

BİL3013 Veri Madenciliğine Giriş – Ödev
Dersin Öğr. Elemanı: Doç. Dr. Üyesi Mete Eminağaoğlu

Ödevin son teslim tarihi: 11 Aralık 2021 Cumartesi, 22:00 (TSİ)

(Ödev teslimi için **ek süre kesinlikle verilmeyecektir**. Herhangi bir nedenle **zamanında iletilmeyen ödevler, hiçbir mazeret kabul edilmeden 0 (sıfır) olarak değerlendirilecektir.**)

DİKKAT: Bu ödev **tek kişiliktir**. Her bir öğrenci, bu ödevi **tek başına** yapmak zorundadır. En ufak bir yardım, vb. kopya / intihal olarak değerlendirilecektir ve **yardım veren / alan öğrencilerin hepsi** bu ödevden **0 (sıfır) alacaktır**.

GENEL HATIRLATMA:

Aşağıdaki yazılanları baştan sona okumanız ve eksiksiz uymanız zorunludur. Ödevle ilişkin kodlamayı tam ve hatasız yapıp, zamanında iletip, hatasız çalıştırsanız bile, aşağıdaki yazılı kural, talimat, vb. bir tanesine bile uyulmazsa, ödevinizden puan / puanlar kılacaktır.

Ödevin Teslim Şekli:

DEÜ Sakai sistemindeki ders sayfasında açılacak olan ödev yükleme (assignment) alanına; tüm dosyalar (tüm kaynak kodları, bağlantılı kütüphane, dizinler, proje dosyaları, vb.), **zip / rar şeklinde sıkıştırılmış tek bir dosya olarak yüklenecektir**.

Ödev Konusu:

Sakai sistemindeki Assignments - Ödev kısmında size iletilmiş olan “trainSet.csv” adlı dosyada bulunan eğitim (train) ve “testSet.csv” adlı dosyada bulunan test verileri kullanılarak karar ağacı algoritması kodlanacak ve ikili sınıflandırma yapılacaktır.

- **Veri Seti ile ilgili bilgiler:**
 - Her değişken, (virgül işareti) ile ayrılmıştır.
 - En sondaki “class” adlı değişken, ikili sınıflandırmaya ilişkin sınıf değişkenidir. “good” etiketi pozitif (birincil öncelikli), “bad” etiketi negatif (ikincil öncelikli) olanları gösterir.
 - Eğitim setinde 750 adet, test setinde 250 adet kayıt bulunmaktadır.
- Karar ağacı algoritması olarak; **CART ya da C4.5 algoritmalarından sadece herhangi birisini öğrenciler kendileri kodlayacaktır.**
- **CART ve C4.5 algoritmalarına ilişkin hazır fonksiyon, kütüphane, hazır araç, vb. kullanımı kesinlikle yasaktır. Kullanılması durumunda ödev notundan 40 puan düşürülür.**
- **Test verisinin sınıflandırmasının Accuracy, True Positive Rate, True Negative Rate, True Positive Adedi, True Negative Adedi ölçümleri ve hesaplamalarını öğrenciler kendileri kodlayacaktır. Bu işlemler için hazır fonksiyon, kütüphane, hazır araç, vb. kullanımı kesinlikle yasaktır. Kullanılması durumunda ödev notundan 40 puan düşürülür.**
- Öte yandan dosya açma, veri okuma, eğitim sonunda oluşturulan karar ağacının görselleştirilmesi, grafik çizim, vb için hazır araç, kütüphane, vb. kullanabilirsiniz.
- Kodlama kısmında, **sadece aşağıdaki programlama dillerinden istediğiniz bir tanesini** seçip kullanabilirsiniz: **C, C++, C#, .Net, Java, Python.**

Veri setinde, öncelikle bazı çeşitli ve farklı veri düzenleme / düzeltme / temizleme (data pre-processing) işlemleri yapılmayacaktır. Gerek yoktur.

Program, önce eğitim dosyasını açarak karar ağacını oluşturacak, sonra da oluşan karar ağacı modeline göre, test veri setini kullanarak **Accuracy, True Positive Rate, True Negative Rate, True Positive Adedi, True Negative Adedi** ölçümlerinin sonucunu ekrana aşağıdakine benzer biçimde yazdıracaktır.

Test sonucu:

Accuracy: 0.753

TPrate: 0.908

TNrate: 0.322

TP adedi: 227

TN adedi: 29

Bu ödevde istenenlerin hatasız biçimde, yukarıda açıklanan kısıt ve koşullara uygun olarak yapılması, ayrıca aşağıdaki 2 seçeneğin de başarıyla yapılması durumunda ödevden en fazla 125 puan alınabilir.

- Eğitim sonunda oluşturulan karar ağacının ekrana çizimi, zorunlu değildir. İsteğe bağlıdır. Yapılması ve hatasız çalışması durumunda ek olarak 10 puan ile ödüllendirilir.
- Eğitim sırasında ağacı oluştururken ön-budama (pre-pruning) ya da sonradan-budama (post-pruning) işlemleri zorunlu değildir. Budama işleminin yapılması ve hatasız çalışması durumunda ek olarak 15 puan ile ödüllendirilir. (Budama yöntemi olarak, literatürde bilinen yöntemlerden istenilen seçilip kodlanabilir).

Ödevde Teslim Edilecekler:

1-Programın tüm kaynak kodları, bağlantılı kütüphane, dizinler, vb.

2-Kullanılan yöntemler, işlemler, vb. ile ilgili kısa bilgiler / notlar (kaynak kod içine kısa açıklamalar olarak eklenmelidir).