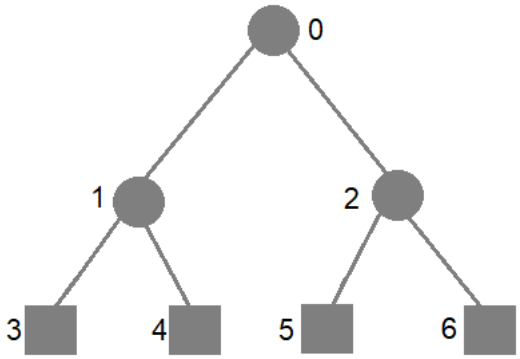
dosya olarak zipli dosya içerisinde gönderiniz. Kaynak kodunda ilgili yerlere gereken açıklamaları yazınız.

6) (75 puan)



Şekil 1: Karar Ağacının Yapısı

Şekil 1' deki yapısı ile gösterilen, belirli bir data topluluğunu sınıflandırmak amacı ile kullanılan karar ağaçları kök düğüm (0 numara), iç düğümler (1 ve 2 numara) ve yaprak düğümlerinden (3, 4, 5 ve 6 numara) oluşur. Kök düğüm ve iç düğümler daire, yaprak düğümleri ise kare ile gösterilmektedir. Datayı sınıflandırmak için karar ağacı öncelikli olarak eğitilir ve eğitimi tamamlandıktan sonra da aynı türden farklı datalar ile test edilir.

Kök düğümü karar ağacının başlangıç yeri olup hem eğitim hem de test esnasında datanın ilk gönderildiği düğümdür. Kök düğümle beraber iç düğümler dataları sola ya da sağa göndererek yaprak düğümlerine kadar ulaşmasını sağlarlar.

Şekil 1' de gösterilen karar ağacının eğitimi şu şekilde yapılmaktadır:

Bir karar ağacı etiketli eğitim dataları ile eğitilirler (Örneğin; bir öğrenci elmaları renklerine göre sınıflandıran bir sınıflandırıcı eğitmek istesin. Burada sarı (s), kırmızı (k) ve yeşil (y) renkler elmanın etiketidir). Eğitim dataları 0 numaralı düğüme (kök düğüm) beslenir ve kök düğüm (0 numara) ile iç düğümlerde (1 ve 2 numaralı düğümler) çeşitli sorulara n defa tabi tutularak sola ya da sağa gönderilir. Ancak bu gönderim esnasında maksimize edilmesi gerekilen bir "bilgi kazanımı" fonksiyonu I mevcuttur:

$$I = H(S) - \sum_{i\epsilon\{L,R\}} \frac{|S^i|}{|S|} H(S^i) \qquad (1)$$

$$H(S) = -\sum_{c\epsilon C} p(c)\log(p(c)) \qquad (2)$$

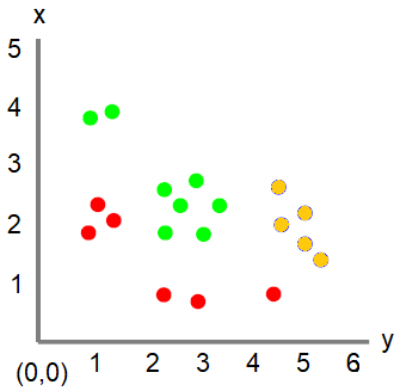
Bu fonksiyonda S eğitim datalarını, $|S|$ eğitim datalarının sayısını, S^L sola dallanan eğitim datalarını, $|S^L|$ sola dallanan eğitim datalarının sayısını, S^R sağa dallanan eğitim datalarını, $|S^R|$ sağa dallanan eğitim datalarının sayısını, c herhangi bir datanın etiketini, C eğitim datalarındaki tüm etiketleri, $H(S)$ S eğitim datalarının entropisini, $H(S^L)$ sola dallanan eğitim datalarının entropisini, $H(S^R)$ sağa dallanan eğitim datalarının entropisini, $p(c)$ ise c kategorisinin olasılığını, $\log(p(c))$ ise $p(c)$ nin $\log_2$ tabanında logaritmasını ifade etmektedir.

İlgili düğümde sorulan n sorudan hangisi bu "bilgi kazanımı" fonksiyonunu maksimize ederek eğitim datalarının sola ya da sağa gönderilmesini sağlıyorsa, o soru bu ilgili düğümde depolanır. Örneğin; sıfırıncı düğümde (0 numara) $n = 4$ için, sırayla birinci soru sorulur, tüm eğitim dataları sola ya da sağa gönderilir ve I fonksiyonu tüm eğitim dataları S, sola ya da sağa gönderilen eğitim dataları S^L ve S^R için hesaplanır. Ardından ikinci soru sorulur, tüm eğitim dataları yeniden sola ya da sağa gönderilir ve I fonksiyonu tüm eğitim dataları ve ikinci soru neticesinde sola ya da sağa gönderilen datalar için hesaplanır. Bu şekilde I fonksiyonu üçüncü ve dördüncü sorular için de hesaplanır. Hangi soru bu I fonksiyonu için maksimum değeri veriyorsa o soru bu kök düğümünde (0 numara) depolanır. I fonksiyonu için maksimum değeri veren sorunun sola ya da sağa gönderdiği eğitim dataları ile birinci ve ikinci düğümler sıfırıncı düğümün eğitildiği şekliyle eğitilir. Yaprak düğümleri ise kendilerine ulaşan datalar göz önünde bulundurularak ilgili yaprak düğümündeki dataların olasılığını **depolar**. Örneğin 3 numaralı yaprak düğümünde 3 sarı, 2 kırmızı ve 1 tane yeşil elma bulunsun. Dolayısı ile burada $C = 3$' tür ve:

$$p(s) = \frac{3}{6}, \qquad p(k) = \frac{2}{6}, \qquad p(y) = \frac{1}{6}$$

dır.

Örnek: Şekil 2' de bir masanın üzerinde duran 5 sarı, 6 kırmızı ve 8 yeşil elma gösterilmiştir. Bu masanın sol alt köşesi (0,0) olarak kabul edilmiştir ve sağa doğru (doğu yönü) ilerledikçe y koordinatı, yukarı doğru ilerledikçe (kuzey yönü) de x koordinatı artış göstermektedir. Sarı (s), kırmızı (k) ve yeşil (y) renkler bu elmaların etiketleridir. Bir öğrenci şu anki konumlarında (0,0) referansına göre duran elmaları kullanarak bir karar ağacı eğitmek istemektedir. Eğitilmesi istenen karar ağacı, kök düğüm ve iç düğümlerde (1) numaralı denklem ile verilen I fonksiyonunu maksimize eden soruları, yaprak düğümlerinde ise eğitim datalarının etiketlerinin (sarı, kırmızı, yeşil) olasılığı bilgisini depolayacaktır.



Şekil 2: Sol alt köşesi (0, 0) kabul edilen bir masanın üzerinde duran 5 sarı (s), 6 kırmızı (k) ve 8 yeşil (y) elma

Sıfırıncı düğümün eğitimi: Karar ağacının sıfırıncı düğümüne beslenen toplam 19 adet sarı, kırmızı ve yeşil elmaya sonsuz adet "eksen-hizalı" sorular içinden **rastgele** seçilecek 4 soru sorulacaktır. Örneğin rastgele seçilen sorular şu şekildedir:

1) Elmanın x pozisyonu 2.1 den küçük mü?

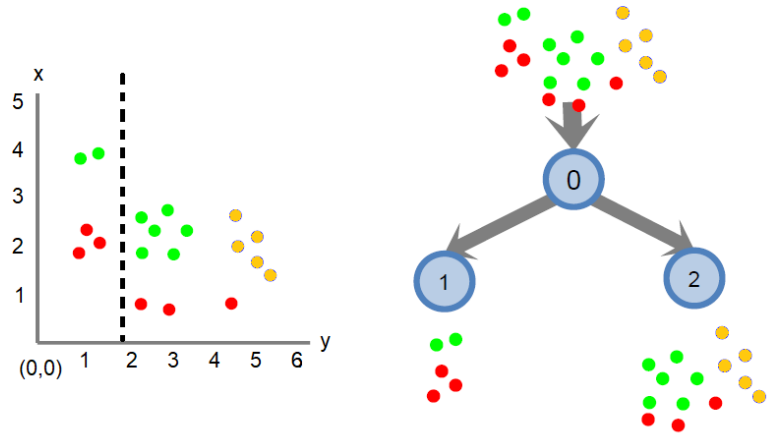
2) Elmanın x pozisyonu 3.3 ten küçük mü?

3) Elmanın y pozisyonun 5.6 dan küçük mü?

4) Elmanın y pozisyonu 1.9 dan küçük mü?

Birinci sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır, ikinci sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır, üçüncü sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R

kullanılarak I fonksiyonu hesaplanır, ve son olarak dördüncü sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır. Bulunan I değerleri birinci soru, ikinci soru, üçüncü soru ve dördüncü soru için farklı değerler üretmekte olup dördüncü soru için hesaplanan I değeri maksimumdur. Dolayısıyla sıfırıncı düğümde depolanan soru dördüncü sorudur (Bakınız Şekil 3).



Şekil 3: Elmanın y pozisyonu 1.9 dan küçük mü? sorusuna cevap olarak sola 3 kırmızı ve 2 yeşil, sağa 5 sarı, 3 kırmızı ve 6 yeşil elma gönderilmiştir.

Şekil 3' e göre 1 numaralı iç düğümde 3 kırmızı ve 2 yeşil olmak üzere toplam 5 elma, 2 numaralı iç düğümde de 5 sarı, 3 kırmızı ve 6 yeşil olmak üzere toplam 14 elma bulunmaktadır.

Birinci düğümün eğitimi: Şu an itibariyle 1 numaralı düğümde toplamda 3 kırmızı ve 2 yeşil elma bulunmaktadır. Bu 5 elmaya sonsuz adet "eksen-hizalı" sorular içinden rastgele seçilecek 4 soru sorulacaktır. Örneğin rastgele seçilen sorular şu şekildedir:

1) Elmanın x pozisyonu 4.3 ten küçük mü?

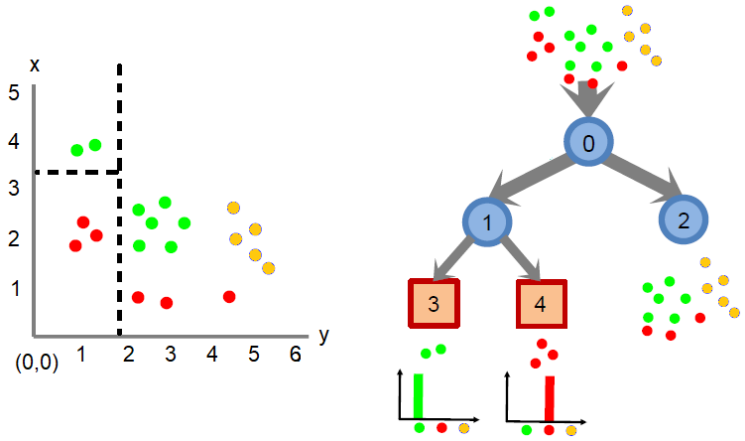
2) Elmanın y pozisyonu 0.3 ten küçük mü?

3) Elmanın x pozisyonu 3.3 ten küçük mü?

4) Elmanın y pozisyonu 1.9 dan küçük mü?

Birinci sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte, lütfen dikkat ediniz: artık birinci düğüm için tüm eğitim dataları S nin sayısı $|S|$=5' tir.) sola ya da sağa gönderilen

eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır, ikinci sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır, üçüncü sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır, ve son olarak dördüncü sorunun sorulması neticesinde (tüm eğitim dataları S ile birlikte) sola ya da sağa gönderilen eğitim dataları S^L ve S^R kullanılarak I fonksiyonu hesaplanır. Bulunan I değerleri birinci soru, ikinci soru, üçüncü soru ve dördüncü soru için farklı değerler üretmekte olup üçüncü soru için hesaplanan I değeri maksimumdur. Dolayısıyla birinci düğümde depolanan soru üçüncü sorudur (Bakınız Şekil 4). Şekil 4' e göre 3 numaralı yaprak düğümünde 2 yeşil elma, 4 numaralı yaprak düğümünde de 3 kırmızı elma bulunmaktadır. Görüldüğü gibi 3 numaralı yaprak düğümünde depolanan $p(c)$ oranları sarı, kırmızı, yeşil için $(p(s),\ p(k),\ p(y)) = (1, 0, 0)$, ve 4 numaralı yaprak düğümünde depolanan $p(c)$ oranları sarı, kırmızı, yeşil için $(p(s),\ p(k),\ p(y)) = (0, 1, 0)$ dır.



Şekil 4: Elmanın x pozisyonu 3.3 ten küçük mü? sorusuna cevap olarak birinci düğümden sola 2 yeşil, sağa 3 kırmızı elma gönderilmiştir. Bu şekle göre 3 ve 4 numaralı düğümler yaprak düğümleridir.

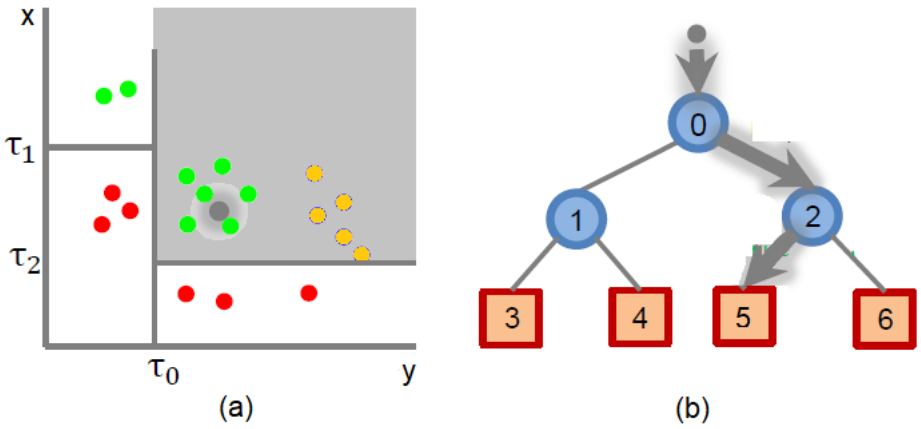
İki numaralı iç düğüm de aynı sıfırıncı ve birinci düğümde olduğu gibi eğitilecek ve bu eğitimin sonunda 5 numaralı ve 6 numaralı yaprak düğümleri oluşturulacaktır (Bu noktada lütfen Şekil 1' e bakınız).

Netice itibariyle karar ağacının eğitimi 5 sarı, 6 kırmızı ve 8 yeşil olmak üzere toplam 19 tane elma kullanılarak tamamlanmıştır ve iç düğümlerde I fonksiyonunu maksimize eden sorular

ve yaprak düğümlerinde ise ilgili yapraklara ulaşan eğitim datalarının olasılıkları depolanmıştır.

Eğitimi yapılmış olan karar ağacının test edilmesi:

Şekil 5 (a)' da test edilecek olan test datası, yani etiketini bilmediğimizi farzettiğimiz elma, gri taralı alanda gri yuvarlak ile gösterilmektedir. Şekil 5 (b)' de ise bu test edilecek elmanın eğitimi yapılmış olan karar ağacında izlediği yol gösterilmiştir. Test anında etiketi (rengi) bilinmeyen test datası (elma) sıfırıncı düğümden başlayarak eğitimi yapılmış olan karar ağacına beslenecektir. Sıfırıncı düğümde, eğitim anında depolanmış olan soruya tabi tutularak sola ya da sağa gönderilecektir. Şekilde gösterildiği üzere sağa gönderilmiştir (ikinci düğüme). İkinci düğümde, eğitim anında depolanmış olan soruya tabi tutulacak ve bu sorudaki neticeye göre sola ya da sağa gönderilecektir. Şekilde gösterildiği gibi sola (beşinci düğüme) gönderilmiştir. Beşinci düğümde (yaprak düğümü), eğitim anında elmaların olasılığı depolandığı için bu test elmasının yüzde kaç ihtimalle sarı, kırmızı ya da yeşil olduğu bilgisi öğrenilmiş olacaktır.



Şekil 5: (a) test datası (etiketi bilinmeyen elma) gri taralı alanda yuvarlak gri renkte gösterilmektedir. (b) Eğitimi yapılmış olan karar ağacına gönderilen test datasının (elma) izlediği yol

“egitimdatasi” isimli “.txt” uzantılı dosyada 19 satır 3 sütun bir eğitim data matrisi verilmiştir. Birinci sütun eğitim için kullanılacak her bir elmanın x koordinatını, ikinci sütun eğitim için kullanılacak her bir elmanın y koordinatını, ve üçüncü sütun ise her bir eğitim datasının (elmanın) etiketini göstermektedir. Üçüncü sütunda sadece 1, 2 ve 3 sayılarını görmektesiniz. 1 sayısı sarı (s), 2 sayısı kırmızı (k), 3 sayısı ise yeşili (y) temsil etmektedir. Bu “.txt” dosyası ile verilen eğitim datalarını kullanarak size teorisi ve örneği verilen karar ağacını önce eğitiniz. Karar ağacını eğitirken kök düğüm (0 numara) ve iç düğümler için (1 ve 2 numara) sorulacak eksen-hizalı soruların sayısını $n = 4$ alınız. Bu 4 sorunun her biri rastgele seçilecektir. Ardından şu pozisyonlarda bulunan elmalar için test ediniz:

Birinci_elma: (3.3, 4.3)

İkinci_elma: (1.3, 2.4)

Üçüncü_elma: (1.0,2.0)

Bu test elmalarının her birinin sarı, kırmızı ya da yeşil olma olasılığı $p(s),\ p(k),\ p(y)$ nedir?

Not: Sizin eğitmiş olduğunuz karar ağacında kök düğüm ve iç düğümlerde sorulan sorular rastgele seçildiği için sağa ve sola gönderilen dataların (elmaların) sayısı ve yaprak düğümlerinde depolanan olasılık değerleri farklı olmak durumundadır.