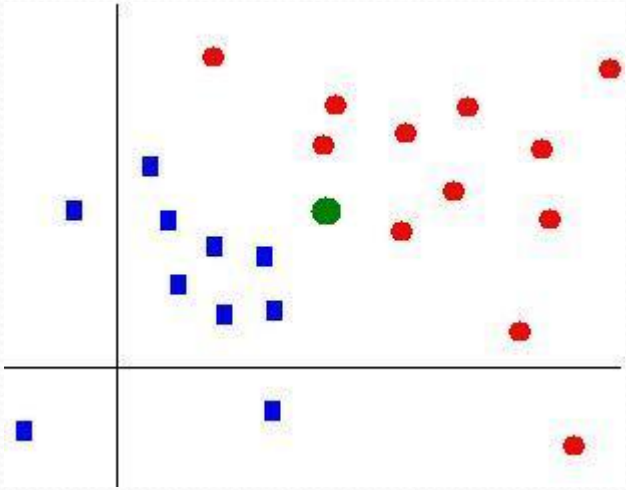


k-nearest neighbor (k-nn) sınıflayıcı

Örnek

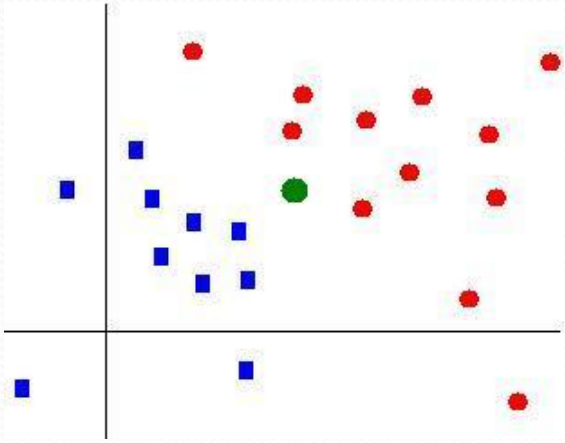


i. KNN yöntemine göre aşağıdaki şekilde yeni bir üyenin geldiğini düşünelim:

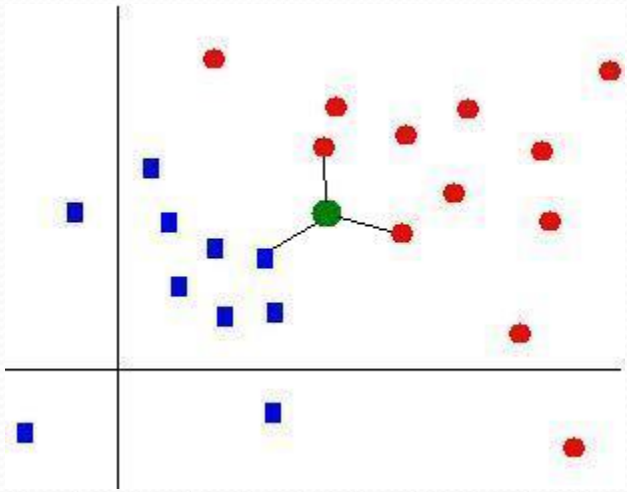


k-nearest neighbor (k-nn) sınıflayıcı

Örnek

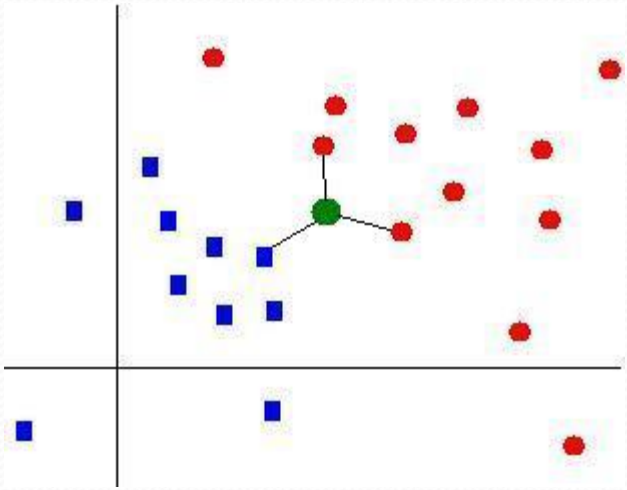


Yukarıdaki bu yeni gelen üyenin en yakın olduğu 3 üyeyi (3 nearest neighbors) tespit edelim.

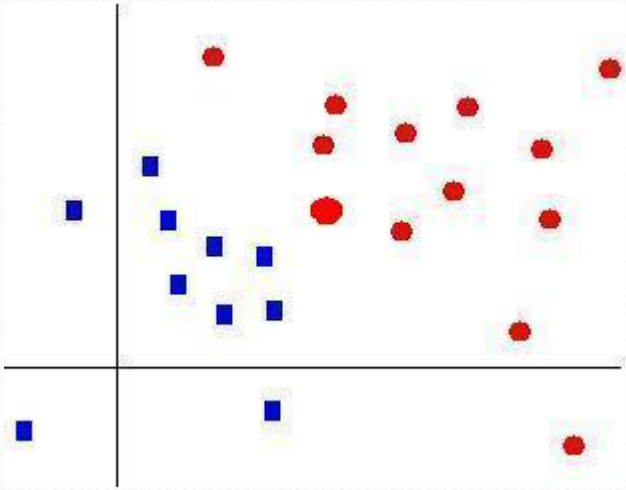


k-nearest neighbor (k-nn) sınıflayıcı

Örnek



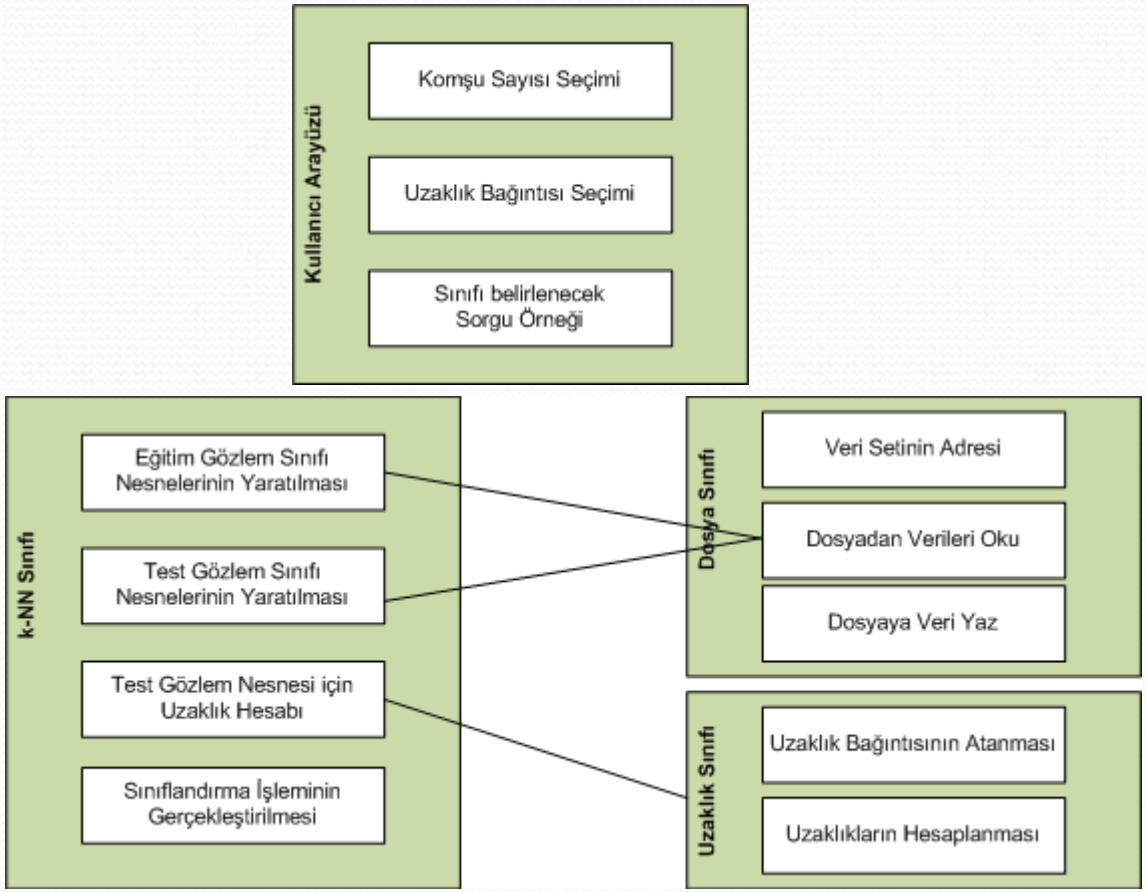
En yakın 3 üyenin iki tanesi kırmızı yuvarlak üyeler olduğuna göre yeni üyemizi bu şekilde sınıflandırabiliriz:



k-nearest neighbor (k-nn) sınıflayıcı

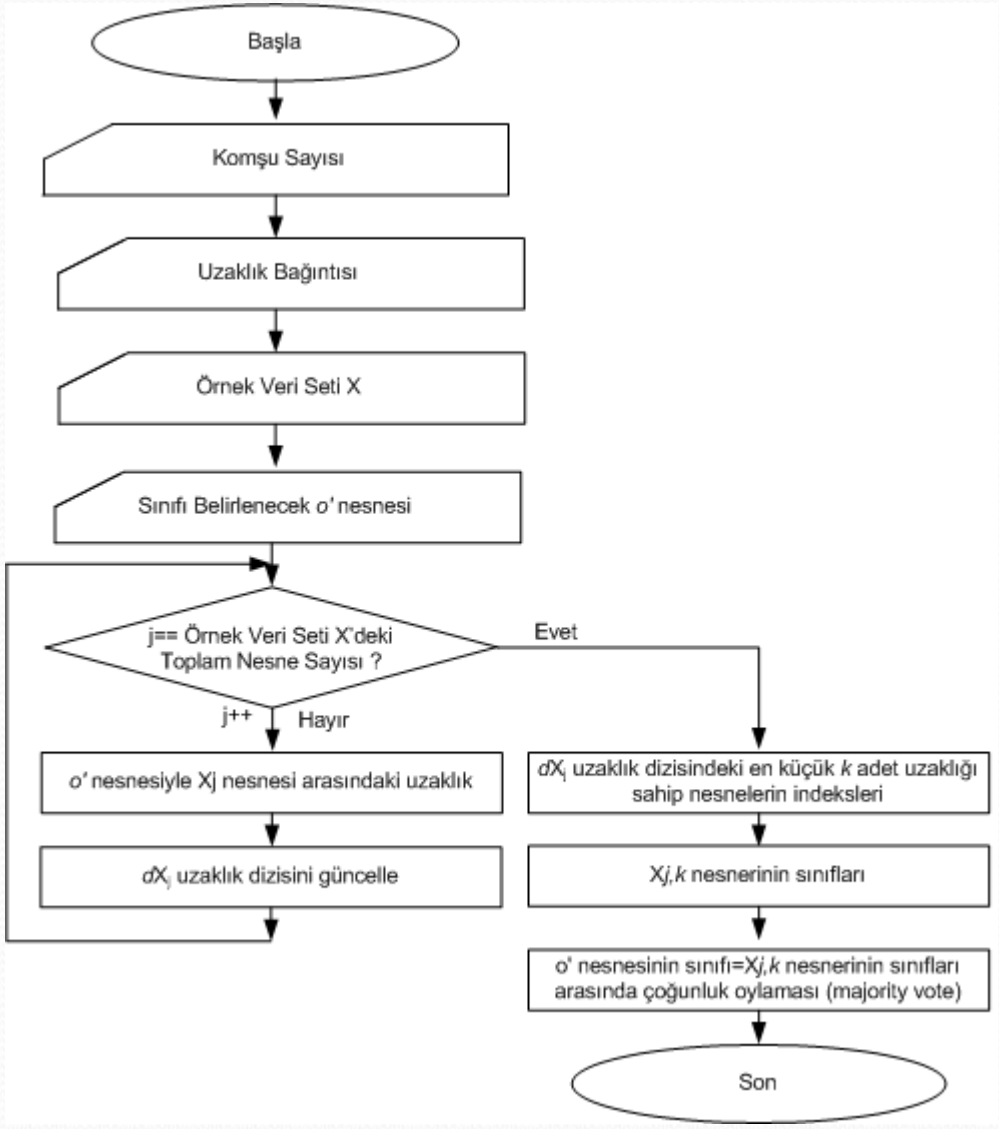
Algoritmanın Analiz Edilmesi

- Öznitelikler: probleme ait örnek veriyi temsil eden giriş parametreleri ve hedef parametre.
- Hedef parametrenin sınıfı aşağıdaki adımlar işletilerek bulunur.



k-nearest neighbor (k-nn) sınıflayıcı

Algoritmanın Akış Diyagramı ve Sözde Kod



1. Komşu sayısı seç k=1,2,3,..,n

2. Uzaklık Bağıntısı seç EU, MA, MI

3. X veri setini oku

4. o' nesnesini gir

5. ***for*** *j=1:length of X*

   d(o',xj) hesapla

6. d(o',xj) içerisinde en küçük *k*-adet nesnenin indekisini bul

7. *Xj,k* nesnelerinin sınıfları arasında "çoğunluk oylaması" yap ve o' nesnesinin sınıfını belirle

Örnek

- Aşağıda verilen veri kümesine göre, yeni bir gözlem olan X1 = 8, X2 = 4 değerlerinin hangi sınıfa ait olduğunu KNN ile bulalım

X1	X2	Y
2	4	KÖTÜ
3	6	İYİ
3	4	İYİ
4	10	KÖTÜ
5	8	KÖTÜ
6	3	İYİ
7	9	İYİ
9	7	KÖTÜ
11	7	KÖTÜ
10	2	KÖTÜ

Örnek

- K nın belirlenmesi
  - k = 4 kabul ediyoruz.
- Uzaklıkların hesaplanması:
  - (8,4) noktası ile gözlem değerlerinin her biri arasındaki Öklit uzaklıkları hesaplanır.

$$d(i,j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}=\sqrt{(2-8)^2+(4-4)^2}=6$$

$$d(i,j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}=\sqrt{(3-8)^2+(6-4)^2}=5{,}39$$

Örnek

Uzaklıkların hesaplanması :

$$d(i,j) = \sqrt{(2-8)^2 + (4-4)^2} = 6.00$$

$$d(i,j) = \sqrt{(3-8)^2 + (6-4)^2} = 5.39$$

$$d(i,j) = \sqrt{(3-8)^2 + (4-4)^2} = 5.00$$

$$d(i,j) = \sqrt{(4-8)^2 + (10-4)^2} = 7.21$$

$$d(i,j) = \sqrt{(5-8)^2 + (8-4)^2} = 5.00$$

$$d(i,j) = \sqrt{(6-8)^2 + (3-4)^2} = 2.24$$

$$d(i,j) = \sqrt{(7-8)^2 + (9-4)^2} = 5.10$$

$$d(i,j) = \sqrt{(9-8)^2 + (7-4)^2} = 3.16$$

$$d(i,j) = \sqrt{(11-8)^2 + (7-4)^2} = 4.24$$

$$d(i,j) = \sqrt{(10-8)^2 + (2-4)^2} = 2.83$$

Örnek

X1	X2	Uzaklık
2	4	6.00
3	6	5.39
3	4	5.00
4	10	7.21
5	8	5.00
6	3	2.24
7	9	5.10
9	7	3.16
11	7	4.24
10	2	2.83

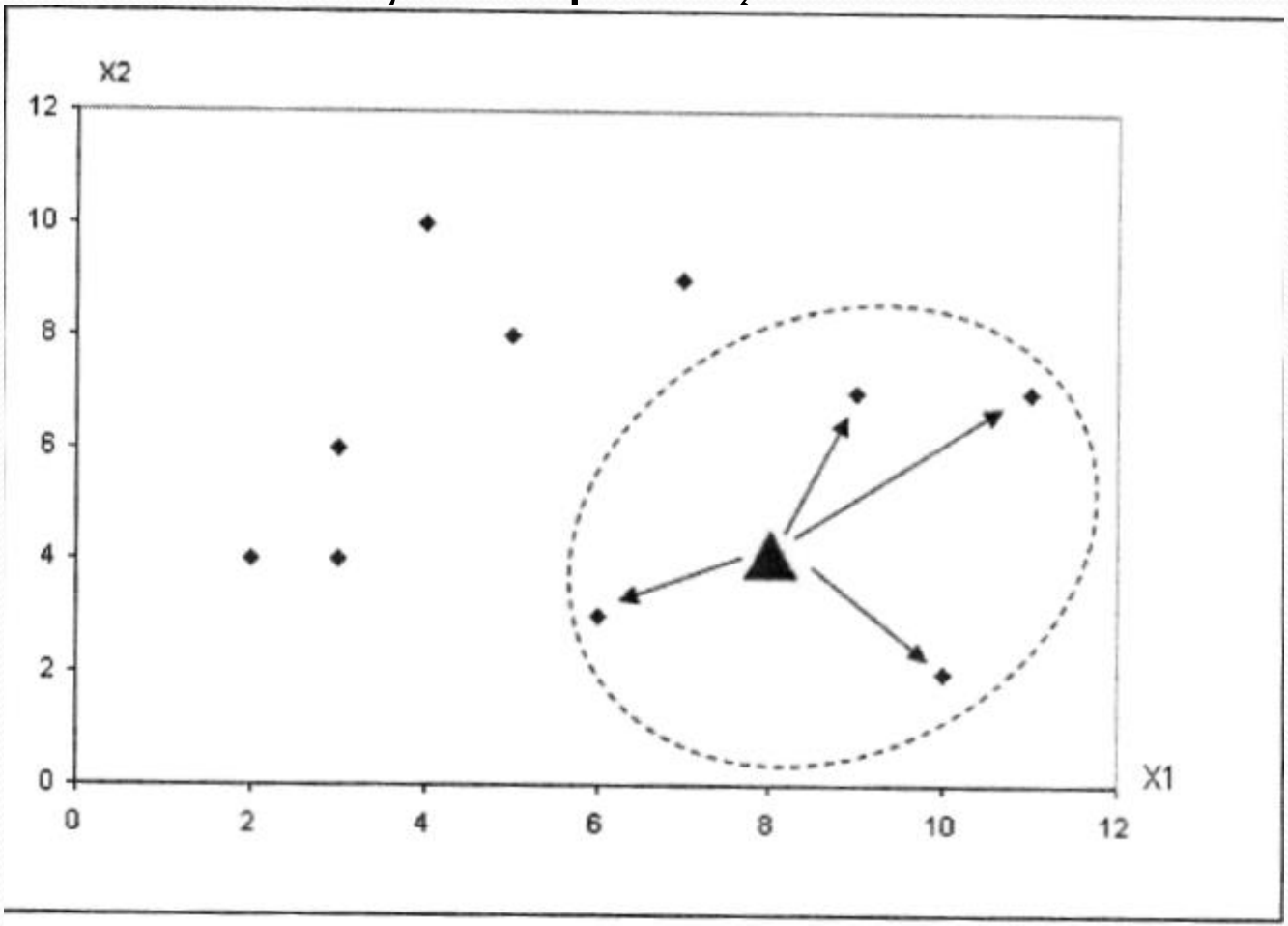
Örnek

- En küçük uzaklıkların belirlenmesi:
  - Satırlar sıralanarak k=4 en yakın komşular belirlenir.

X1	X2	Uzaklık	Sıra
2	4	6.00	9
3	6	5.39	8
3	4	5.00	6
4	10	7.21	10
5	8	5.00	5
6	3	2.24	1
7	9	5.10	7
9	7	3.16	3
11	7	4.24	4
10	2	2.83	2

Örnek

- (8,4) noktasına ek yakın 4 komşu:



Örnek

- Seçilen satırlara ilişkin sınıfların belirlenmesi
  - En yakın komşuların sınıf etiketleri göz önüne alınır
  - (8,4) test verisinin sınıfı “KÖTÜ” olarak belirlenir.

X1	X2	Uzaklık	Sıra	Y
2	4	6.00	9	
3	6	5.39	8	
3	4	5.00	6	
4	10	7.21	10	
5	8	5.00	5	
6	3	2.24	1	İYİ
7	9	5.10	7	
9	7	3.16	3	KÖTÜ
11	7	4.24	4	KÖTÜ
10	2	2.83	2	KÖTÜ

Örnek

- Ağırlıklı Oylama:

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

- Burada yer alan d(i,j) ifadesi i ve j gözlemleri arasındaki Öklid uzaklığıdır.
- Herbir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir.
- En büyük ağırlıklı oylama değerine sahip olan sınıf etiketi yeni test verisinin sınıfı kabul edilir.

Uygulama1

- K’nın belirlenmesi: K-en yakın komşu algoritması için k = 3
- Böylece (0.10, 0.50) gözlemine en yakın 3 komşu aranır.

Uygulama1

- Aşağıda verilen gözlem tablosuna göre, yeni bir gözlem olan (0.10, 0.50) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu algoritması ile bulalım

X1	X2	Cinsiyet
0.08	0.20	ERKEK
0.07	0.07	ERKEK
0.20	0.09	ERKEK
1.00	0.20	KADIN
0.05	0.06	ERKEK
0.20	0.25	ERKEK
0.17	0.07	ERKEK
0.15	0.55	KADIN
0.50	0.08	ERKEK
0.10	0.06	KADIN

Uygulama1

- Uzaklıkların hesaplanması: (0.10, 0.50) gözlemi ile diğer gözlem değerlerinin herbirisi arasındaki uzaklıkları Öklid uzaklık formülünü kullanarak hesaplanır

X1	X2	Uzaklık
0.08	0.20	0.30
0.07	0.07	0.43
0.20	0.09	0.42
1.00	0.20	0.95
0.05	0.06	0.44
0.20	0.25	0.27
0.17	0.07	0.43
0.15	0.55	0.07
0.50	0.08	0.58
0.10	0.06	0.44

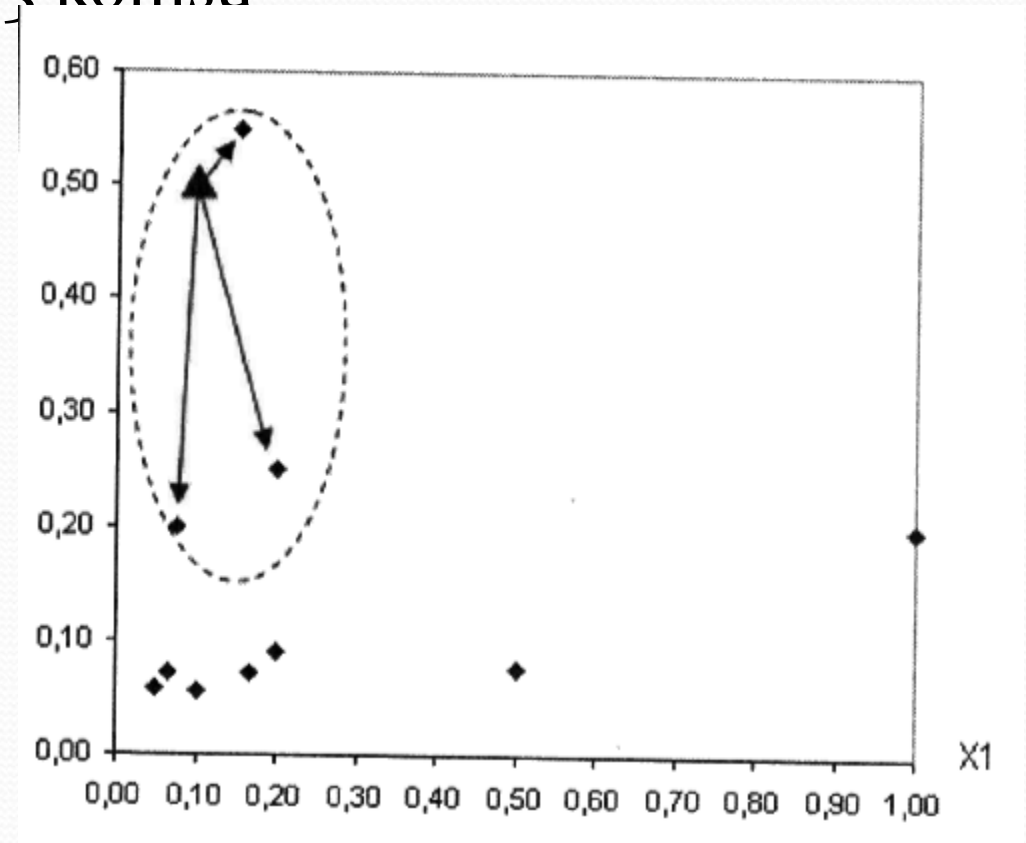
Uygulama1

- En küçük uzaklıkların belirlenmesi
  - Satırlar sıralanarak, en küçük k = 3 tanesi belirlenir.Bu üç nokta yeni gözlem noktasına en yakın noktalardır

X1	X2	Uzaklık	Sıra
0.08	0.20	0.30	3
0.07	0.07	0.43	5
0.20	0.09	0.42	4
1.00	0.20	0.95	10
0.05	0.06	0.44	7
0.20	0.25	0.27	2
0.17	0.07	0.43	6
0.15	0.55	0.07	1
0.50	0.08	0.58	9
0.10	0.06	0.44	8

Uygulama1

- En yakın 3 komsu



Uygulama1

- Seçilen satırlara ilişkin sınıfların belirlenmesi:

X_1	X_2	Uzaklık	Sıra	Cinsiyet
0.08	0.20	0.30	3	ERKEK
0.07	0.07	0.43	5	
0.20	0.09	0.42	4	
1.00	0.20	0.95	10	
0.05	0.06	0.44	7	
0.20	0.25	0.27	2	ERKEK
0.17	0.07	0.43	6	
0.15	0.55	0.07	1	KADIN
0.50	0.08	0.58	9	
0.10	0.06	0.44	8	

Uygulama1

- Ağırlıklı oylama yönteminin uygulanması:
- Bu aşamada, seçme işlemi “ağırlıklı oylama” yöntemiyle yapılıyor. O halde son tabloya d(i,j)’ ağırlıklı ortalamaları
- eklemek gerekiyor.

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

$$d(1, yeni\ gözlem)' = \frac{1}{0.30^2} = 11.046$$

$$d(6, yeni\ gözlem)' = \frac{1}{0.27^2} = 13.793$$

$$d(8, yeni\ gözlem)' = \frac{1}{0.07^2} = 200$$

Uygulama1

X1	X2	Uzaklık	Sıra	Cinsiyet	Ağırlıklı Uzaklık
0.08	0.20	0.30	3	ERKEK	11,05
0.07	0.07	0.43	5		
0.20	0.09	0.42	4		
1.00	0.20	0.95	10		
0.05	0.06	0.44	7		
0.20	0.25	0.27	2	ERKEK	13,79
0.17	0.07	0.43	6		
0.15	0.55	0.07	1	KADIN	200
0.50	0.08	0.58	9		
0.10	0.06	0.44	8		

Uygulama1

- Cinsiyet sınıfının “ERKEK” değerleri için ağırlıklı oylama değeri hesaplanır.

$$d(1,gözlem)'+d(6,gözlem)'=11,046+13,793=24,84$$

- Sonuç:
  - “Kadın” değeri için elde edilen ağırlıklı oylama değeri “erkek” değeri için elde edilenden daha büyük olduğundan yeni gözlem değerinin “kadın” sınıfına ait olduğu belirlenir.

Uygulama 2

- Aşağıda verilen gözlem tablosuna göre (7,8,5) noktasının hangi sınıf değerine sahip olduğunu bulalım.

X1	X2	X3	Y
10	5	19	EVET
8	2	4	HAYIR
18	16	6	HAYIR
12	15	8	EVET
3	15	15	EVET

Uygulama 2

- Gözlem değerlerini (0,1) aralığına göre dönüştürmek için min-max normalleştirme yöntemini uygulayalım

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

	X1	X2	X3
X_m	3	2	4
X_{max}	18	16	19

Uygulama 2

Birinci satırda X1 için X^* şu şekilde elde edilir:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$$= \frac{10-3}{18-3} = 0,47$$

Benzer biçimde birinci satırda X2 için X^* şu şekilde elde edilir:

$$X^* = \frac{5-2}{16-2} = 0,21$$

Birinci satırla ilgili olarak son hesaplama X3 için şu şekilde yapılır:

$$X^* = \frac{19-4}{19-4} = 1$$

Uygulama 2

- Dönüştürülmüş veriler:

X1	X2	X3	Y
0.47	0.21	1.00	EVET
0.33	0.00	0.00	HAYIR
1.00	1.00	0.13	HAYIR
0.60	0.93	0.27	EVET
0.00	0.93	0.73	EVET

- Dönüştürülmüş test değeri:
  - (7,8,5) → (0.26 , 0.43, 0.07) elde edilir.

Uygulama 2

- K nın belirlenmesi:
  - K-en yakın komşu algoritması için k = 3 kabul edilir.
- Uzaklıkların hesaplanması:
  - (0.26, 0.43, 0.07) noktası ile dönüştürülmüş değerlerin herbirisi arasındaki Öklid uzaklıkları hesaplanır.

Uygulama 2

- Gözlem değerlerinin (0.26, 0.43, 0.07) noktasına olan uzaklıkları:

X1	X2	X3	Uzaklık
0.47	0.21	1.00	0.98
0.33	0.00	0.00	0.44
1.00	1.00	0.13	0.93
0.60	0.93	0.27	0.63
0.00	0.93	0.73	0.87

- En küçük uzaklıkların belirlenmesi: Uzaklık gözönüne alınarak k = 3 komşu gözlemin belirlenmesi

X1	X2	X3	Uzaklık	Sıra
0.47	0.21	1.00	0.98	5
0.33	0.00	0.00	0.44	1
1.00	1.00	0.13	0.93	4
0.60	0.93	0.27	0.63	2
0.00	0.93	0.73	0.87	3

Uygulama 2

- Y sınıfına ilişkin ilk 3 değerin belirlenmesi: Seçilenler arasında 'EVET' lerin sayısı 'HAYIR' dan daha fazla olduğu için yeni gözlemin sınıfı 'EVET'dir.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri
0.47	0.21	1.00	0.98	5	
0.33	0.00	0.00	0.44	1	HAYIR
1.00	1.00	0.13	0.93	4	
0.60	0.93	0.27	0.63	2	EVET
0.00	0.93	0.73	0.87	3	EVET