JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -2 Makeup EXAMINATION- April 2018

B.Tech (Bioinformatics) VI Semester

COURSE CODE: 16B11BI612   MAX. MARKS: 25

COURSE NAME: Datawarehousing and Mining for Bioinformatics

COURSE CREDITS: 4   MAX. TIME: 1.5 Hrs

Note: All questions are compulsory. Carrying of mobile phone during examinations will be treated as case of unfair means.

1. Narayana et al. derived the following rules for the prediction of cleavage site in HIV protease. (Their dataset consisted of 114 cleavage substrates and 249 non-cleavage substrates.) (7)
   (a) If position 4 is Phenylalanine then cleavage (35/5)
   (b) If position 4 is Leucine then cleavage (38/9)
   (c) If position 4 is Serine then non cleavage (26/1)
   (d) If position 4 is Tyrosine and position 5 is Proline then cleavage (32/5)
   (e) If position 6 is Glutamate then cleavage (44/8)

   Answer the following w.r.t the above.

   (a) Is it possible to calculate the overall accuracy of the above method using the data given? Justify and calculate it if it is possible.
   (b) Explain the scientific rationale behind developing this method.
   (c) Besides prediction of cleavage and non-cleavage, what are the other inferences from this study?

2. Describe the application of identification trees for the prediction of (5)
   (i)   Protein secondary structure
   (ii)   HIV protease cleavage site

3. Explain the meaning of significant and minor rules with respect to Li's committee approach of decision trees .How is this approach better than using single decision trees. (5)

4. Elaborate how the deterministic split of data is a limitation of identification trees. How can it be overcome? (4)

5. Why do we perform clustering on samples as well as features? (1)
6. Describe the various approaches used for clustering. (3)