Project - I Report

ECE 880 : Machine Learning - II

G.S.Venkatesh (UIN: 01095565)

October 15, 2018

1 A Brief about the Work

The assignment was to train a Support Vector Machine that classify the three different datasets,

- SELDI Spectrum data of Ovarian Cancer
- MADELON data
- SMK CAN 187 data

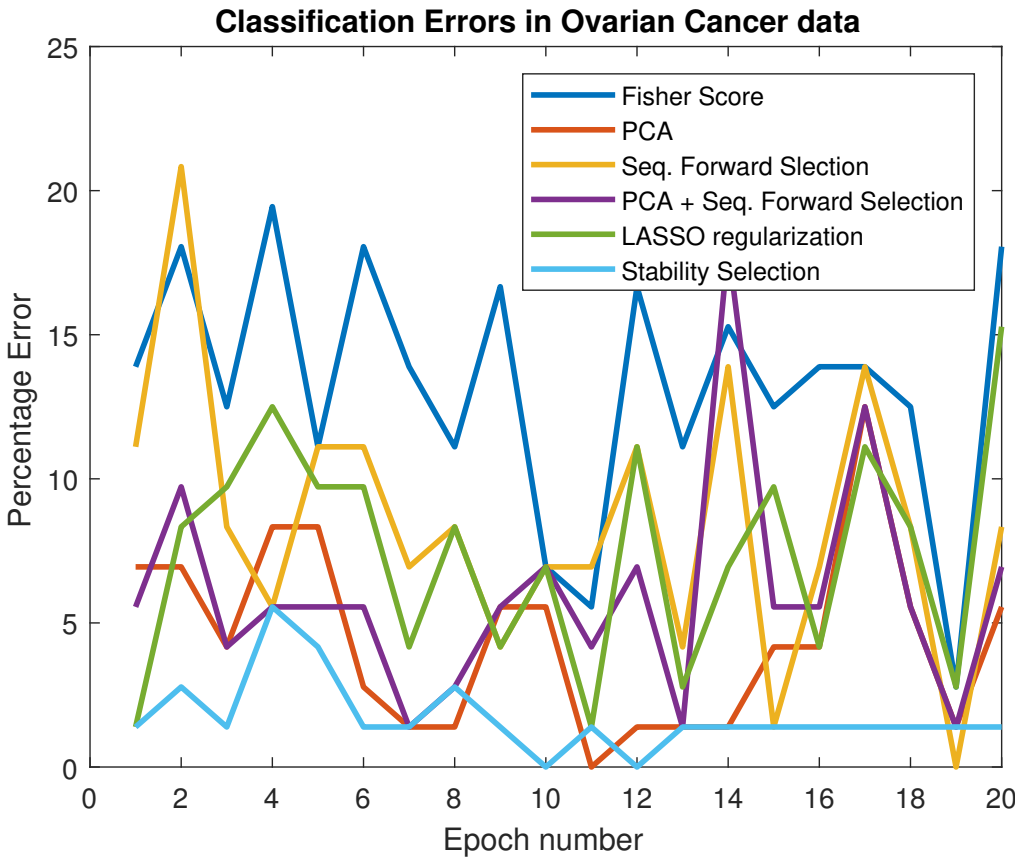
Prior to classification, a feature selection or dimensionality reduction was carried out by six different methods:

- Fisher Score
- Principal Component analysis
- Sequential Forward Selection
- PCA and then, Sequential Forward Selection
- LASSO Regularization
- Stability Selection

A $\frac{1}{3}^{rd}$ of the total dataset is partitioned as a Holdout set, each time the above feature selection and classification is carried out. A 20 such epochs were carried out and the statistics of the accuracy (Percentage Error) has been tabulated, comparing the above mentioned feature selection procedures.

2 SELDI Spectrum Data of Ovarian Cancer

For the Ovarian Cancer data, the LASSO regularization has been carried out with $\lambda = 0.3$ and the threshold probability for stability selection is 0.75, with a subsample size of 5 for 100 values of $\lambda \in (0, 1)$.

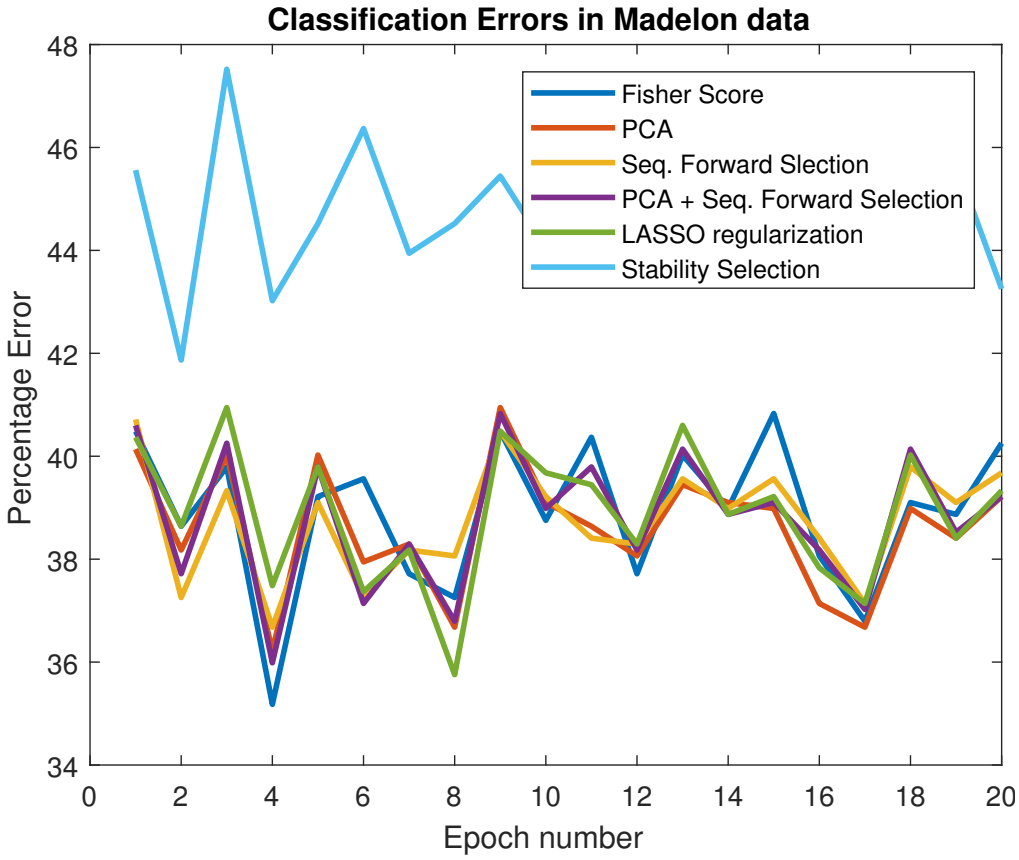


Selection Method	Mean Percentage Error	Standard Deviation
Fisher Score	13.19	4.26
PCA	4.44	3.17
Sequential Forward Selection	8.47	4.68
PCA + Seq. Forward Selection	6.04	3.88
LASSO Regularization	7.43	3.88
Stability Selection	1.74	1.26

Table 1: Ovarian Cancer data statistics (over 20 epochs)

3 Madelon data

For the Madelon data, the LASSO regularization has been carried out with $\lambda = 0.1$ and the threshold probability for stability selection is 0.6, with a subsample size of 5 for 100 values of $\lambda \in (0, 1)$.

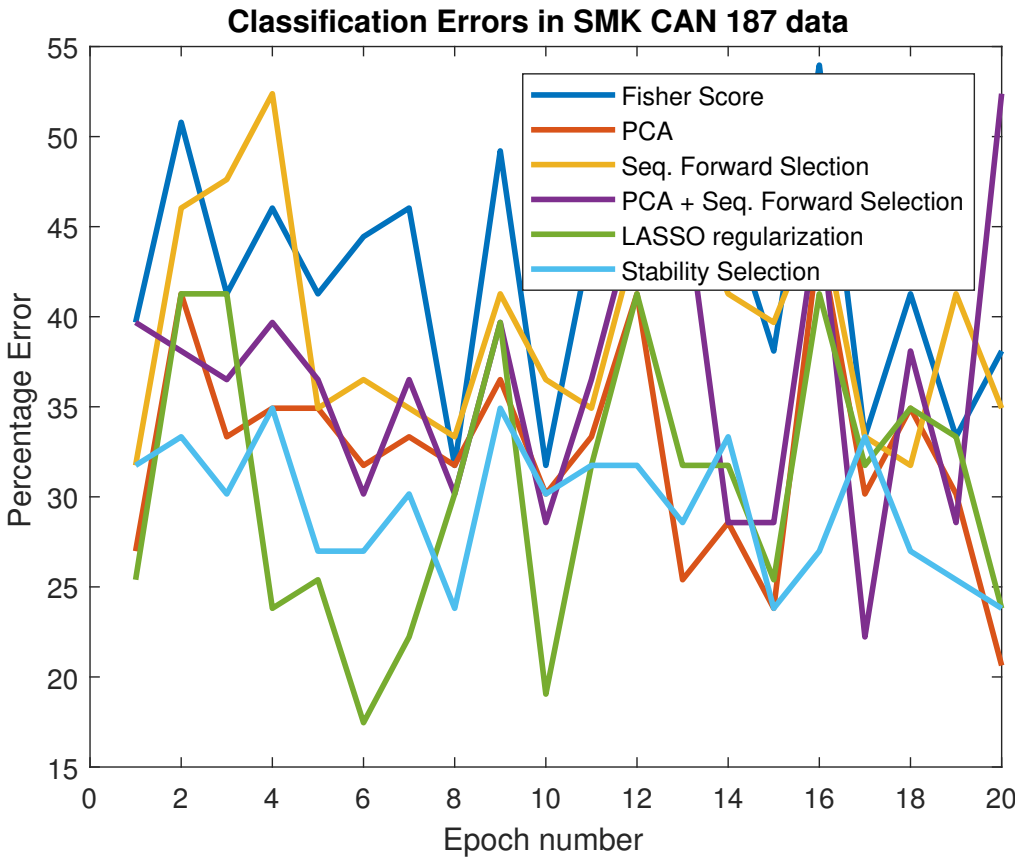


Selection Method	Mean Percentage Error	Standard Deviation
Fisher Score	38.90	1.45
PCA	38.61	1.26
Sequential Forward Selection	38.76	1.11
PCA + Seq. Forward Selection	38.77	1.37
LASSO Regularization	38.89	1.35
Stability Selection	44.80	1.26

Table 2: Madelon data statistics (over 20 epochs)

4 SMK CAN 187 Data

For the SMK CAN 187 data, the LASSO regularization has been carried out with $\lambda = 0.1$ and the threshold probability for stability selection is 0.75, with a subsample size of 5 for 100 values of $\lambda \in (0, 1)$.



Selection Method	Mean Percentage Error	Standard Deviation
Fisher Score	41.83	6.21
PCA	32.38	5.90
Sequential Forward Selection	39.52	6.13
PCA + Seq. Forward Selection	36.51	7.71
LASSO Regularization	30.63	7.68
Stability Selection	29.44	3.70

Table 3: SMK CAN 187 data statistics (over 20 epochs)