Math6373 Final (take home exam) Due date May 6th 2020 at midnight

1 Prediction Task : Select one major stock on the US stockmarket. On day "t", let S(t) be the price of this stock at closing time. On each day "t" , we want *to predict* the future stock price S(t+1) given the last 20 oberved stock prices S(t) , S(t-1), S(t-2), ..., S(t-19)

2 Data Set : Let t=1, 2, ... N be the days on which the US stock exchange was open during the time period 2014-2015-2016-2017 . Download the time series S(t) for t= 1,2, ..., N

2 PreProcessing: Replace *isolated* missing values S(t) by the mean of two actual values closest to time t. If there are too many missing values, download another stock . For $20 \leq t \leq N-1$, compute the following three moving averages of the time series S :

MA5(t) = [S(t-4) + S(t-3) + S(t-2) +S(t-1) + S(t)] /5

MA10(t) = [S(t-9) + S(t-8) + ... + S(t)] /10

MA20(t) = [S(t-19) + S(t-18) + ... + S(t)] /20

Plot the 4 curves S(t) , MA5(t), MA10(t), MA20(t), on the same graph

*3 Training and Test sets for an MLP predictor :*On each day $t \geq 20$, the *recent past* of the series S will be defined as the 1x18 line vector

Vt = [MA5(t), MA10(t), MA20(t), S(t), S(t-1), S(t-2)..., S(t-13), S(t-14)]

For $20 \leq t \leq N-1$ the input vector Vt will be the input of our MLP predictor , which will have a *single* output neuron with state Zt. This output Zt will be the MLP prediction computed on day t for the *target* TARGt = S(t+1), which is not known at time t.

For this prediction task , we have a data set of (N-20) "cases" $Case_{20}$ $Case_{21}$ $Case_{22}$ ... $Case_{N-1}$, indexed by t= 20, 21, ... N-1. Each $Case_t$ is described by 18 features = 18 coordinates of vector Vt. The TRUE output to be predicted at time t is the yet unknown TARGt = S(t+1).

The data set of (N-20) cases for MLP prediction learning is denoted

PredCases = { all pairs (Vt, TARGt) with t= 20, 21, ... N-1 }

Randomly Split the set *PredCases,* with 90% cases in the training set *PredTRAIN*, and 10 % cases in the test set *PredTEST*

4 MLP predictor : Our MLP predictor *(MLPpred)* will have the simple 3 layers architecture

INPUT ==> HiddenLayer K ==> OUTPUT with dim(INPUT) = 18 , dim(OUTPUT) = 1

dim(K) = k *to be selected below*. For each training input Vt we want the MLP output Zt to be close to TARGt= S(t+1).

Implement PCA on the set of all input vectors Vt , with t= 20 ,21, ..., N . Determine the number **k** of principal components which preserves *95%* of the variance (see HW3) and fix *dim(K)= k*.

Compute the number w of weights and thresholds in this MLP, and compare w to the number of informations provided by the training set.

5 Training of the MLP predictor: Implement an automatic training on the training set PredTRAIN, with the options :

RELU response, Loss = "MSE", Stochastic Gradient Descent *or* ADAM, Batch Learning, Early Stopping

Let RMSE be the root mean squared error $\sqrt{MSE}$. Plot the evolution of RMSE versus the number of batches (one curve for the training set and one for the test set) . Compare these two curves.

Plot on the same graph the true values TARGt = S(t+1) and the predicted values Zt . Comments.

Compute the Mean Relative Errors of Prediction MREP on the training set :

MREP= average (| Zt - TARG t | / TARGt) over all cases in the Training set

Compute similarly MREP on the test set. Comments .

6. Denote NOD1 NOD2 ... NODk the hidden neurons . For j= 1...k, compute and display the mean activity Yj of NODj over all cases in the Training set. Display all the weights W1 ... Wk linking the neurons NOD1 ... NODk to the output node.

For each hidden NODj compute IMPj = Wj Yj = *average impact* of NODj on the prediction Zt. Display these k impacts and comment. Identify the hidden neuron NOD* with maximal impact on Zt

7. Denote INP1 INP2 ... INP18 the 18 input neurons. Compute and display the mean activities X1 ... X18 of the 18 input neurons. Display all the weights U1 ... U18 linking the input nodes INP1 ... INP18 to the neuron NOD*. For each input neuron INPs compute Fs= Us Xs which is the *average impact* of input feature "s" on the key hidden neuron NOD*. Identify the 5 input features with the largest impact on NOD*. Comments.