Data Mining: LassoNet + Neural Networks

By ***Team 13***:

Alvin Alex

Valerie Le

Annshu Prajapati

Anton Vernikov

Goal of LassoNet: A Neural Network with Feature Sparsity

- Introduction of LassoNet: Presenting a novel neural network framework integrating direct feature selection.
- Extension of Lasso Regression: Expanding feature sparsity from linear to non-linear neural networks.

Problem Addressed:

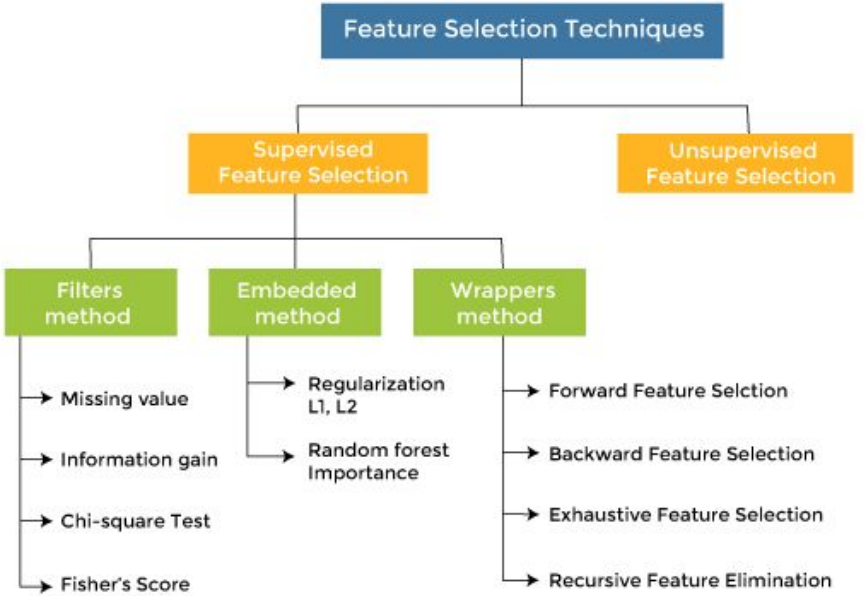
- Limitations of Traditional Lasso Regression: Inapplicable to neural networks due to their non-linear nature.
- Complexity in Neural Network Interpretation: Absence of integrated feature selection mechanisms leads to intricate and opaque models.

Background Overview

- Lasso(l1-regularized) regression assigns zero weights to the most irrelevant features, and is widely used in data science.
- The drawback of Lasso is that it only offers solutions to linear models.
- The authors proposed a new approach that extends lasso regression and its feature sparsity to feed-forward neural networks.
- The solution was LassoNet, a neural network framework with global feature selection that can be easily applied.
- The method allows capturing arbitrary nonlinearity in a nonparametric way while simultaneously performing feature selection.

Benefits of Feature Selection

- Allows shorter training time
- Reduces overfitting
- Helps overcome dimensionality
- Increases interpretability
- Improves accuracy

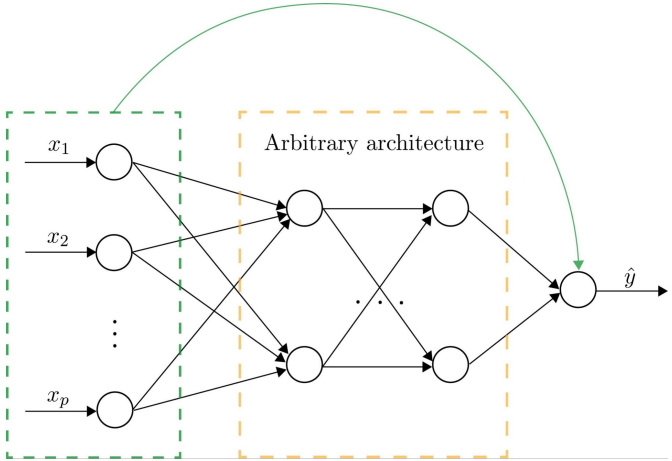


Formulation of the Lassonet

- We consider the class of all fully connected feed-forward residual neural networks, abbreviated as $\mathscr{F} = \left\{ f : f(\mathbf{x}) = \theta^T \mathbf{x} + \mathbf{NN}(\mathbf{x}, W) \right\}$.
- **Notation:** W denotes the network parameters, K denotes the size of the first hidden layer, $W^{(0)} \in \mathbb{R}^{d \times K}$ denotes the first hidden layer, $x \in \mathbb{R}^d$ denotes an input data point, and $L(\theta, W) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x_i}, y_i; \theta, W)$ denotes the loss on the training data set.
- The LassoNet objective function is defined as

$$\begin{aligned} &\underset{\theta, W}{\text{minimize}}\ L(\theta, W) + \lambda \|\theta\|_1 \\ &\text{subject to}\ \left\| W_j^{(0)} \right\|_\infty \leq M|\theta_j|,\ j = 1, \ldots, d. \end{aligned} \tag{1}$$

- The constraint $|W_{jk}^{(0)}| \leq M \cdot |\theta_j|$, $k = 1, \ldots, K$, budgets the total amount of non-linearity involving variable j according to the relative importance of X_j as a linear variable.
- It follows that $W_j = 0$ as soon as $\theta_j = 0$. In other words, variable j is completely inactive from the model, without the need for an explicit penalty on W.
- The only hyper-parameters are the hierarchy coefficient, M, and the regularization strength, λ.

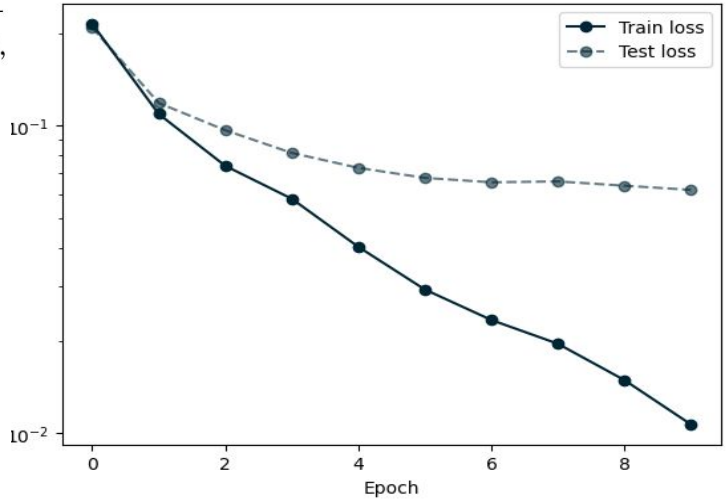


The architecture of LassoNet consists of a single residual connection, shown in green, and an **arbitrary feed-forward neural network**, shown in **black**. The residual layer and the first hidden layer are optimized jointly using a hierarchical operator.

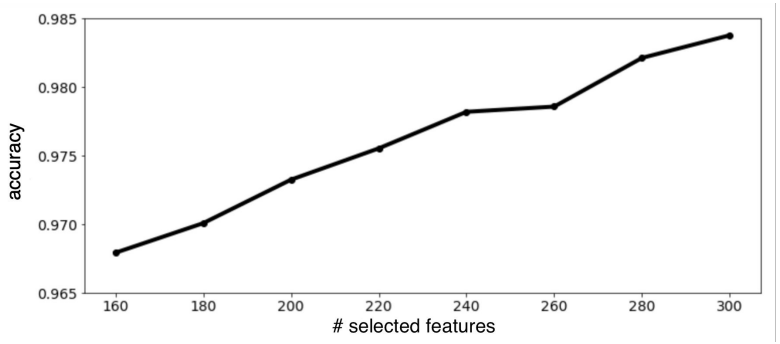
Algorithm

Algorithm 1 Training LassoNet

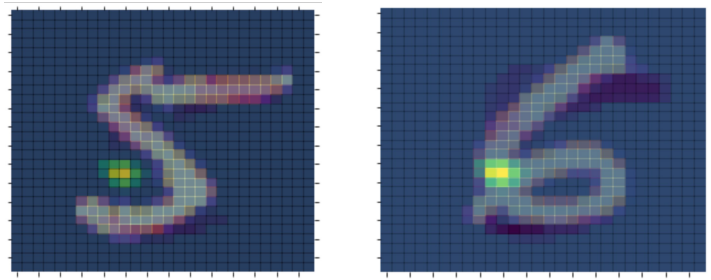
Input: training dataset $X \in \mathbb{R}^{n \times d}$, training labels Y, feed-forward neural network $f_W(\cdot)$, number of epochs B, hierarchy multiplier M, path multiplier ϵ, learning rate α
Initialize and train the feed-forward network on the loss $L(X, Y; \theta, W)$
Initialize the penalty, $\lambda = \epsilon$, and the number of active features, $k = d$
while $k > 0$ **do**
    Update $\lambda \leftarrow (1 + \epsilon)\lambda$
    for $b \in \{1 \dots B\}$ **do**
        Compute gradient of the loss w.r.t to θ and W using backpropagation
        Update $\theta \leftarrow \theta - \alpha \nabla_\theta L$ and $W \leftarrow W - \alpha \nabla_W L$
        Update $(\theta, W^{(0)}) = \text{HIER-PROX}(\theta, W^{(0)}, \lambda, M)$
        Apply early-stopping criterion
    end for
    Update k to be the number of non-zero coordinates of θ
end while



Demonstrating LassoNet on MNIST



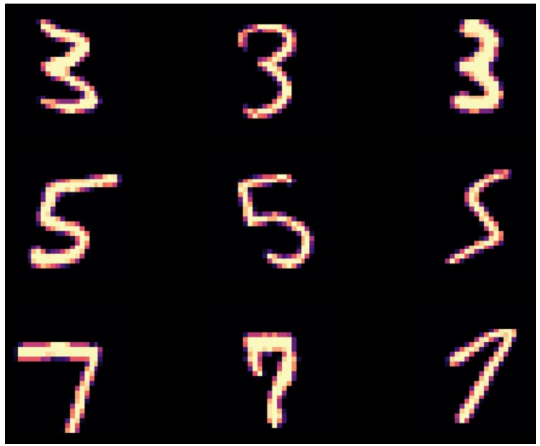
The classification accuracy by number of selected features



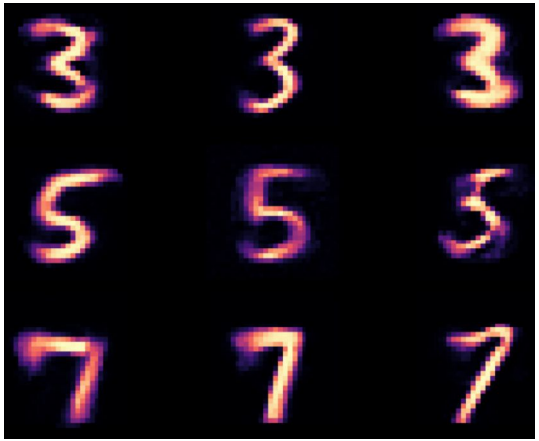
Individual pixel importance for the model with 220 active features

The results of using LassoNet to simultaneously select informative pixels and classify digits 5 and 6 from the MNIST dataset. The classification accuracy by number of selected features is given above

Unsupervised Learning using LassoNet on the MNIST dataset



3 test images from each class of digits are shown, sampled at random.



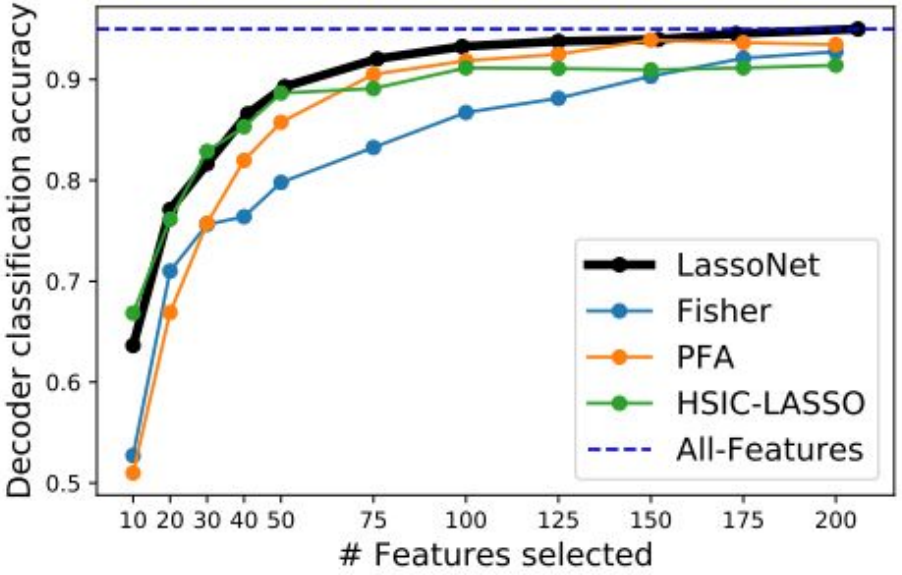
The reconstructed versions of the test images using LassoNet with an intermediate penalty level (corresponding to about 50 active features)

Model Evaluation in Paper

On the ISOLET dataset for each method

1. Select k features using the feature selection method
2. Feed those features to a downstream classifier
3. Measure the accuracy of the classifier
4. Repeat for varying k

Dataset	(n, d)	# Classes
ISOLET	(7797, 617)	26



Recreating results for new datasets

Dataset	(n, d)	# Classes
MNIST	(10000, 784)	10
MNIST-Fashion	(10000, 784)	10