IE 506: Machine Learning: Principles And Techniques

PROJECT: MULTI-CLASS FEATURE SELECTION VIA SPARSE SOFTMAX WITH A DISCRIMINATIVE REGULARIZATION

TEAM: OUTLIERS

SAYANTAN MAITI (24M1503), IEOR (MTech 1st Year)

AISHWARYA JAISWAL (24M1511), IEOR (MTech 1st Year)

OUTLINE OF PRESENTATION

PROBLEM RECAP & OBJECTIVES

- Many real-world applications such as text categorization, face recognition, handwritten digit recognition, and gene detection involve high-dimensional data.

- These datasets suffer from the ”curse of dimensionality,” as they often contain many features that are irrelevant or redundant,

- As the number of features increases, **computational cost also increases** which can also lead to **overfitting** and **poor generalization.**

PROBLEM RECAP & OBJECTIVES

Existing multi-class feature selection (MFS) methods face three critical limitations:

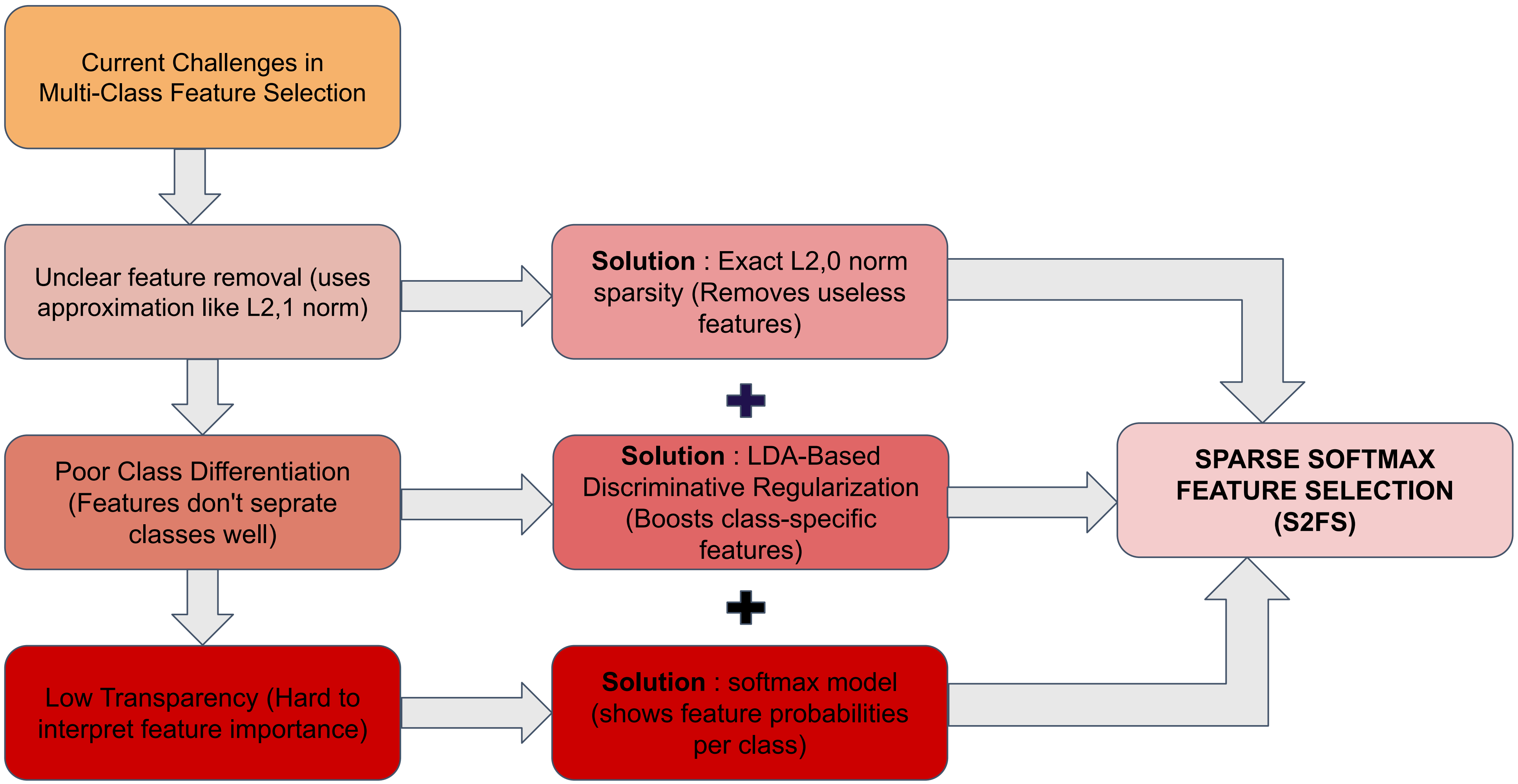
Ineffective Sparsity: Reliance on approximations (e.g., $L_{2,1}$ -norm) fails to achieve precise feature elimination, leading to redundant or noisy features.

Poor Discriminative Power: Features lack class-separability due to inadequate regularization, harming multi-class classification accuracy.

Weak Interpretability: Simple regression-based models obscure probabilistic insights into feature relevance across classes.

The proposed framework addresses these gaps by integrating exact sparsity, discriminative regularization, and probabilistic modeling for robust and interpretable MFS.

APPROACH PROPOSED



SUMMARY OF STAGE-1 WORK

During Stage 1, we focused on:

- **Paper Analysis:**
  - Studied the paper's motivation, mathematical formulation of S2FS, ADMM optimization algorithm, and experiments.
  - Reviewed 2 related works to understand prior state-of-the-art methods.

- **Algorithm Understanding:**
  - Analyzed the integration of $\ell_{2,0}$-norm regularization with Softmax and LDA-based discriminative regularization.
  - Explored ADMM's role in solving the non-convex problem via auxiliary variables.

- **Implementation Prep:**
  - Investigated datasets used in the paper.
  - Initiated partial algorithm recreation using AI tools for coding.

Outcome: Solid theoretical groundwork for implementing S^2FS in Stage 2.

SUMMARY OF STAGE 1 WORK

The optimization problem is **non-convex** due to L2,0-norm regularization. To solve this, we use an optimization algorithm based on **Alternating Direction Method of Multipliers (ADMM)**

Problem Reformulation:

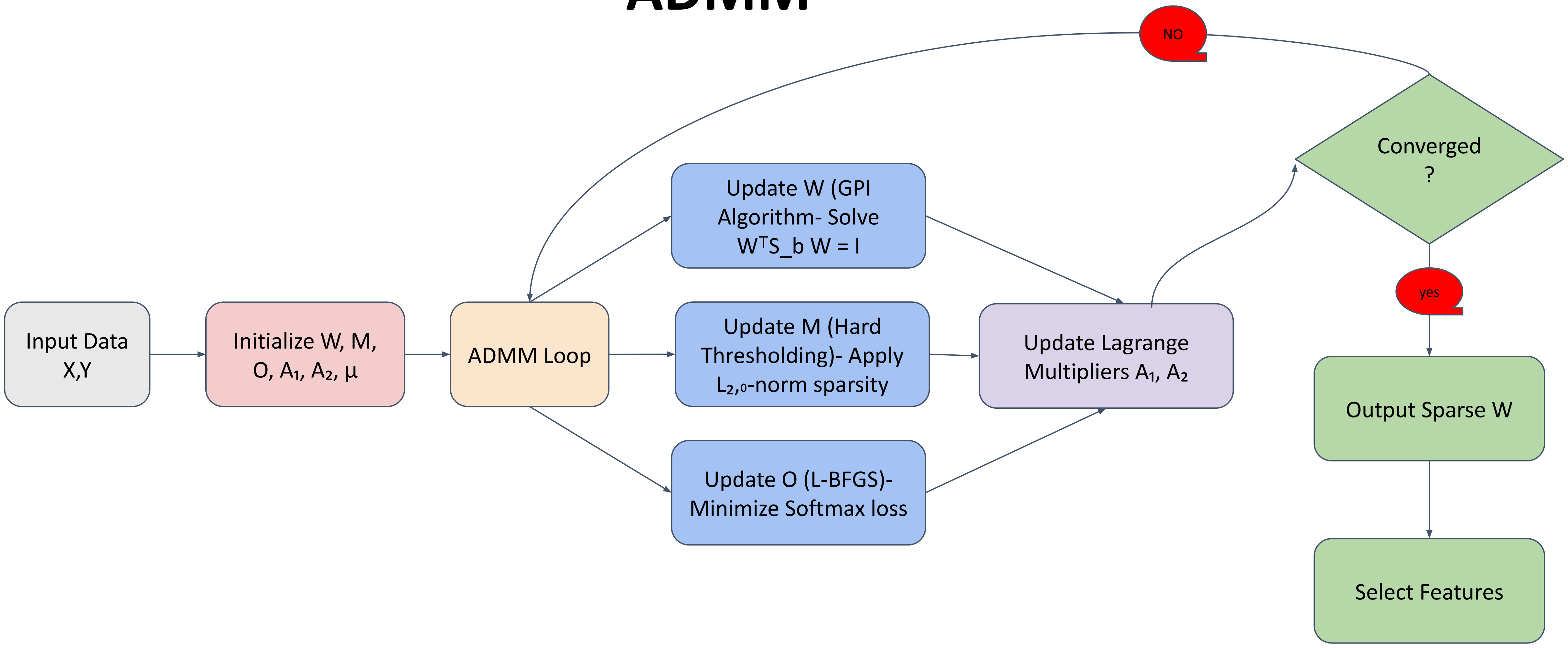
$$\min_{W,M,O,W^TS_bW=I} \mathcal{L}(O) + \alpha tr\left(W^TS_wW\right) + \lambda\|M\|_{2,0}$$
$$s.t. \quad W = M$$
$$W = O,$$

Constrained Optimization Problem
(using Auxiliary variables)

$$\min_{W,M,O,W^TS_bW=I} \mathcal{L}(O) + \alpha tr\left(W^TS_wW\right) + \frac{\mu}{2}\|W - M + \frac{\Lambda_1}{\mu}\|_F^2 + \frac{\mu}{2}\|W - O + \frac{\Lambda_2}{\mu}\|_F^2 + \lambda\|M\|_{2,0},$$

Unconstrained Optimization Problem
(Augmented Lagrangian Function)

ADMM


Input Data X,Y
Initialize W, M, O, A_1, A_2, μ
ADMM Loop
Update W (GPI Algorithm- Solve $W^T S\_b W = I$
Update M (Hard Thresholding)- Apply $L_{2,0}$-norm sparsity
Update O (L-BFGS)- Minimize Softmax loss
Update Lagrange Multipliers A_1, A_2
Converged ?
NO
yes
Output Sparse W
Select Features

COMMENTS GIVEN DURING STAGE-1

During the Stage-1 review, the following comments and suggestions were provided by the instructor and TAs:

- As no code was given in the original paper, we were advised to implement the entire S2FS algorithm from scratch, including all key components such as the Softmax model with $\ell 2,0$-norm regularization, the discriminative regularization term, and the ADMM-based optimization strategy.

- Our initial implementation produced suboptimal results. We were instructed to perform hyperparameter tuning using cross-validation techniques to improve model performance.

- We were asked to implement the S2FS method on 5 datasets out of the 15 datasets used in the original paper.

ADDRESSING COMMENTS

- Developed Python implementation of S^2FS with:
    - $\ell_{2,0}$-norm sparsity
    - ADMM optimization
    - Discriminative regularization
- Conducted hyperparameter tuning (λ, α) using paper-specified ranges
- Validated on 5 datasets:
    - Madelon, MNIST, Semeion, Musk, Lung

POST STAGE-1 WORK

1. Implementation

Integrated Softmax classifier with $\ell_{2,0}$-norm sparsity and LDA regularization.
Optimized via ADMM with GPU-accelerated matrix ops (CuPy).

2. Key Steps

ADMM Variables: Updated weights (`W`), sparsity (`M`), and Softmax outputs (`O`).
Hyperparameters: Tuned `λ` (sparsity) and `α` (LDA weight) via grid search.

3. Evaluation

5-fold CV: KNN (`k=5`) for validation accuracy.
Results: Matched original paper’s accuracy trends.
Ablation: Confirmed LDA’s impact in the Accuracy.

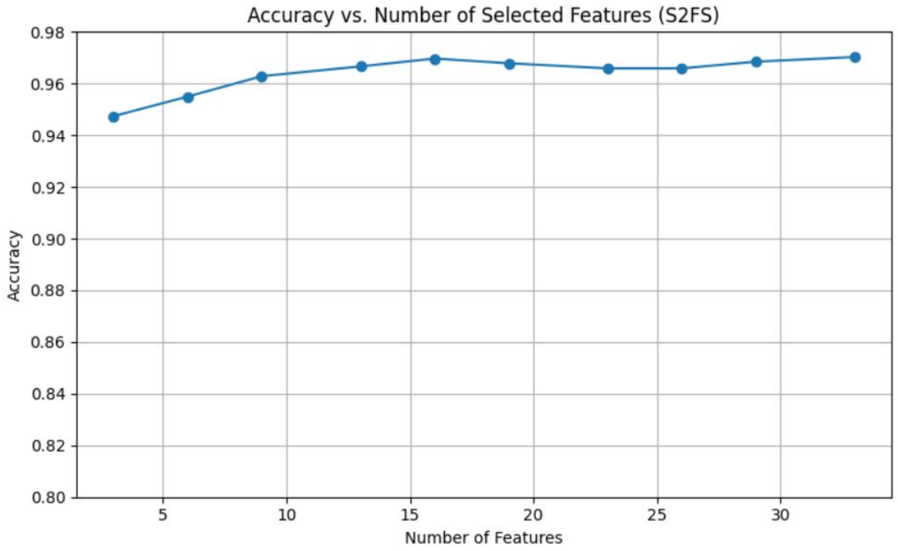
4. Outcome

Successfully replicated S^2FS, validating its feature selection efficacy.

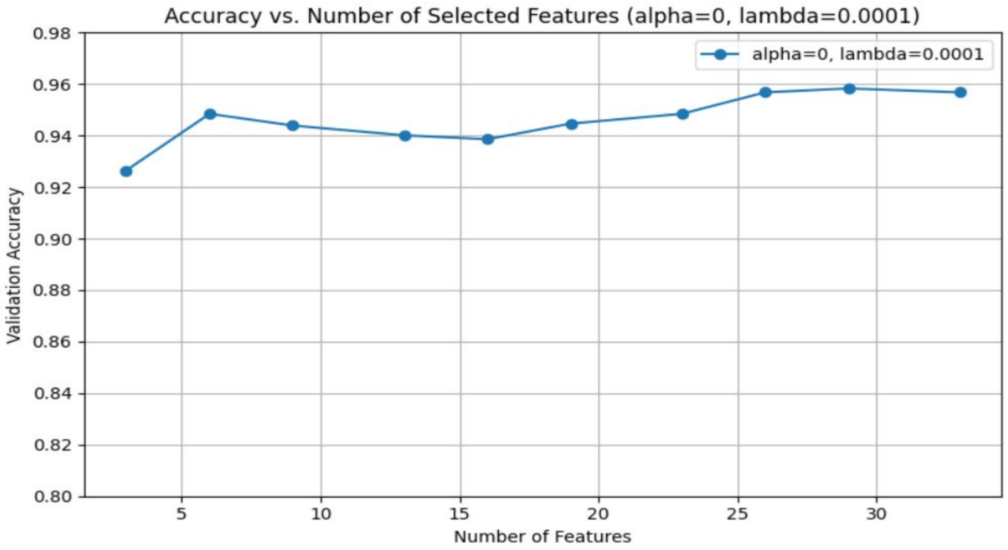
POST STAGE-1 WORK

1. MUSK

Best Hyperparameters: Alpha = 10,
Lambda = 0.001
Best Average Validation Accuracy = 0.9492 ,
Number of Selected Features = 28



Ablation Study (Alpha = 0)
Best Lambda with Alpha equal to 0 = 0.001
Best Average Validation Accuracy = 0.9463 ,

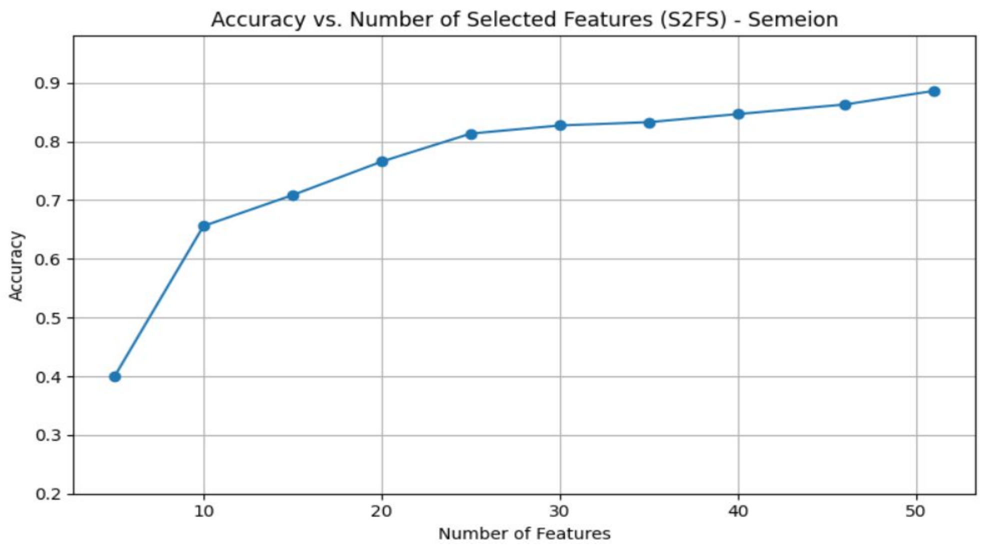


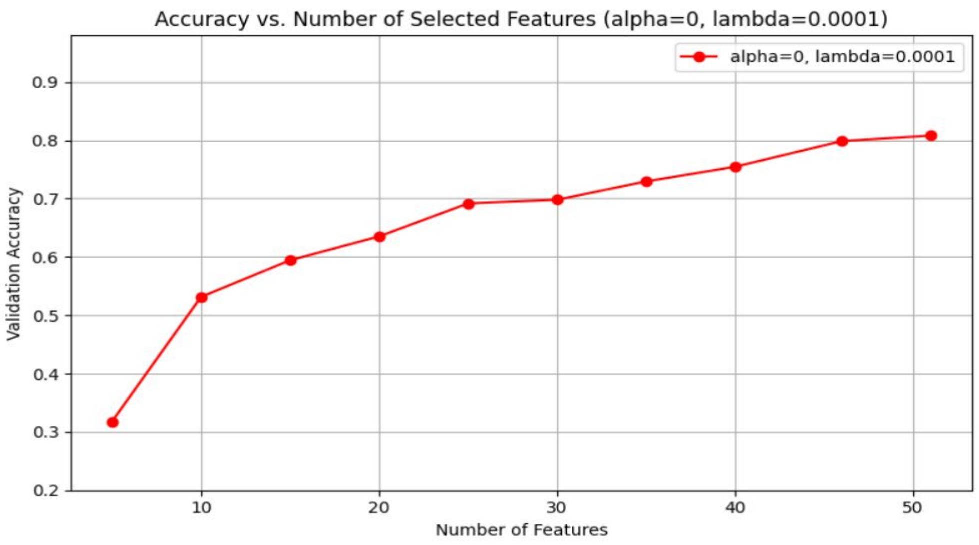
POST STAGE-1 WORK

2. SEMEION

Best Hyperparameters: Alpha = 0.0001, Lambda = 0.0001
Best Average Validation Accuracy = 0.7099 ,
Number of Selected Features = 252

Ablation Study (Alpha = 0)
Best Lambda with Alpha equal to 0 = 0.0001
Best Average Validation Accuracy = 0.6711 ,

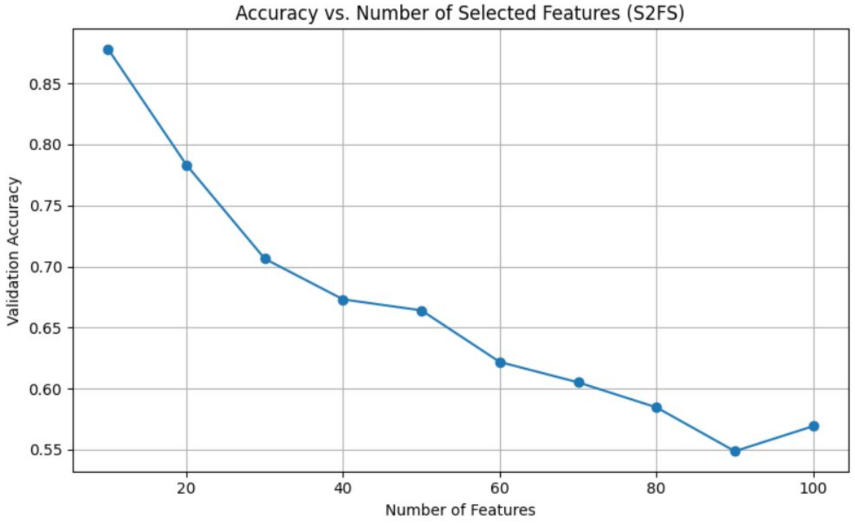




POST STAGE-1 WORK

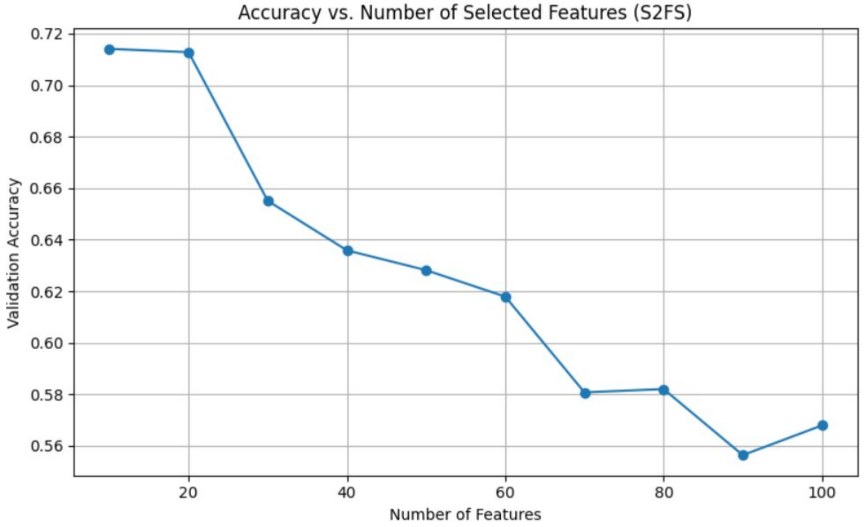
3. MADELON

Best Hyperparameters: Alpha = 0.1,
Lambda = 0.01
Best Average Validation Accuracy = 0.6641 ,
Number of Selected Features = 92



Ablation study (Alpha =0)

Best Hyperparameters: Alpha = 0, Lambda = 0.01
Best Average Validation Accuracy = 0.5718 ,

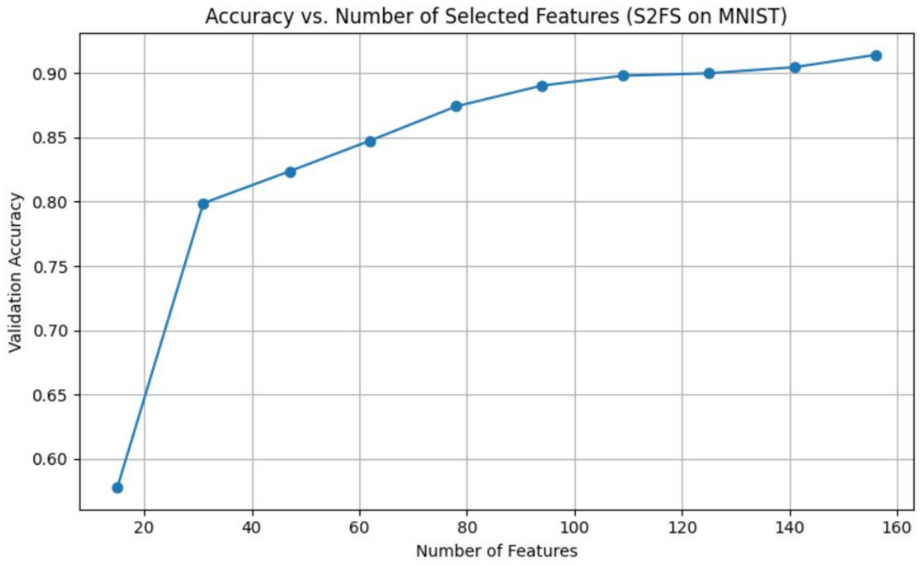


POST STAGE-1 WORK

4. MNIST

Best Hyperparameters: Alpha = 10^-6, Lambda = 10^-5
Number of Selected Features = 605

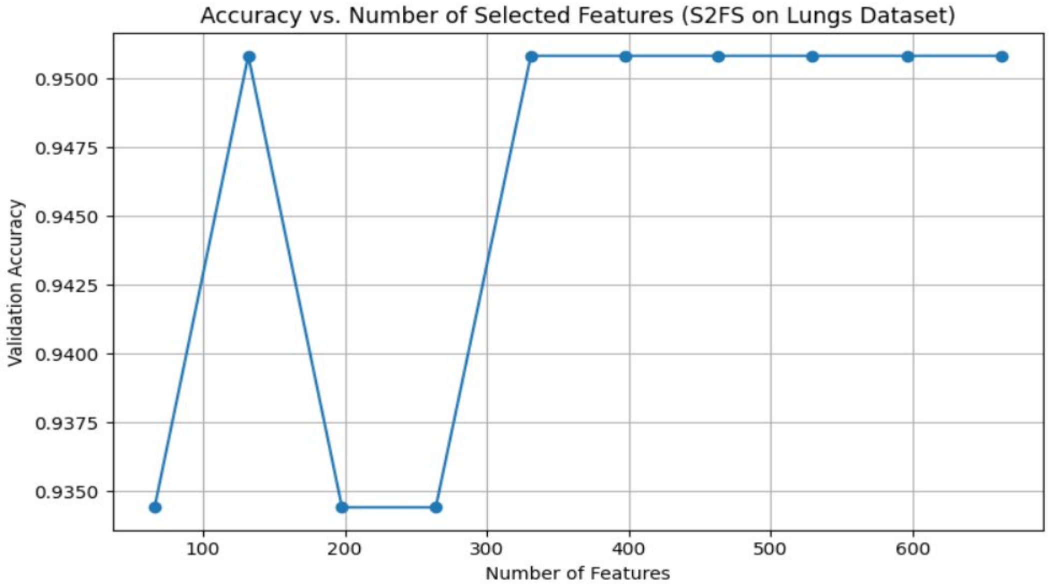
Best Average Validation Accuracy: 0.8742



5. LUNGS

Best Hyperparameters: Alpha = 1, Lambda = 0.0001
Number of Selected Features = 443

Best Average Validation Accuracy: 0.9508



SUMMARY OF EXPERIMENTS REPLICATED

Datasets

- Five benchmark datasets were selected from the original paper for evaluation:
  - Madelon, MNIST, Semeion, Musk, Lung

Experimental Settings

- **Data Splitting**:
  - 70% training, 30% testing (stratified split).

- **Hyperparameter Tuning:**
  - Sparsity parameter (λ): Tested range $[10^{-6}, 10^{-2}]$ (log scale).
  - Discriminative regularization (α): Tested range $[10^{-6}, 10^{2}]$.
  - Optimal values selected via validation set performance (5-fold CV).

- **Evaluation Protocol:**
  - KNN classifier (k=5) on selected features.
  - Metrics: Classification accuracy, sparsity rate.

SUMMARY OF EXPERIMENTS REPLICATED

- **Reproducibility:**
  - Plotted accuracy vs. no. of selected features for all datasets.
  - Ablation study on Madelon (with/without discriminative term).

Implementation Details

- **ADMM Optimization:**
  - Penalty parameter μ tuned for convergence.
  - Early stopping if relative change < 1e-4 for 10 iterations.

- **Codebase:** Python (NumPy, SciPy), shared publicly with documentation.

Ablation Study

- **Objective:** Isolate impact of discriminative regularization (LDA term).
- **Setup:**
  - Trained S^2FS on Madelon with $\alpha=0$ vs. $\alpha>0$.
  - Fixed λ, compared accuracy/sparsity.

SUMMARY OF NOVELTY EXPERIMENT

We incorporated the addition of **L1 regularizer** with weight β into the loss function as a novel aspect in our project. The L1 term enforces element wise sparsity.

Thus, the new loss function becomes:

$$\min_{\boldsymbol{W},\boldsymbol{M},\boldsymbol{N},\boldsymbol{O}} \mathcal{L}(\boldsymbol{O}) + \alpha \operatorname{tr}(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}) + \lambda \|\boldsymbol{M}\|_{2,0} + \beta \|\boldsymbol{N}\|_1$$

$$\text{s.t.} \quad \boldsymbol{W} = \boldsymbol{M}, \quad \boldsymbol{W} = \boldsymbol{N}, \quad \boldsymbol{W} = \boldsymbol{O}, \quad \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W} = \boldsymbol{I}$$

$$\downarrow$$

$$\mathcal{L}_{\text{aug}}(\boldsymbol{W},\boldsymbol{M},\boldsymbol{N},\boldsymbol{O},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2,\boldsymbol{\Delta}_3) = \mathcal{L}(\boldsymbol{O}) + \alpha \operatorname{tr}(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}) + \lambda \|\boldsymbol{M}\|_{2,0} + \beta \|\boldsymbol{N}\|_1$$
$$+ \frac{\mu}{3} \left\| \boldsymbol{W} - \boldsymbol{M} + \frac{\boldsymbol{\Delta}_1}{\mu} \right\|_F^2 + \frac{\mu}{3} \left\| \boldsymbol{W} - \boldsymbol{N} + \frac{\boldsymbol{\Delta}_2}{\mu} \right\|_F^2 + \frac{\mu}{3} \left\| \boldsymbol{W} - \boldsymbol{O} + \frac{\boldsymbol{\Delta}_3}{\mu} \right\|_F^2$$

Update rule for N:
Soft thresholding

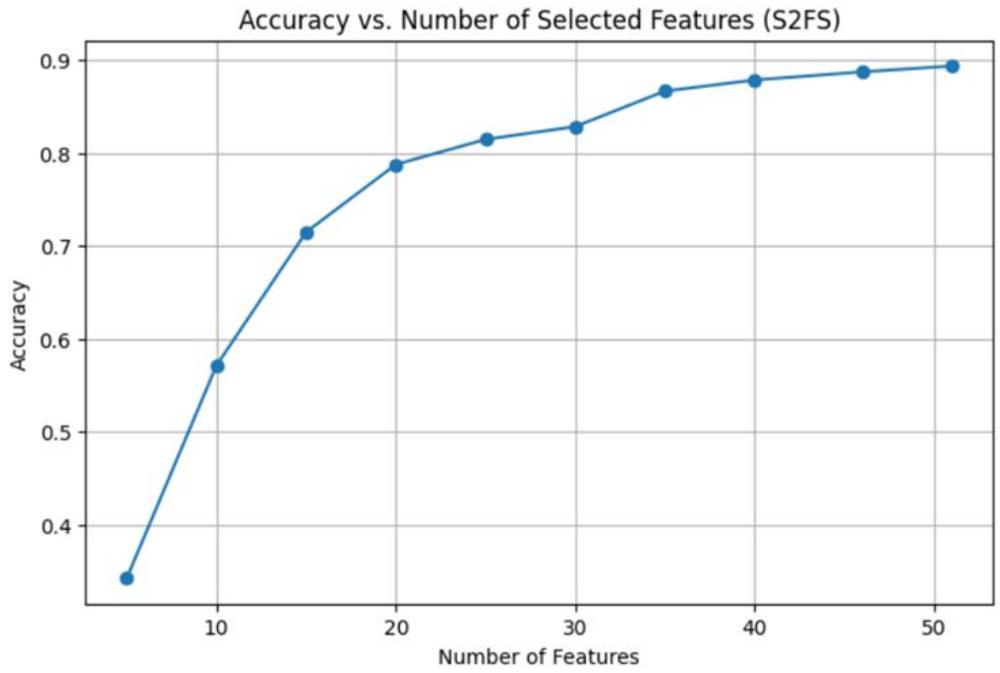
$$N_{ij} = \operatorname{sign}(a_{ij}) \cdot \max(|a_{ij}| - \beta/\mu, 0)$$
$$\text{where} \quad a_{ij} = W_{ij} - (\Lambda_2)_{ij}/\mu$$

SUMMARY OF NOVELTY EXPERIMENT

Results:
We implemented it on the Semeion Dataset. The results obtained were as follows:

Best Hyperparameters: α = 0.01, λ 0.01, β=0.0001
Best Average Validation Accuracy: 0.7389
Number of Selected Features: 255

Interpretation of Results:
•L-2,0 enforced row wise sparsity selecting features that are uniformly relevant across all classes.
•L1 enforces element wise sparsity enabling the algorithm to selectively choose features for specific classes.
•As a result, the **average validation accuracy improves by 2.9%**



CONTRIBUTIONS

Aishwarya Jaiswal (24M1511):

- Rewrote and debugged entire Python implementation post-Stage 1
- Formulated problem statement and solution approach slides
- Validated framework on 3 datasets
- Executed critical ablation studies
- Designed final presentation slides

Sayantan Maiti(24M1503):

- Established foundational understanding of algorithms
- Enhanced AI-generated code for accuracy:
- Fine-tuned hyperparameters
- Tested framework on 2 additional datasets
- Co-authored final project report

Collaborative Outcomes:

- Successful replication of S^2FS with verified results
- Comprehensive ablation analysis validating LDA's impact

CONCLUSION

This project successfully implemented and evaluated the Sparse Softmax Feature Selection (S^2FS) algorithm for high-dimensional multi-class datasets. Through rigorous experimentation, **we demonstrated that S^2FS effectively combines:**

- $\ell_{2,0}$-norm regularization for exact sparsity control
- Softmax classification for probabilistic feature weighting
- LDA-based discriminative regularization for enhanced class separation

Key findings from our implementation include:

- we have implemented the code on the 5 datasets and found out that the 4 dataset have the consistent performance which is almost similar to original paper's result
- Ablation studies confirming the critical role of LDA regularization

FUTURE DIRECTIONS

Extension to Multi-Label Learning

- Adapt the $\ell_{2,0}$-norm regularization and Softmax framework for multi-label classification tasks, where each instance can belong to multiple classes simultaneously.

Integration with Deep Learning

- Incorporate S^2FS into deep neural networks (e.g., as a feature selection layer) to enhance interpretability and reduce redundancy in high-dimensional deep features.

Implementing Elastic Net

- Instead of giving separate weights to L-2,0 and L1 term, we can use Elastic Net and check the performance of our model.

REFERENCES

1. 1. Zhenzhen Sun, Zexiang Chen, Jinghua Liu, and Yuanlong Y. Multi-class feature selection via Sparse Softmax with a discriminative regularization. International Journal of Machine Learning and Cybernetics, 2025. https://link.springer.com/content/pdf/10.1007/s13042-024-02185-5.pdf

1. Tianji Pang, Feiping Nie, Junwei Han, and Xuelong Li. Efficient Feature Selection via L2,0-norm Constrained Sparse Regression. IEEE Transactions on Knowledge and Data Engineering, 2019. https://ieeexplore.ieee.org/abstract/document/8386668

1. Zheng Wang, Feiping Nie, Lai Tian, Rong Wang, and Xuelong Li. Discriminative Feature Selection via A Structured Sparse Subspace Learning Module. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), 2020. https://www.ijcai.org/proceedings/2020/0416.pdf

THANK YOU