Partial Exam: Statistical Learning- Bioinformatics (2023-24)

Questions (Total 7.5 Pnt)

1) Explain in your own words what is a feature vector and a feature table (0.5 Pnt)

2) Explain what a decision boundary is in the context of classifiers. (0.5 Pnt)

3) Explain the roles of internal and external validation sets in predictive model development. (0.5 Pnt)

4) Name two figures of merit for regression problems. One of them should be robust to outliers. Explain the reason why. (0.5 Pnt)

5) A data analyst is doing supervised feature selection with all the dataset. Then he does a data partition to develop the classifier and test it. Is this methodologically correct? Yes/No. Motivate your answer. (0.5 Pnt)

6) Explain the basic differences between a clustering using k-means and a clustering based on a mixture of Gaussian distributions (0.5 Pnt).

7) What is the "curse of dimensionality" phenomenon? How does this relate to the overfitting of a model? (0.5 Pnt)

8) According to the session of unsupervised clustering, and as it is summarized in the article "*K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data*", which are the limitations of the K-means clustering? (0.5 Pnt)

9) Implement the MISSING parts of the pseudocode described in the article "*K-means clustering algorithms […]*" (1 Pnt)

Algorithm 1: Standard K-means clustering algorithm pseudocode

```
Input:    Array X{x1, x2, ···, xn} // Dataset to be clustered
Output:   MISSING

1.        A set of k clusters
2:        // Initialize Parameters
3:        X = {x1, x2, ···, xn}
4:        C = {c1, c2, ···, ck}
5:        Repeat
6:        //Distance calculations
7:          fori = 1 to n do
8:          forj = 1 to k do
9:            Compute the Euclidean distance from a data object to all cluster
10:         end j
11:         //Data object assignment
12:         Add data objects to the closest cluster
13:       end i
14:       //Update cluster centroid
15        MISSING
16:       Until the difference between the cluster centroids of two consecutive iterations remains the same
          End
```

10) As described in the session "Feature selection" and according to the paper "*A review of feature selection techniques in bioinformatics*", which are the objectives of a feature selection approach? (0.5 Pnt)

11) Which are the different taxonomies/strategies of feature selection techniques? (0.5 Pnt)

12) According to the paper "*A review of feature selection techniques in bioinformatics*", which is one of the applications of feature selection for sequence analysis? (0.25 Pnt) And for microarrays (0.25 Pnt)?

13) Which are the minimal steps that any algorithm based on the evolutionary paradigm should implement? (0.5 Pnt)

14) What is the purpose of the EM algorithm in the mixture of Gaussian distributions approach) (0.25 Pnt)

15) What is the difference between a partitional algorithm and a hierarchical algorithm in unsupervised clustering? (0.25 Pnt)