Task2WriteUp

2024-12-09

Task 2: Feature Selection

Abstract Task 2

The focus of the following analysis is to attempt to achieve high classification accuracy, while using a minimal number of features on very highly dimensional data. We will use a combination of six different feature selection and classification techniques to accomplish the goal. The models and methods used are a combination of methods taught in our ST443 Machine Learning and Data Mining Course taught at the London School of Economics and Political Sciences, and other techniques used from personal research.

Data Description

We were given one dataset that contains binary features that describe the three-dimensional properties of a molecule in a compound or a random probe across different compounds. The data contains 800 observations, which represent 800 compounds, and 100,000 columns (50,000 real variables and 50,000 random probes). The first column named *label* represents whether a compound bound to a target site on thrombin[1].

Exploratory Data Analysis

Models

Lasso The first model we attempted to run was logistic classifier with a Lasso-Penalty term. The lasso penalty term is a common regularization technique used for feature selection as it shrinks some of the coefficient estimates to be exactly equal to zero and removes them from the model. The ability to shrink coefficients to zero depends on the magnitude of the tuning parameter, *lambda*, which we tune in our model. #### Forward Stepwise Selection (FSS) #### Random Forest #### Elastic Net #### XGBoost #### Support Vector Machines (SVM)

Results

Model	Num. of Features	Bal. Accuracy	Accuracy	F1
Logistic Lasso	**47**	0.8189	0.925	0.6
Lin. Lasso	53	0.8189	0.925	0.6
FSS	208	0.7805	0.9188	0.5517
Random Forest	100	0.8223	0.9313	0.6207
Log. Elastic Net.	88	**0.8608**	**0.9375**	**0.6667**
Lin. Elastic Net	57	0.8223	0.9312	0.6207
XGBoost	73	0.8257	**0.9375**	0.6429

[1] Thrombin is an enzyme that plays a key role in blood clotting and other biological processes.

Model	Num. of Features	Bal. Accuracy	Accuracy	F1
SVM (Linear Kernel)	1	0.8257	0.9375	0.6429

Refrences

Guyon, I., Gunn, S., Ben-Hur, A. and Dror, G. (2004) Result Analysis of the NIPS 2003 Feature Selection Challenge. In Advances in Neural Information Processing Systems (Saul, L., Weiss, Y. and Bottou, L., eds.), vol. 17, MIT Press.

Shlobin NA, Har-Even M, Itsekson-Hayosh Z, Harnof S, Pick CG. Role of Thrombin in Central Nervous System Injury and Disease. Biomolecules. 2021 Apr 12;11(4):562. doi: 10.3390/biom11040562. PMID: 33921354; PMCID: PMC8070021.