Computer Science 3202/6915
Extra Assignment 3 – Explore the impact of data encodings

Due date: Friday April 5th by 11:30pm.

Learning goals:

1. Practice applying various data encodings/representations.
2. Practice using cross-validation to select the most appropriate encoding.
3. Be aware of the impact of data encodings in classifier's performance.

Instructions:
In the Brigthspace folder for this extra assignment 3, there is a training dataset available (extraA3_Data.tsv). This dataset consists of a single string (DNA sequence) attribute and one categorical label for 872 instances. The dataset is given as a tab-delimited text files with one instance per line. There is **not** a column header.

Your job is to work with this data to find a suitable data encoding to generate a classification model.

1. This extra assignment can be done individually or in groups of two students. You need to self-enroll in a group in Brightspace (even if you will work alone) before submitting your assignment.
2. Read the input data. [Suggest to explore the data: what's the class distribution?]
3. Generate three different data representations/encodings of the single string attribute. Note that the strings are of different length across the instances (instances have different number of characters per string) and thus you would like to use encodings that generate the same number of features irrespective of the string length.
    1. You can consider using k-mer frequencies (also called nucleotide composition). Different values of k will count as different encodings (i.e., if you use di-nucleotides (k=2) and tri-nucleotides (k=2) that's two encodings). Check the scikitbio function DNA.kmer_frequencies https://scikit.bio/docs/latest/generated/skbio.sequence.DNA.kmer_frequencies.html. If you use this function make sure to set relative = True.
    2. You might also want to explore the Python package MathFeature (https://github.com/Bonidia/MathFeature/blob/master/documentation/descriptors.md) that can be used to obtain several DNA sequence encodings.
4. Choose a tree-based ensemble classification method (such as those listed in this page https://scikit-learn.org/stable/modules/ensemble.html)
5. Use 10-fold stratified cross-validation (CV) to assess model performance (choose the performance metric from among these: https://scikit-learn.org/stable/modules/model_evaluation.html#scoring-parameter) of your chosen ML method with the different encodings from step 2 as input. Use the same hyper-parameters for all the encodings, and remember to use the same cross-validation partitions for all the encodings.
6. Generate the precision-recall curves of the CV-performance of the classifier per encoding.
7. Obtain the mean average-precision-score and F1-score and their standard deviation per encoding.

Computer Science 3202/6915
Extra Assignment 3 – Explore the impact of data encodings

Submission:
Submit through Brightspace the following:

a) Your python code in a single file with instructions on how to run your code.
b) A PDF file containing:
    1. short and clear description and justification of the data encodings evaluated,
    2. short and clear description of the tree-based ensemble method used,
    3. short and clear justification of the performance metric used to assess model performance,
    4. a table with the mean average-precision-score and F1-score and their standard deviation per encoding indicating the best encoding found,
    5. the figures your created in step 5 of the instructions, and
    6. an acknowledgement section listing your collaborations and sources.
c) If done in a group, submit a PDF listing the contributions of each team member and describing how did you work as a team for this assignment – this PDF must be signed by both group members.