Data Mining
Classification: Alternative Techniques

Lecture Notes for Chapter 4

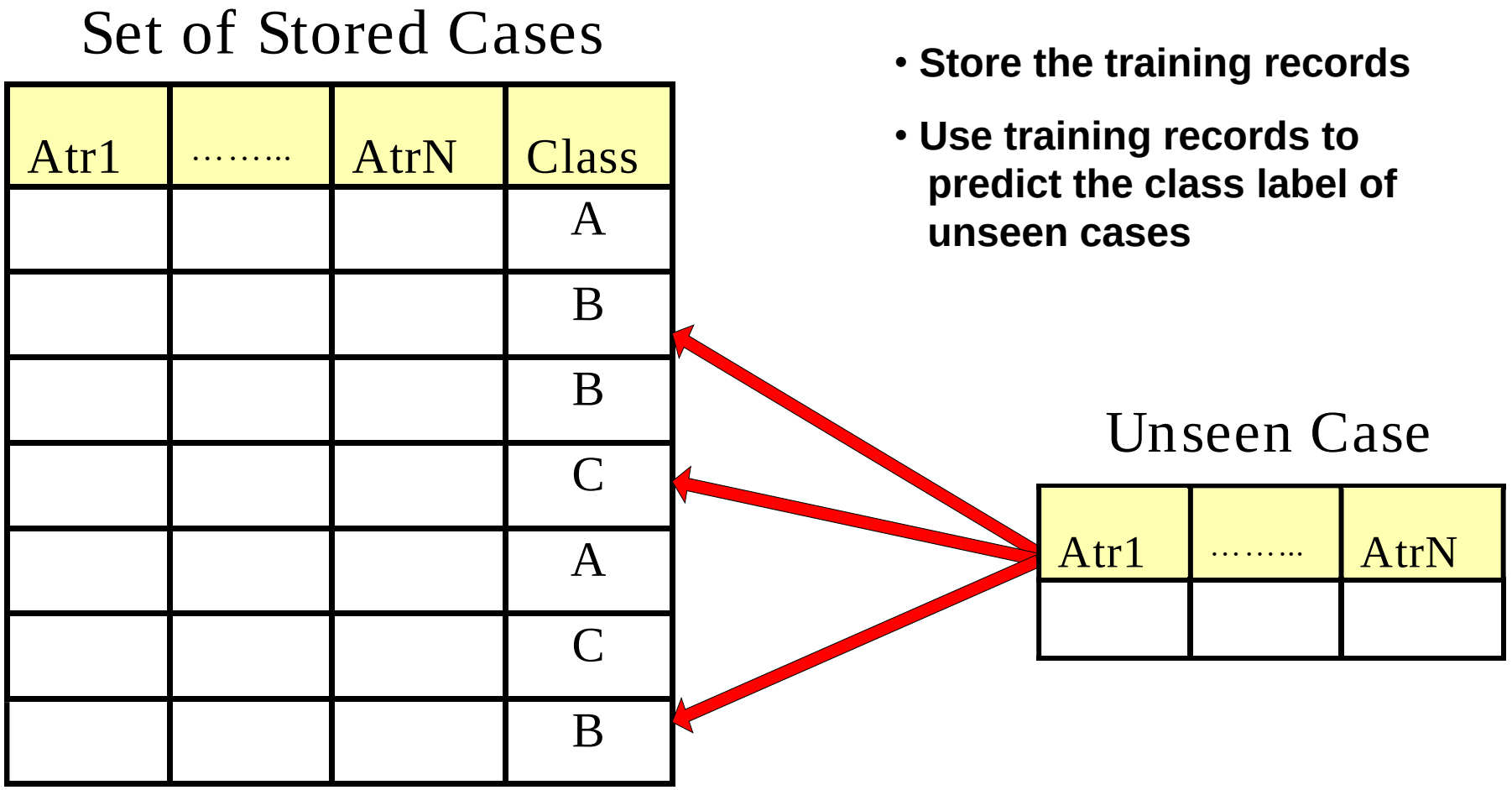
Instance-Based Learning

Introduction to Data Mining , 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar

Instance-Based Classifiers



Instance Based Classifiers

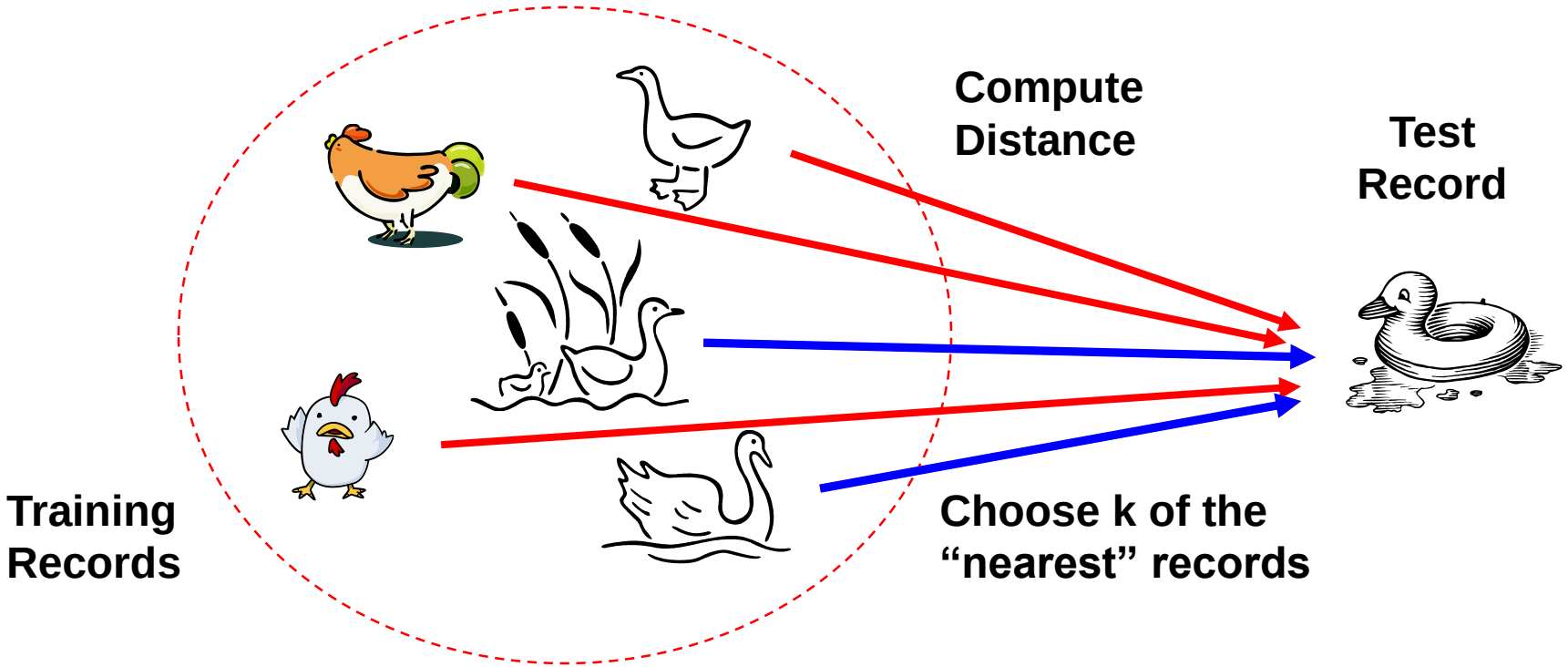
Karar ağacı ve kural tabanlı sınıflandırıcılar, istekli öğreniciler (**eager learners**) olarak adlandırılır. Çünkü eğitim verileri kullanılabilir hale gelir gelmez giriş özniteliklerini sınıf etiketine eşleyen bir modeli öğrenmek için tasarlanmışlardır.

Tam tersi bir strateji de test örneklerini sınıflandırmak gerekene kadar eğitim verilerini modelleme sürecini ertelemektir. Bu stratejiyi kullanan teknikler, tembel öğreniciler (**lazy learners**) olarak bilinir.

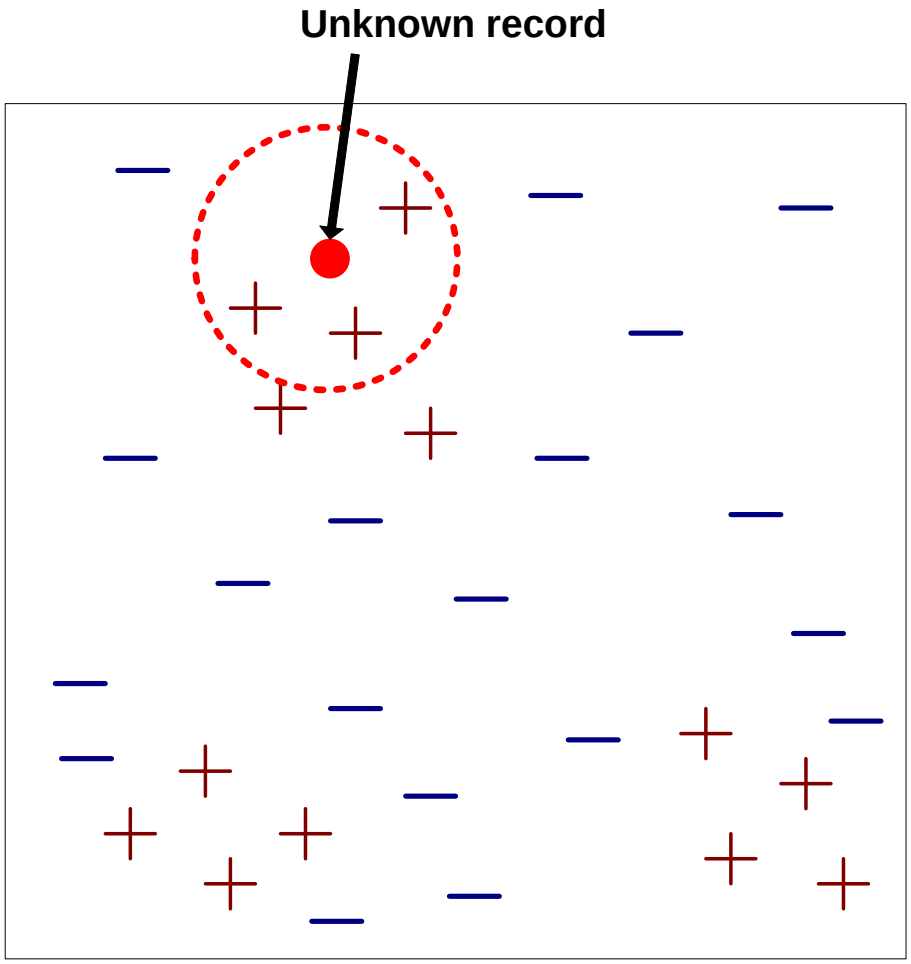
- Examples:
  - Rote-learner
    - Tüm eğitim verilerini ezberler ve yalnızca kaydın özellikleri, eğitim örneklerinden biriyle tam olarak eşleşirse sınıflandırmayı gerçekleştirir
    - Bariz dezavantajı (drawback)
      - bazı test kayıtları, herhangi bir eğitim örneğiyle eşleşmediği için sınıflandırılmayabilir
  - Nearest neighbor
    - Sınıflandırma yapmak için "en yakın" k tane noktayı (en yakın komşular) kullanır

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



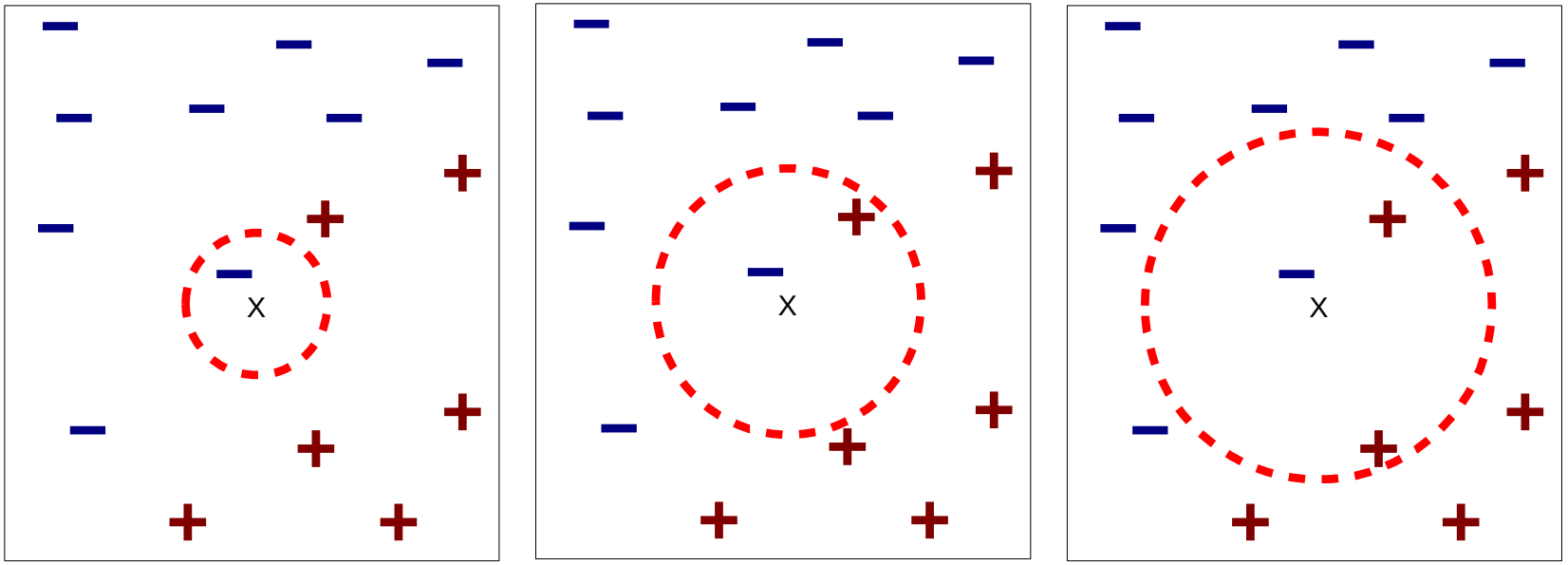
Nearest-Neighbor Classifiers



- Üç şey gerektirir
  - **The set of labeled records**
  - **Distance Metric** (kayıtlar arasındaki mesafeyi hesaplamak için)
  - **The value of *k***, (alınacak en yakın komşuların sayısı)
- Bilinmeyen bir kaydı (unknown record) sınıflandırmak için:
  - Diğer eğitim kayıtlarına olan mesafeyi hesapla
  - En yakın K komşuyu belirle
  - Bilinmeyen kayıtların sınıf etiketini belirlemek için en yakın komşuların sınıf etiketlerini kullanın (örneğin, çoğunluk oyu alarak, majority vote)

Definition of Nearest Neighbor

The 1-, 2-, and 3-nearest neighbors of an instance.



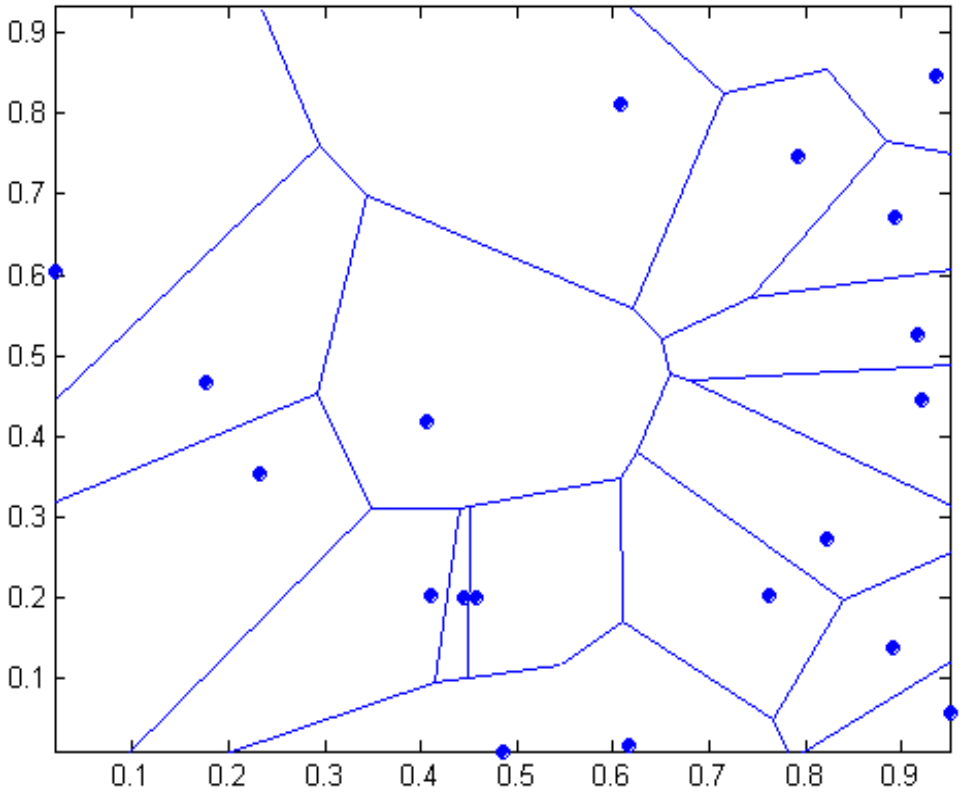
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Bir x kaydının K-en yakın komşuları, x'e en küçük mesafeli k tane veri noktasıdır.

1 nearest-neighbor

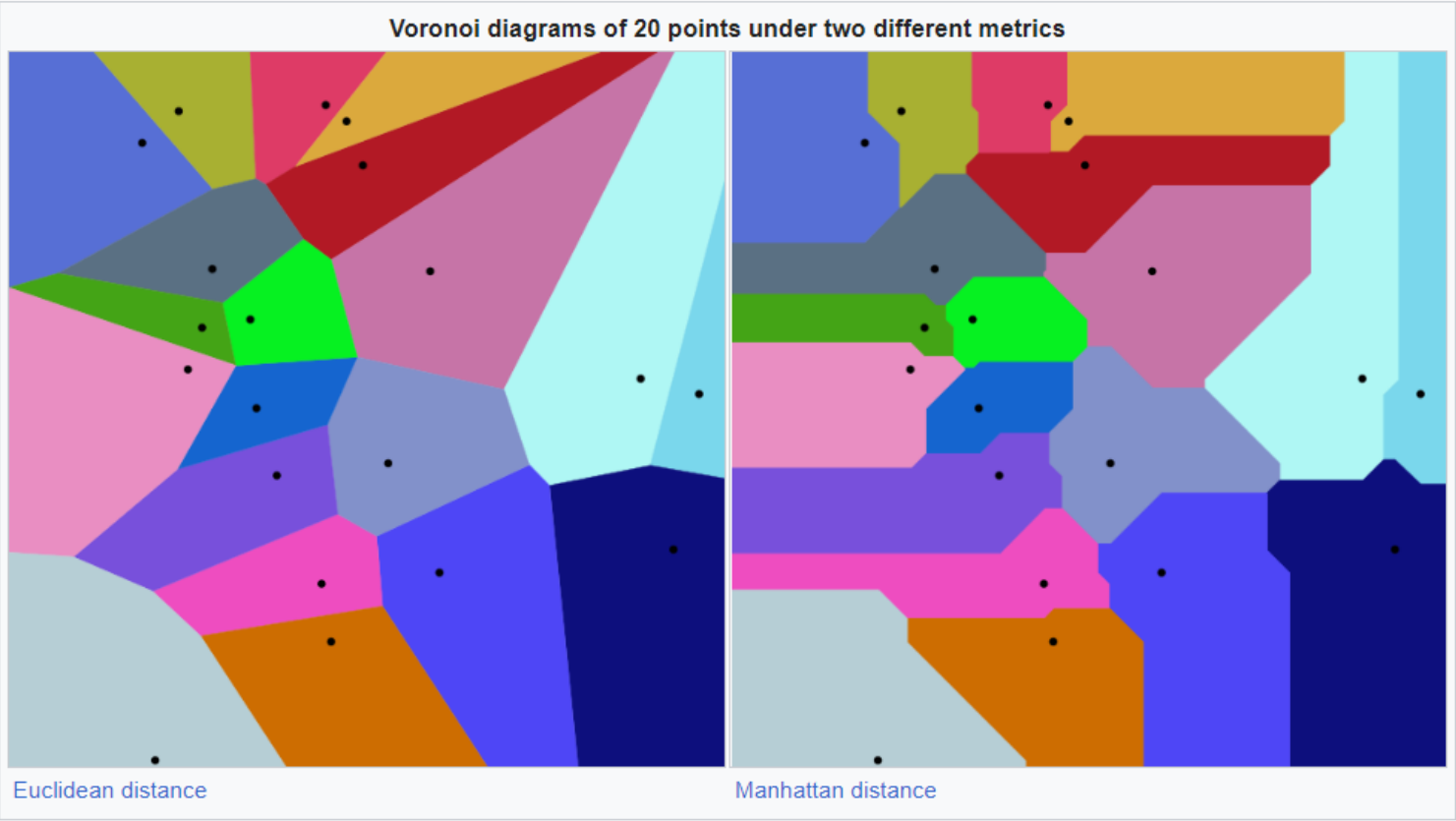
Voronoi Diagram

Matematikte, bir Voronoi diyagramı, bir düzlemin belirli bir nesne kümesinin her birine yakın bölgelere bölünmesidir. En basit durumda, bu nesneler düzlemde yalnızca sonlu sayıda noktadır (called seeds, sites, or generators**). Her çekirdek (**seed**) için, düzlemin o çekirdeğe diğerlerinden daha yakın olan tüm noktalarından oluşan, Voronoi hücresi adı verilen karşılık gelen bir bölge vardır.**



1 nearest-neighbor

Voronoi Diagram



Nearest Neighbor Classification

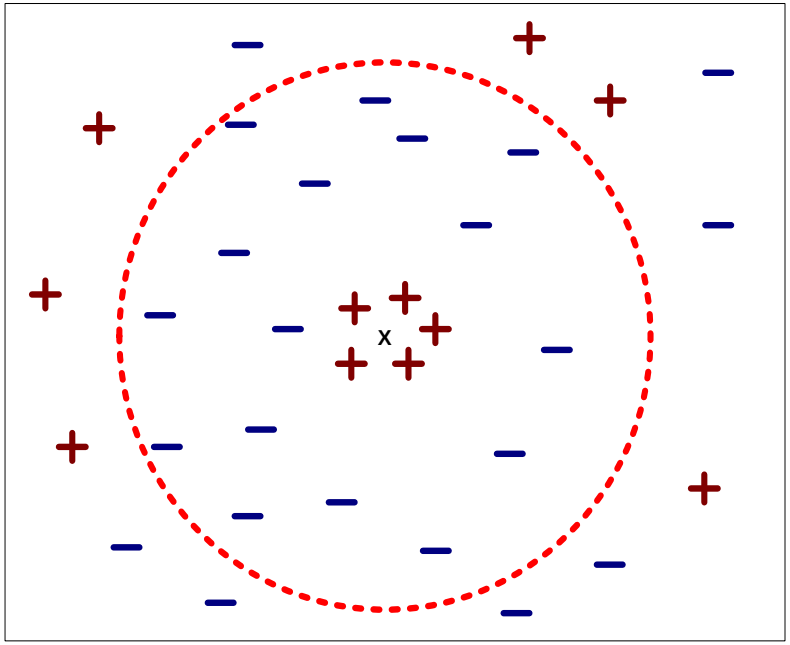
- İki nokta arasındaki mesafeyi hesaplamak için:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- En yakın komşu listesinden sınıfı belirle
  - K en yakın komşular arasında sınıf etiketlerinin çoğunluk oylaması yapılır (majority vote)
  - Oyu mesafeye göre ağrlıklandırma
    - weight factor, $w = 1/d^2$

Nearest Neighbor Classification…

- Choosing the value of k:
  - K çok küçükse gürültü noktalarına duyarlı (***sensitive to noise points***)
  - K çok büyükse, komşuluk diğer sınıflardan noktalar içerebilir



Nearest Neighbor Classification…

- Scaling issues
  - Uzaklık ölçülerine özniteliklerden birinin hakim olmasını önlemek için özniteliklerin ölçeklendirilmesi gerekebilir
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

Use a scaling function

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Bu işlem, orijinal veri kümesinin sütunlarını ölçeklendirir, böylece her bir özniteliğin değerleri 0-1'e eşlenir.

Nearest Neighbor Classification…

- Selection of the right similarity measure is critical:

1 1 1 1 1 1 1 1 1 1 1 0		0 0 0 0 0 0 0 0 0 0 0 1
	vs	
0 1 1 1 1 1 1 1 1 1 1 1		1 0 0 0 0 0 0 0 0 0 0 0

Euclidean distance = 1.4142 for both pairs

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (10+0)/12=**0.833** for both pairs

Jaccard = $(M_{11}) / (M_{01} + M_{10} + M_{11})$ = 10/12 =**0.833** for the first pair , 0/2=**0** the other

Nearest Neighbor Classification…

Algorithm 5.2 The k-nearest neighbor classification algorithm.

1: Let k be the number of nearest neighbors and D be the set of training examples.
2: **for** each test example $z = (\mathbf{x}', y')$ **do**
3: Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4: Select $D_z \subseteq D$, the set of k closest training examples to z.
5: $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
6: **end for**

$$\text{Majority Voting: } y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i), \qquad (5.7)$$

where v is a class label, y_i is the class label for one of the nearest neighbors, and $I(\cdot)$ is an indicator function that returns the value 1 if its argument is true and 0 otherwise.

$$\text{Distance-Weighted Voting: } y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i \times I(v = y_i). \qquad (5.8)$$

$$w_i = 1/d(\mathbf{x}', \mathbf{x}_i)^2.$$

Characteristics of Nearest-Neighbor Classifiers

- k-NN sınıflandırıcılar, modeli açıkça oluşturmadıkları için tembel öğrenicilerdir (**lazy learners**)
- Bilinmeyen kayıtları sınıflandırmak nispeten maliyetlidir (Karar ağacı indüksiyonu ve kural tabanlı sistemler gibi istekli (*eager*) öğrencilerin aksine)
- Keyfi şekillendirilmiş karar sınırları üretebilir **(arbitrarily shaped decision boundaries)**
- Kararlar yerel bilgilere dayandığından, değişken etkileşimlerini yönetmek kolaydır

Characteristics of Nearest-Neighbor Classifiers

- Sınıflandırma kararları yerel olarak verildiği için, en yakın komşu sınıflandırıcılar (küçük k değerleri için) gürültüye oldukça duyarlıdır.
- Doğru yakınlık ölçüsünün seçimi önemlidir
- Gereksiz ve fazlalık öznitelikler sorun yaratabilir
- Eksik özelliklerin üstesinden gelmek zordur