

# VERİ MADENCİLİĞİ YILIÇI SINAVI

Fen Bilimleri Enstitüsü

Yrd. Doç. Dr. Şule Ögüdücü

Süre 90 dakikadır.

10.11.2005

- (15 puan) Veritabanı ve veri madenciliği işlemleri arasındaki farkları yazınız. Örneklerle açıklayınız.
- (15 puan) Veri madenciliği uygulamalarında kullanılan modeller kaç'a ayrılır? Açıklayınız.
- Bir veri madenciliği uygulaması tasarlanacaktır. Çalışılacak veri kümesindeki niteliklerden biri *yaş* bilgisidir. Bu niteliğe ait değerler şu şekildedir: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 45
  - (10 puan) *yaş* niteliğine ait veri için gürültüyü azaltmak amacıyla eşit aralıklarla bölme genişliği 3 olacak şekilde bölmeleme (*binning*) yapılacaktır. Bu işlemi yapmak için gerekli adımları gösteriniz. Yukarıdaki veri için bu işlemin etkilerini yorumlayınız.
  - (10 puan) Verideki aykırılıkları belirlemek için hangi işlemler yapılabilir?
  - (10 puan) Verideki gürültüyü düzeltmek için başka hangi yöntemler uygulanabilir?
- Aşağıda verilen tablodaki verileri sınıflandırmak için karar ağacı oluşturulacaktır.

## İPUCU:

Entropi kullanarak:

| A | B | Sınıf |
|---|---|-------|
| T | F | +     |
| T | T | +     |
| T | T | +     |
| T | F | -     |
| T | T | +     |
| F | F | -     |
| F | F | -     |
| F | F | -     |
| T | T | -     |
| T | F | -     |

$$Entropy(S) = - \sum_{i=1}^c p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{\nu \in values(A)} \frac{|S_\nu|}{|S|} Entropy(S_\nu)$$

Gini index kullanarak:

$$gini(S) = 1 - \sum_{i=1}^c p_i^2$$

$$Gain(S, A) = gini(S) - \sum_{\nu \in values(A)} \frac{|S_\nu|}{|S|} gini(S_\nu)$$

- (15 puan) A ve B nitelikleri için bilgi kazancı hesaplayın. Karar ağacı algoritması hangi niteliği kullanarak bölmeleme yapar?
  - (15 puan) A ve B nitelikleri için Gini index değerini hesaplayınız. Karar ağacı algoritması hangi niteliği kullanarak bölmeleme yapar?
  - (10 puan) Bilgi kazancı ve gini index değerlerinin bölmeleme için farklı nitelikleri seçmesi mümkün müdür? Açıklayınız.
5. (ödürl sorusu) Aşağıdaki şekilde  $C_1$ ,  $C_2$  ve  $C_3$  üç demedin merkezidir ve  $A, B, D, E$  nesneleri de Öklid uzaklığı kullanılarak demetlenecektir. Düz çizgiler demet merkezleri arasındaki

uzaklığı göstermektedir. K-means algoritmasına göre her nesne en yakın merkezin bulunduğu demede atanır. Üç demet merkezi olduğuna göre bir nesnenin hangi demette olduğunu bulmak için 3 uzaklık hesaplaması yapmak gerekir (örneğin:  $d(A, C_1)$ ,  $d(A, C_2)$ ,  $d(A, C_3)$ ). Toplam dört nesne bulunduğuna göre K-means algoritmasının her yinelenmesinde toplam  $4 \times 3$  uzaklık hesaplaması yapmak gerekir. Ancak, uzaklık hesaplamalarını azaltmak için üçgen eşitsizliğinden yararlanılabilir.

- (10 puan) Üçgen eşitsizliği kullanılarak gereksiz uzaklık hesaplamalarının azaltılabileceğini gösteriniz.
- (10 puan) Şekildeki nesneler için hangi uzaklık hesaplamalarının yapılmasına gerek kalmaz?

