Machine Learning in Computational Biology (Fall 2023)

Assignment #3

Deadline: 23:59:59 23th December, 2023 (Delayed submission is not allowed for any reason)

Purpose: to enhance the learning outcomes for the topics in "**Protein Sequence Modeling**" , "**Supervised Learning Methods**" and "**Performance Evaluation**".

After the removal of homologous sequences in both positive and negative dataset using CD-Hit, in order to carry out a binary classification between ubiquitination and non-ubiquitination sites, please accomplish the tasks described as follows.

1. [Performance Comparison of Different Feature Encoding Methods] (50 points)

You have learning a couple of feature encoding methods, such as One-hot encoding, Amino acid composition (AAC), Amino acid pair composition (AAPC), Positional Weighted Matrix (PWM), Position-specific scoring matrix (PSSM) and BBLOSUM62, from this course. According to the k-fold cross-validation (**k should be larger than five**), please do the performance comparison between the models trained using different feature encoding methods (**at least five feature types**), based on a supervised learning method (e.g. support vector machine). The comparison results should be provided in terms of a table and bar-charts, as the example shown below.

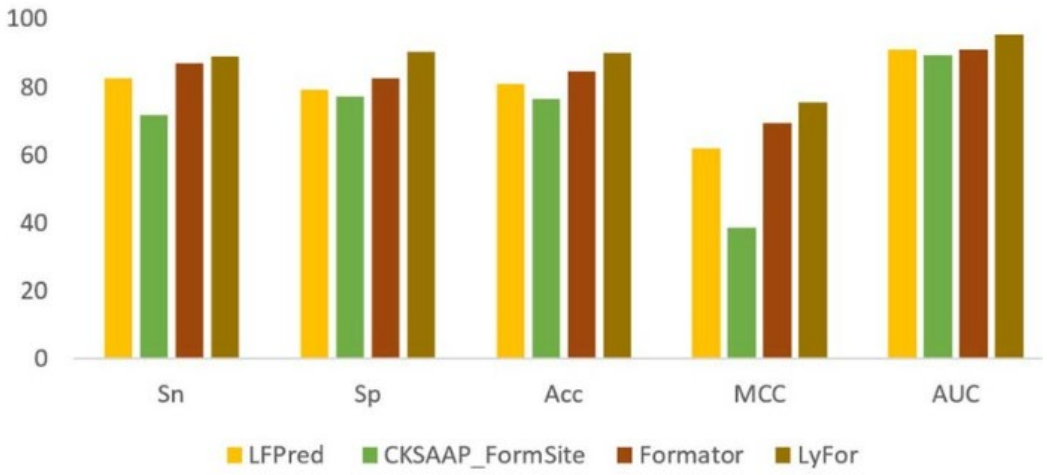
You can use:
either (1) Balanced Data
or (2) Imbalanced Data

Table 2. Five-fold cross validation results for SVM models trained with various features individually. A total of 1145 positive data and 8368 negative data were used in the cross validation process. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews Correlation Coefficient.

Training features	Sn	Sp	Acc	MCC
20D Binary code	0.66	0.68	0.68	0.23
BLOSUM62	0.68	0.70	0.69	0.26
Amino Acid Composition (AAC)	0.64	0.65	0.65	0.19
Amino Acid Pair Composition (AAPC)	0.64	0.67	0.67	0.21
Accessible Surface Area (ASA)	0.60	0.61	0.61	0.14
Secondary structure (SS)	0.56	0.56	0.56	0.08
Position Weight Matrix (PWM)	0.64	0.66	0.66	0.20
Position-specific scoring matrix (PSSM)	0.71	0.72	0.72	0.30

how to choose features

how to choose models



(Different feature encoding methods)

2. [Performance Comparison of Different Supervised Learning Methods] (50 points)

After the performance evaluation of different feature encoding methods, you can then try to consider different supervised learning methods (**at least five classifiers**) into the construction of predictive models using the best feature or hybrid features, with an attempt to identify the best model. The comparison results should be provided in terms of a table, bar-chart representation, and ROC curves, as the example shown below.