BURSA TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

BİM0101 – Hesaplamalı Biyolojiye Giriş
Bütünleme Sınavı

Ad&Soyad	**: CEVAP ANAHTARI**
Öğrenci Numarası	**:**

Akademik yıl : **2021-2022**

Dönem : **Bahar**

Tarih : **25 Temmuz 2022 – 15:00**

Sınav süresi : **90 dakika**

Öğr. görevlisi : **Dr. Öğr. Üyesi Ergün GÜMÜŞ**

Soru	1	2	3	4	5	Toplam
Puan	25	16	15	25	19	100
Not						

KURALLAR

- Sınava başlamadan önce Ad&Soyad ve Öğrenci numarası alanlarını doldurunuz.
- Sınav öncesinde ve süresince sınav gözetmenlerinin tüm uyarılarına uymanız gerekmektedir.
- Sınav öncesinde cep telefonlarınızı KAPATINIZ!
- Soruları yanıtlamak için sadece sınav kâğıdınızı kullanmanız gerekmektedir. Yanıtlarınız açık ve okunaklı olmalıdır.
- Sınav boyunca masanızın üzerinde bulunabilecek malzemeler sadece sınav kâğıdınız, kalem ve silgidir.
- Sınav süresince herhangi bir nedenle birbirinizle konuşmak ve malzeme (silgi, kalem, kâğıt vb.) alışverişi yasaktır.
- Bu kuralların herhangi birine uymamak kopya çekmeye yönelik bir hareket olarak değerlendirilir ve ilgili makamlara bildirilir.

1) [25p] A, B, C, D, E isimli beş adet canlı türü ve her canlıyı ifade etmek için de X_1, X_2, X_3, X_4, X_5 şeklinde beş adet karakter olduğunu, canlı-karakter matrisinin de aşağıdaki gibi olduğunu kabul edelim.

	X_1	X_2	X_3	X_4	X_5
A	6	3	4	7	5
B	7	7	6	5	5
C	6	2	8	4	8
D	5	7	1	5	8
E	5	2	9	3	6

Buna göre;

a) Canlı çiftleri arasındaki mesafeyi gösteren *D1* mesafe (distance) matrisini oluşturunuz. Mesafe hesabında öklid (euclidean) uzaklığını kullanınız [10p].

b) a şıkkında oluşturduğunuz *D1* mesafe matrisini kullanarak canlı türlerini aşağıdan yukarıya yaklaşımla kümeleyiniz. Kümeleme işleminin her adımında sadece birbirine en yakın olan iki kümeyi birleştirerek bir sonraki adımın mesafe matrisini (D2, D3, …) oluşturunuz. İki kümenin yakın olup olmadıklarını belirlerken bağlama (linkage) tekniği olarak tekil bağlama (single linkage) kullanınız. Puan alabilmeniz için kümeleme işleminin her adımını sonucuyla beraber göstermeniz gerekmektedir [10p].

c) Yaptığınız kümeleme işlemine karşılık gelen dendrogramı çiziniz [5p].

1) a)

D1 =

	A	B	C	D	E
A	0	5	5,9161	6,245	6,6332
B	5	0	6,3246	6,1644	6,5574
C	5,9161	6,3246	0	8,7178	2,6458
D	6,245	6,1644	8,7178	0	9,8489
E	6,6332	6,5574	2,6458	9,8489	0

C ve E birleşecek.

10 puan

b)

D2 =

	A	B	{C,E}	D
A	0	5	5,9161	6,245
B	5	0	6,3246	6,1644
{C,E}	5,9161	6,3246	0	8,7178
D	6,245	6,1644	8,7178	0

A ve B birleşecek

4 puan

D3 =

	{A,B}	{C,E}	D
{A,B}	0	5,9161	6,1644
{C,E}	5,9161	0	8,7178
D	6,1644	8,7178	0

{A,B} ve {C,E} birleşecek.

3 puan

D4 =

	{A,B,C,E}	D
{A,B,C,E}	0	6,1644
D	6,1644	0

3 puan

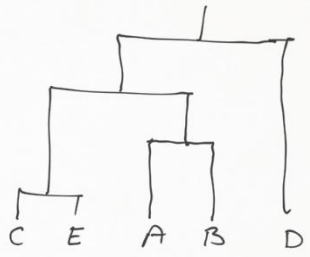
{A,B,C,E} ve D birleşecek.

D5 =

	{A,B,C,D,E}
{A,B,C,D,E}	0

puansız.

c)



5 puan

2) [16p] Aşağıda, iki sınıftan oluşan veri üzerinde çalışan sınıflandırma modelinin sınıflandırma sonuçları görülmektedir. Buna göre,

		Modelin tahmini	
		Pozitif	Negatif
Gerçek sınıf etiketi	Pozitif	23	27
	Negatif	41	9

a) True Positive (TP), False Positive (FP), True Negative (TN) ve False Negative (FN) değerlerini belirleyiniz [4p].
b) Modelin duyarlılığını (sensitivity) hesaplayınız [4p].
c) Modelin özgüllüğünü (specificity) hesaplayınız [4p].
d) Modelin hatalı tahmin oranını (false discovery rate) hesaplayınız [4p].

2) a) $TP=23,\ TN=9,\ FP=41,\ FN=27$ Herbiri 1 puan

b) $\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{23}{23+27} = \frac{23}{50} = 0,46$ 4 puan

c) $\text{Specificity} = \frac{TN}{TN+FP} = \frac{9}{9+41} = \frac{9}{50} = 0,18$ 4 puan

d) $FDR = \frac{FP}{FP+TP} = \frac{41}{41+23} = \frac{41}{64} = 0,6406$ 4 puan

3) [15p] G = 10^9 bp uzunluklu bir genomun sekanslandığını ve ortalama L = 1500bp uzunluklu, N = 9×10^5 adet klon (fragman) elde edildiğini kabul edelim. Buna göre aşağıdaki soruları cevaplayınız:

a) Bu sekanslama işleminin sonunda coverage (c) kaç olur [5p]?
b) Genomdan rastgele seçilen bir bazın sekanslama işlemi sonunda en az 1 kez <u>okunmuş olma</u> ihtimali kaçtır?, hesaplayınız [5p].
c) Klonları ucuca ekleyerek layout'lar oluşturmak istiyoruz. Bir klonun ucuyla diğer bir klonun ucunu eşleştirmek istersek bu eşleşmenin, en az klon uzunluğunun %10'u kadar olması gerektiğini varsayalım. Bu durumda kaç adet adacık (island) bulmayı bekleriz?, hesaplayınız [5p].

3) a) $c = \frac{N.L}{G} = \frac{9.10^5 . 15.10^2}{10^9} = \frac{135}{100} = 1,35$ 5 puan

b) $f_c = 1-(1-L/G)^N = 1-\left(1-\frac{15.10^2}{10^9}\right)^{9.10^5} = 0,7408$ 5 puan

c) $\theta = 0,1$

Beklenen adacık sayısı $= \Gamma = N.e^{-(1-\theta).c} = 9.10^5 . e^{-0,9.1,35} \cong 267039$ tane

5 puan

4) [25p] Bir canlının DNA'sının herhangi bir pozisyonundaki baz görülme olasılıkları şu şekilde olsun: P(A) = 0,27 P(C) = 0,30 P(G) = 0,21 P(T) = 0,22. Bu canlının DNA'sının herhangi bir bölgesindeki X=0 baz pozisyonunun 2 baz solu ve 2 baz sağını çevreleyen alandaki baz görülme olasılıkları ise şu şekilde verilmiş olsun:

	X=-2	X=-1	X=0	X=+1	X=+2
P(A)	0,18	0,31	0,37	0,15	0,14
P(C)	0,39	0,29	0,13	0,56	0,20
P(G)	0,26	0,19	0,19	0,13	0,38
P(T)	0,17	0,21	0,31	0,16	0,28

Buna göre;

a) X=-2 → X=+2 aralığındaki her baz pozisyonu için bağıl entropi değerlerini hesaplayınız [15p].

b) Canlının X=-2 → X=+2 aralığını ifade eden sekans logosunu oluşturunuz (Logoyu çizmeniz gerekmez, aralıktaki her baz poziyonu için o pozisyondaki baz logolarının yüksekliğini hesaplamanız yeterlidir) [10p].

4) a)

$H = \sum_{a \in \{A,C,G,T\}} p_a \cdot \log_2(p_a/q_a)$

3p

$X=-2 \rightarrow H = [0,18 \cdot \log_2(0,18/0,27) + 0,39 \cdot \log_2(0,39/0,3) + 0,26 \cdot \log_2(0,26/0,21) + 0,17 \cdot \log_2(0,17/0,22)] = 0,0592$

$X=-1 \rightarrow H = [0,31 \cdot \log_2(0,31/0,27) + 0,29 \cdot \log_2(0,29/0,3) + 0,19 \cdot \log_2(0,19/0,21) + 0,21 \cdot \log_2(0,21/0,22)] = 0,0061$ 3p

$X=0 \rightarrow H = [0,37 \cdot \log_2(0,37/0,27) + 0,13 \cdot \log_2(0,13/0,3) + 0,19 \cdot \log_2(0,19/0,21) + 0,31 \cdot \log_2(0,31/0,22)] = 0,1373$ 3p

$X=+1 \rightarrow H = [0,15 \cdot \log_2(0,15/0,27) + 0,56 \cdot \log_2(0,56/0,3) + 0,13 \cdot \log_2(0,13/0,21) + 0,16 \cdot \log_2(0,16/0,22)] = 0,2136$ 3p

$X=+2 \rightarrow H = [0,14 \cdot \log_2(0,14/0,27) + 0,2 \cdot \log_2(0,2/0,3) + 0,38 \cdot \log_2(0,38/0,21) + 0,28 \cdot \log_2(0,28/0,22)] = 0,1729$ 3p

4) b)

L yüksekliği ifade etmek üzere,

$X=-2 \rightarrow L(A) = 0,18 \cdot 0,0592 = 0,0107$
$L(C) = 0,39 \cdot 0,0592 = 0,0231$
$L(G) = 0,26 \cdot 0,0592 = 0,0154$
$L(T) = 0,17 \cdot 0,0592 = 0,0101$

$X=-1 \rightarrow L(A) = 0,31 \cdot 0,0061 = 0,0019$
$L(C) = 0,29 \cdot 0,0061 = 0,0018$
$L(G) = 0,19 \cdot 0,0061 = 0,0012$
$L(T) = 0,21 \cdot 0,0061 = 0,0013$

$X=0 \rightarrow L(A) = 0,37 \cdot 0,1373 = 0,0508$
$L(C) = 0,13 \cdot 0,1373 = 0,0178$
$L(G) = 0,19 \cdot 0,1373 = 0,0261$
$L(T) = 0,31 \cdot 0,1373 = 0,0426$

$X=+1 \rightarrow L(A) = 0,15 \cdot 0,2136 = 0,032$
$L(C) = 0,56 \cdot 0,2136 = 0,1196$
$L(G) = 0,13 \cdot 0,2136 = 0,0278$
$L(T) = 0,16 \cdot 0,2136 = 0,0342$

$X=+2 \rightarrow L(A) = 0,14 \cdot 0,1729 = 0,0242$
$L(C) = 0,2 \cdot 0,1729 = 0,0346$
$L(G) = 0,38 \cdot 0,1729 = 0,0657$
$L(T) = 0,28 \cdot 0,1729 = 0,0484$

Hesbiri 0,5 puan

5) [19p]
a) K-means kümeleme algoritmasını işlem adımlarıyla beraber sözde kod halinde yazınız [13p].
b) Algoritmada ideal K değerini seçmek için hangi yaklaşımı uygulamamız gerektiğini anlatınız [6p].

a)
Adım1: Başlangıç.

Adım2: Kabul edilebilir olan bir kümeiçi varyans değeri belirle.

Adım3: İstenilen küme adedini simgeleyen K değerini seç.

Adım4: Eldeki örneklerden rastgele K tanesini K kümenin merkez noktası olarak seç.

Adım5: Kalan örneklerin her birini, merkez noktalarına olan uzaklığa göre en yakın kümeye dahil et.

Adım6: Kümelerin merkezlerini yeniden hesapla.

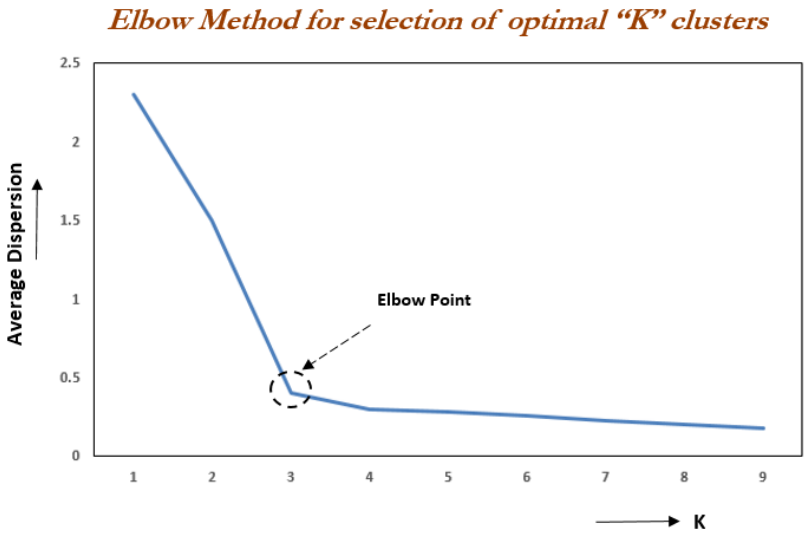
Adım7: Kümelerin toplam kümeiçi varyansını hesapla.

Adım8: Eğer kümeleme işleminin sonucu bir önceki denemenin sonucundan farklı ise Adım4'e dön.

Adım9: Eğer toplam kümeiçi varyans kabul edilebilir bir değer değil ise Adım4'e dön.

Adım10: Bitiş.

b) Denemelerde kullanılan her K değeri için bir toplam kümeiçi varyans değeri elde edilir. K arttıkça bu kümeiçi varyansın düşmesi beklenir ve aşağıdakine benzer bir grafik[1] elde edilir.



K arttıkça, bir noktada toplam kümeiçi varyans'daki hızlı düşüş kırılacak ve yataya yakın bir seyre girecektir. Kırılmanın olduğu bu noktaya "elbow point" (dirsek noktası) adı verilir. Bu nokta, ideal K değerinin belirlenmesi için kullanılır. Grafikteki örnek için ideal K değeri 3'dür.

[1]: https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml