ML01 – Spring 2021
Lab 8: Neural networks

The `Boston` dataset in package `MASS` records median house values (variable `medv`) for 506 neighborhoods around Boston. We will use neural networks to predict `medv` from 13 predictors such as `rm` (average number of rooms per house), `age` (average age of houses), and `lstat` (percent of households with low socioeconomic status), etc., and we will compare the results with those of other machine learning methods. We will use the neural network implementation in package `nnet`.

Data preprocessing

Standardize the 12 quantitative predictors and randomly split the data into a training set of size $n = 300$ and a test set.

Part I

In this first part, we use only variable `lstat` as a predictor.

1. Train networks of different sizes. For each one, compute the test MSE and plot the prediction function $y = \widehat{f}(x)$.
2. Train a network with 10 hidden units and weight decay, with different values of the regularization coefficient λ. For each trained network, compute the test MSE and plot the prediction function $y = \widehat{f}(x)$.
3. Find the optimal λ by cross-validation, and compute the corresponding test MSE.

Part II

We now use all 13 predictors.

1. Train a network with 30 hidden units and weight decay, with different values of the regularization coefficient λ. For each trained network, compute the test MSE and plot the prediction function $y = \widehat{f}(x)$.

2. Find the optimal λ by 5-fold cross-validation, and compute the corresponding test MSE. (Repeat the cross-validation and average the cross-validation estimates to reduce the variance).
3. Compare the results with those of linear regression, Lasso and random forests. Which predictors seem to be the most important to predict the median house value ?