Makine Öğrenmesi
Kümeleme

Doç.Dr. İlhan AYDIN

Kümeleme sorunu

- Kümeleme, denetimsiz makine öğrenme bir sorunudur

Kümeleme sorunu

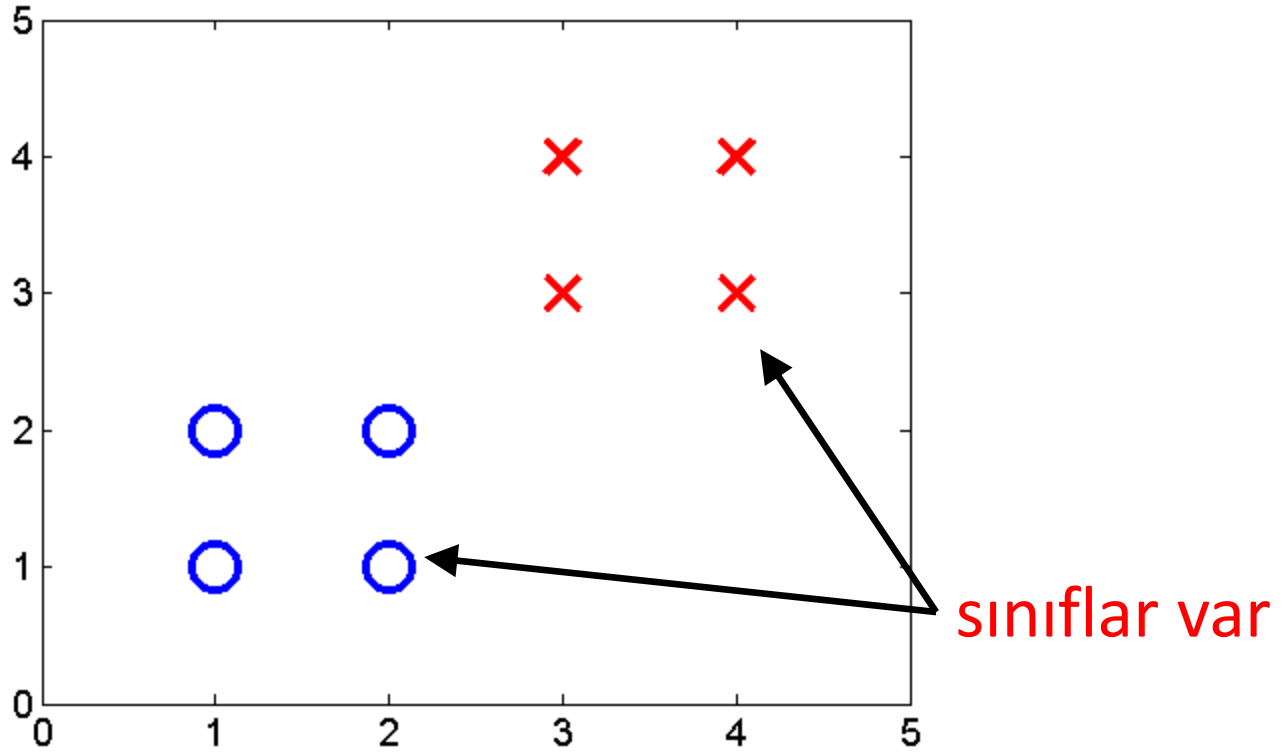
- Denetimli öğrenme, hatırlatma:
  - Modellenecek ilişki için girdi-çıktı, neden-sonuç, durum-sınıf, vb örnekler var
  - Var olan örnekleri kullanarak bir model ve gelecek durumlar için karar etme yöntemini geliştirmek lazım

Kümeleme sorunu

- Denetimsiz öğrenme, hatırlatma:
  - Modellenecek ilişki için girdi-çıktı, neden-sonuç, durum-sınıf, vb örnekleri yok
  - Sadece “etiketsiz” veri kümesi var
  - Hem verilerin yapısını anlamak hem de ilişki model ve veriler sınıflandırılmasını bulmak lazım

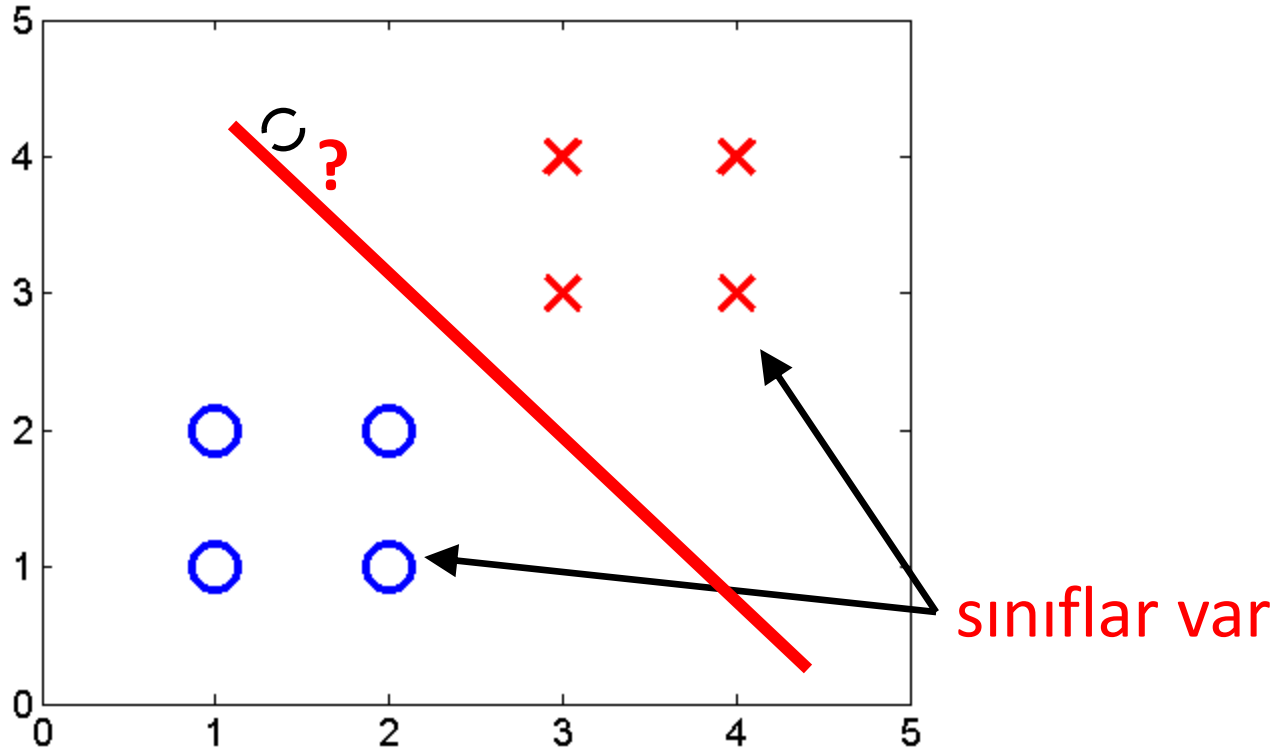
Kümeleme sorunu

Denetimli öğrenme: sınıfların örnekleri var



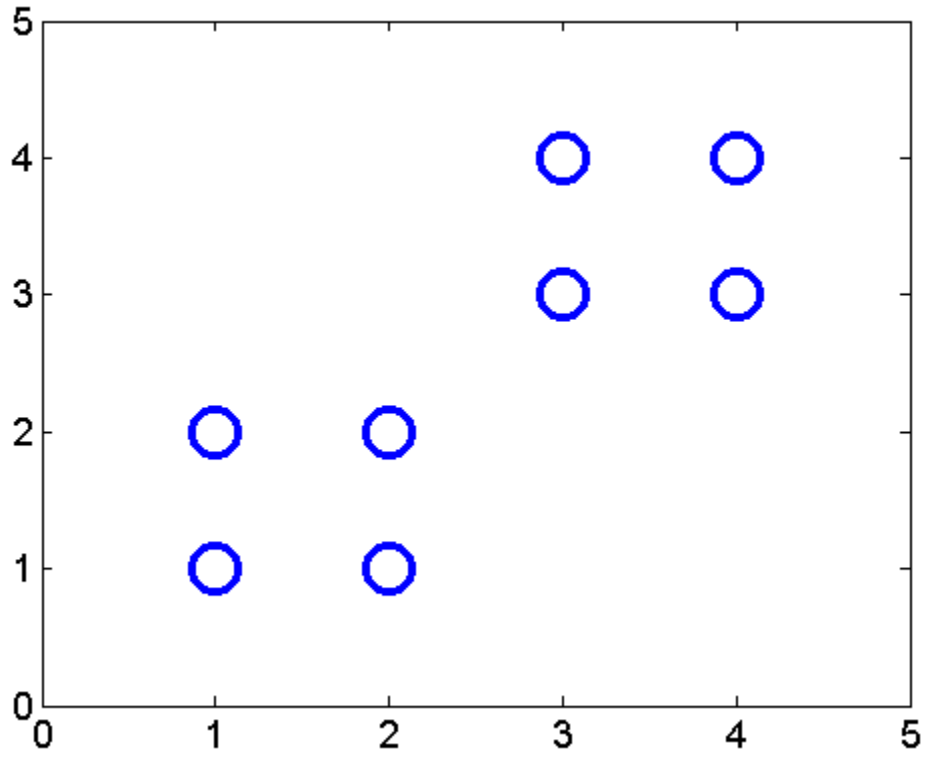
Kümeleme sorunu

Bu örnekleri kullanarak karar modeli (mesela, lineer karar sınırı) oluşturulabilir



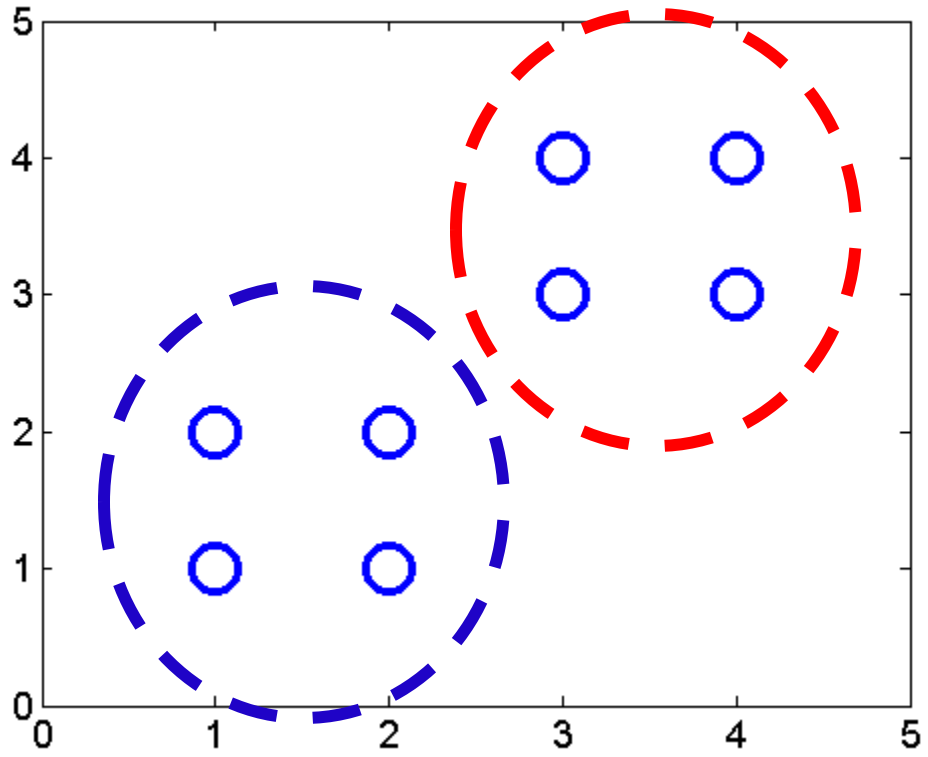
Kümeleme sorunu

Denetimsiz öğrenme: sınıflar yok



Kümeleme sorunu

Hem sınıflar hem de onlarındaki noktaları belirtmek lazım



Kümeleme sorunu

- Genel veriler için, yani etiketsiz veriler için, hem olabilir sınıfılar etiketleri hem de örneklerin sınıflandırılması yöntemi belirtmek lazım

Kümeleme sorunu

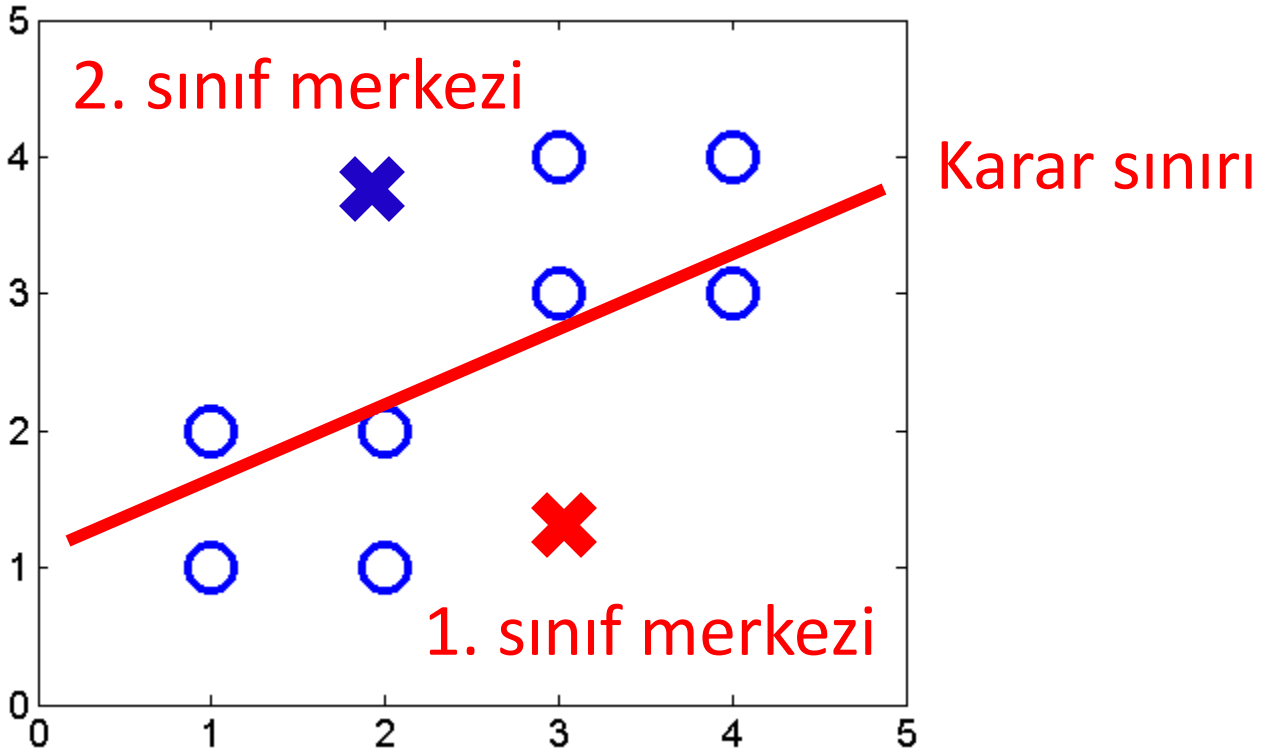
- Kümeleme önemli ve yaygın sorundur, bir çok yöntem de var
- Biz, en kolay ve çok popüler K-mean yöntemine bakacağız

Kümeleme sorunu

- K-means yöntemi, lineer sınıflandırma yaklaşımıdır
- Bu yönteme göre, iki veri sınıfı tanımlamak için iki merkez noktası belirtilir
- Örnekler, her zaman en yakın merkezinin sınıfına konulmuştur
  - Bu sınıflandırma yöntemi, lineer karar sınırıyla temsil edilebilir

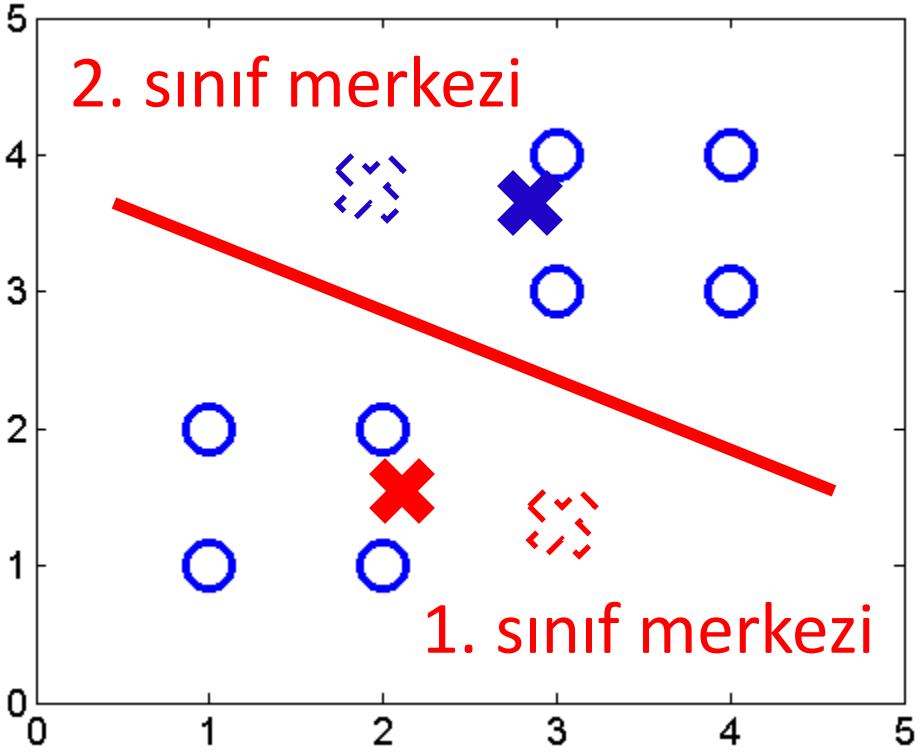
Kümeleme sorunu

1. İki merkez noktasını belirtip örnekler merkez noktasına göre bölünür



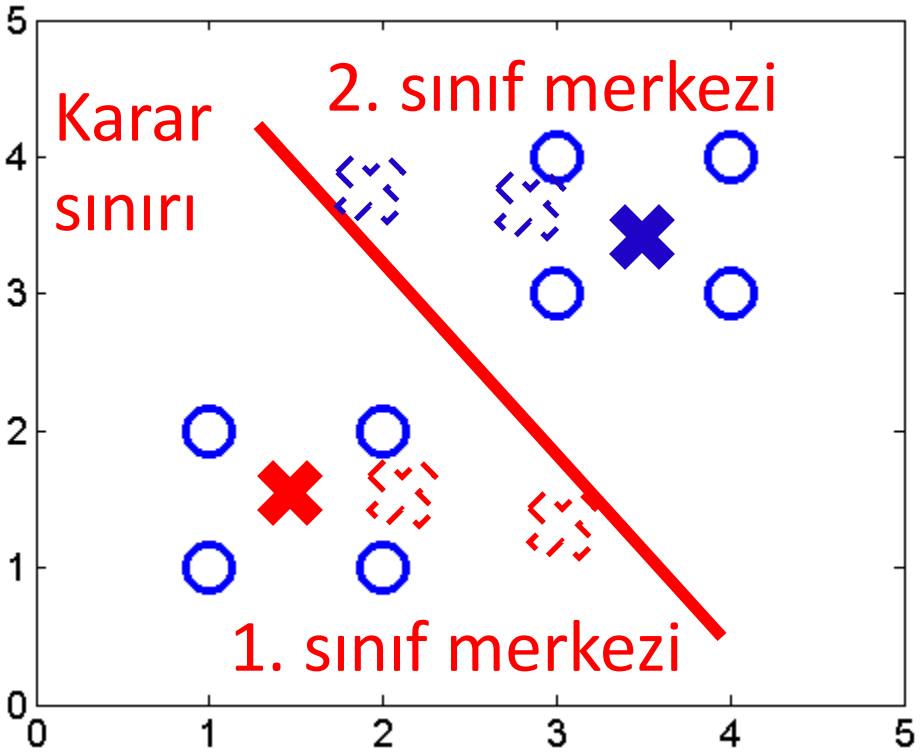
Kümeleme sorunu

2. Bu şekilde örnekleri bölüp merkez noktalarını örneklerin ortalama pozisyonu olarak güncelleştirilir



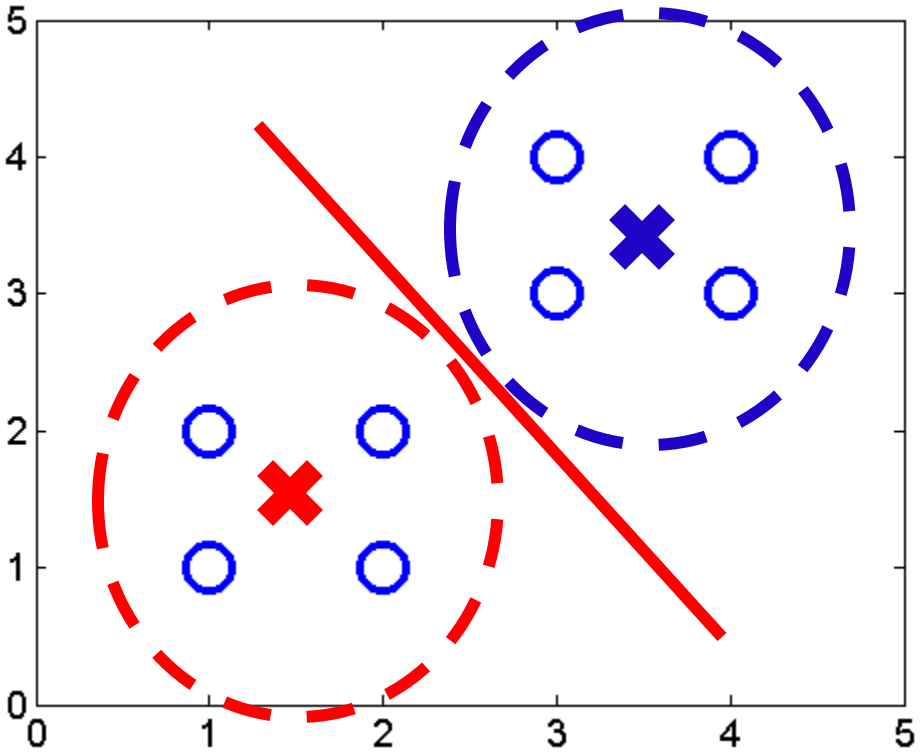
Kümeleme sorunu

3. Yeniden örnekleri bölüp merkez noktalarını tekrar güncelleştirilir



Kümeleme sorunu

Sonuçta, iki küme, merkezleri ile belirtilir, ve bu kümelere göre örneklerin atanması belirtilir



Kümeleme sorunu

- Matematiksel şekilde, K-means yöntemi bu şekilde tanımlanır;
  - Birkaç (2 yada daha çok) rasgele sınıf merkezi seçilir
  - Bütün örnekler, en yakın merkezlerin sınıflarına konulur
  - Atanmış örneklere göre yeni sınıf merkezleri ortalama pozisyon olarak hesaplanır
  - Tekrarlanır

Kümeleme sorunu

- K-means yöntemi;

$$\mu_1,\ldots,\mu_k = rand(\Re^n)$$

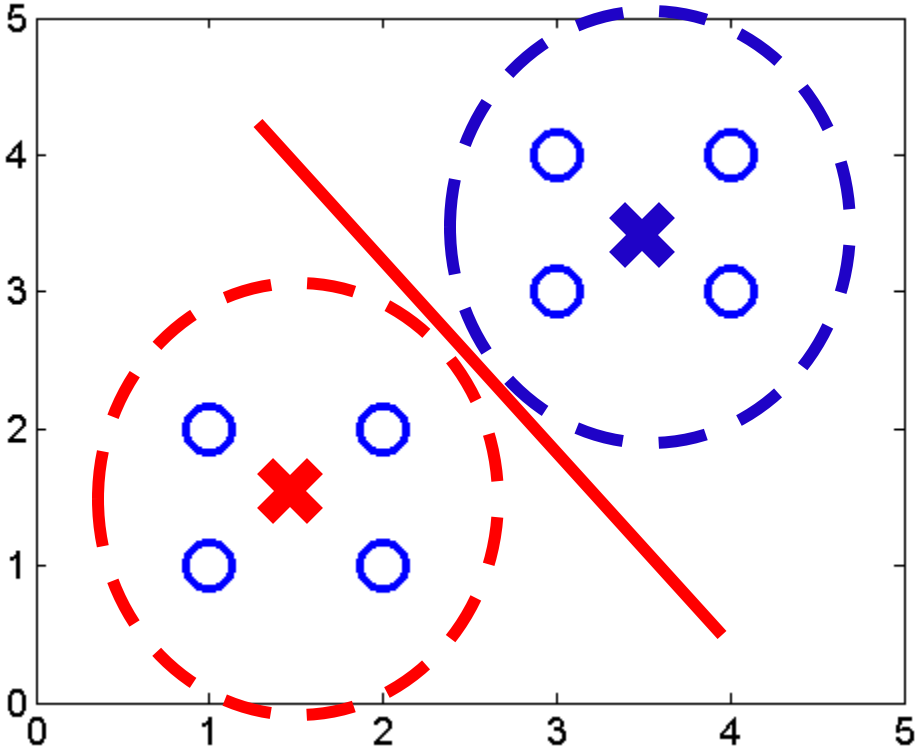
$$butun\ ornek\ icin\ etiket\ atayin$$

$$c^i = \{x^i\text{'}ye\ en\ yakin\ \mu_j\}$$

$$\mu_j \rightarrow ortalama(x^i : c^i = \mu_j)$$

Kümeleme sorunu

- Verilerin birkaç iyi ayrılmış kümesi varsaydı, K-means yöntemi bu kümeleri bulabiliyor



Kümeleme sorunu

- Verilerin birkaç iyi ayrılmış kümesi yoksa, K-means yöntemi bir optimal şekilde verilerin bölgelenmesi verecek

Kümeleme sorunu

- K-mean yöntemi için maliyet fonksiyonu şekilde tanımlanabilir; böyle maliyet fonksiyonuna "*distorisyon*" denir;

$$J(\{c\},\{\mu\}) = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu^i)^2$$

Yani örneklerin merkezlerden toplam ortalama mesafesi

Kümeleme sorunu

- Distorsiyonu adım adım azaltırken K-means yöntemi çıkmaktadır

$$J(\{c\},\{\mu\}) = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu^i)^2 \rightarrow \min$$

Kümeleme sorunu

- Uygulama sorunları:
  - K-means algoritmasını başlatmak için sınıflar için ilk merkezleri elleriyle seçilmesi lazım
  - Örneklerin sırasından iki yada birkaç rasgele nokta seçilir ve onları sınıf merkezleri olarak kullanılır
  - K-means algoritması birkaç defa tekrarlanması gerekebilir: distorsiyonun birçok lokal minimumu olması yüzden algoritmanın bir geçişi iyi kümeler vermeyebilir

Kümeleme sorunu

- Uygulama sorunları:
  - K-means yöntemi, veriler iki yada birkaç, K, sınıfa bölüyor (şundan, **K**-means yöntemi) ama ...
  - Sınıf sayısını, K'yı, önceden seçmek lazım;
    - K, aşağı yukarı gerçekten var olan sınıf sayısına eşit olmalıdır !
    - Bunu bilmeyebiliriz hiç ...

Kümeleme sorunu

- Uygulama sorunları:
  - Sınıf sayısını belirtmek için “eğme noktası” (*elbow point*) metodu kullanılabilir
    - K-means algoritması, K=1,2,3,4,... ile çalıştırılır
    - Farklı K için, en iyi J distorsiyon değerleri çizilir
    - Bu grafikte bir “eğme noktası” olabilir
    - Bu şekilde, eğme noktası gerçek sınıf sayısını belirtir

Kümeleme sorunu

- Eğme noktası:

Kümeleme sorunu

- Diğer taraftan, K kümelemenin son amacına göre seçilebilir
- K-means, örnekleri en uygun birkaç sınıfa bölüyor; bu iş farklı bir amacıyla yapılabiliyor
- Bu “son” amacı bazen kullanılacak K değerini de belirtebiliyor

Kümeleme sorunu

- Örneğin: giyim üretmede üretilecek giyim birkaç ayrık boyutlarda üretilmelidir (Small-Medium-Large)
- Tabii ki gerçek insanlar 3 boyutta değil; bu demek ki, gerçek insanların boyut dağılımına göre üretici üç en uyugun orta noktası seçip bütün giyimleri şu boyutta üretmek zorundadır

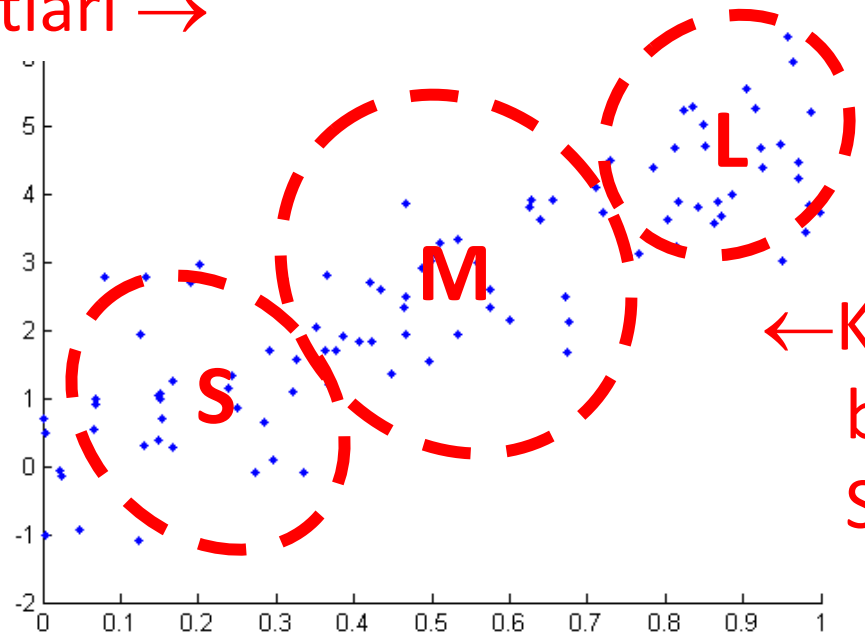
Kümeleme sorunu

- Bu en uygun kullanılacak boyutlar, K-means algoritmasıyla seçilebilir
- Bu durumda, K=3 olması lazım, çünkü (son amacımız olarak) giyim üç boyutta üretilecek
- İyi ayrılmış sınıf burada yok, ama K-means yine de verileri üç en uygun sınıfa bölebiliyor

Kümeleme sorunu

K-means'deki K değerini kümeleme son amacına göre 3 değerinde seçilmiştir

Gerçek insanların
giyim boyutları →



←K-means tarafından
bulunmuş en uygun
S-M-L boyut sınıfları