Using ann_tab_cv() and transfer learning

Walter K. Kremers, Mayo Clinic, Rochester MN

28 July 2023

The functions

The ann_cv_lin() function fits an artificial neural network model, or neural network (NN) for short, to tabular data. The network has a simple "sequential" structure with linear components and by default ReLU activations. It has two hidden layers where the number of terms in each of these two layers can be specified by the user. Data for input are in a format as with the nested.glmnetr() function. Models can be fit as generalizations to Cox, logistic or linear regressions.

We wrote the ann_cv_lin() function in part to better understand how a NN model which begins its "numerical optimization" near a model informed by a linear model might perform compared to a standardly fit NN model. The program is not optimized to find the "best algorithm" for model fitting so other NN programs may perform better. Still, it does show how this "transfer learning" from a linear model to the NN can dramatically improve model fit. Here we use the relaxed lasso regression model (tuned on lambda and gamma) as the linear model for the transfer learning. In that both the relaxed lasso and NN models are fit in nested.glmneter() we use this function to fit these transfer learning NN models. Because nested cross validations can have long run times, as an option, one may fit these models without actually performing the outer nested loop.

The nested.glmnetr() function also fits lasso, XGBoost and p-value tuned stepwise regression models while using one level of cross validation to inform hyper parameters and another level of cross validation to evaluate model performance, that is it does nested cross validation. Measures of model performance include deviance and agreement (concordance or R-square). We also calculate linear calibration coefficients obtained by regressing the original outcome variable on the predicteds in the hold out data of the nested cross validation. In this formulation the predicteds are taken before any "final" activation so they are analogous to the X * Beta term from a Cox, logistic or linear model. A linear calibration regression coefficient greater than 1 suggests a bias in the model in under estimating the relative differences in risk between sample units and a coefficient less than 1 suggests a bias in over estimating these relative differences.

Data requirements

The data elements for input to the ann_tab_cv() function are basically the same as for the other *glmnetr* functions like nested.glmenter() and cv.glmneter(). Input data should comprise of 1) a (numerical) matrix of predictors and 2) an outcome variable or variables in (numerical) vector form. NULL and Na's are not allowed in input matrices or vectors. For the estimation of the "fully" relaxed parts (where gamma=0) of the relaxed lasso models the package is set up to fit the "gaussian" and "binomial" models using the *stats* glm() function and Cox survival models using the the coxph() function of the *survival* pacakge. When fitting the Cox model or an extension like with NNs the "outcome" variable is interpreted as the "time" variable in the Cox model, and one must also specify 3) an indicator variable for event, again in vector form. Start times are not at this time accounted for in ann_tab_cv(). Row i of the predictor matrix and element i of the outcome vector(s) are to include the data for the same sampling unit.

R torch

For fitting the NNs we use the R torch package. We refer to the package reference manual and the book https://skeydan.github.io/Deep-Learning-and-Scientific-Computing-with-R-torch/ for general information on this package. To run the NN models one needs to install the R torch package. When first loading torch from the R library, the package may prompt the user to allow torch to download some tensor libraries which is necessary to run torch. The NN torch library generally requires input data to be in torch tensor format, and provides output in torch tensor format as well. Here we convert data from the usual R format to the torch format so the user does not have to. In some functions, too, we convert outputs from torch format back to standard R format for the user.

An example dataset

To demonstrate usage of ann_tab_cv() we first generate a data set for analysis. The code

```
# Simulate data for use in an example Neural Network fit of survival data
# first, optionally, assign seeds for random number generation to get replicable results
set.seed(116291949)
torch_manual_seed(77421177)
simdata=glmnetr.simdata(nrows=1000, ncols=100, beta=NULL, intr=c(1,0,1,0))
```

generates simulated data with interactions (specified by the intr=c(1,0,1,0) term) for analysis. We extract data in the format required for input to the ann_tab_cv() program.

```
# Extract simulated survival data
xs = simdata$xs         # matrix of predictors
yt = simdata$yt         # vector of survival times
event = simdata$event   # indicator of event vs. censoring from survival data
y_ = simdata$y_         # vector of quantitative values as an outcome variable
yb = simdata$yb         # vector of 0 and 1 values for an outcome variable
```

Inspecting the predictor matrix we see

```
# Check the sample size and number of predictors
print(dim(xs))
```

```
## [1] 1000  100
```

```
# Check the rank of the design matrix, i.e. the degrees of freedom in the predictors
rankMatrix(xs)[[1]]
```

```
## [1] 94
```

```
# Inspect the first few rows and some select columns
print(xs[1:10,c(1:12,18:20)])
```

```
##       X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12        X18        X19         X20
##  [1,]  1  1  0  0  0  0  0  0  0   1   0   1  0.1513225 -0.4034383  0.35250844
##  [2,]  1  0  0  0  1  0  0  1  0   0   0   0 -1.1610480  0.5533030  0.14578868
```

```
##  [3,]  1  0  0  1  0  0  1  0  0   0   0   0 -0.3292269  0.3086399 -0.48443836
##  [4,]  1  1  0  0  0  0  0  0  0   1   0   0  2.0635214 -0.5500741 -0.02173104
##  [5,]  1  0  0  0  1  0  0  1  0   0   0   0  0.3905722 -0.6836452 -0.37643201
##  [6,]  1  0  1  0  0  0  0  0  1   0   0   0 -0.2397597  1.6909447  0.49599945
##  [7,]  1  0  1  0  0  0  0  1  0   0   0   0 -0.5592424  0.2314638 -0.53198341
##  [8,]  1  0  0  1  0  0  0  0  0   0   1   0 -1.0050514  0.5319574  0.54287646
##  [9,]  1  0  0  1  0  0  0  0  0   0   1   0  1.2548034  0.8213164  0.17067691
## [10,]  1  0  0  0  1  0  0  0  1   0   0   0 -0.3079151 -0.6105910 -0.88711869
```

Fitting a basic neural network model to "tabular" data using ann_tab_cv()

To fit a NN model we can use a simple function call as in

```
## fit a model with some monitoring to the R console
set.seed( 67213041 )
torch_manual_seed( 7221250 )
ann_tab_cox_ex1 = ann_tab_cv(myxs=xs, myy=yt, myevent=event, family="cox", fold_n=10,
                             eppr=-2)
```

```
##  Epoch:   0 Full data Loss:  6.205775 Train Concordance: 0.5323026
##   imax= 42   minloss= 1   which_loss= 42   which_agree= 47   gotoend= 0   l1= 0
##  Epoch: 42 Train Loss:  5.198872 Train Concordance: 0.8798803
```

We see there is little the user needs to specify, which is basically the data one includes when fitting a regular Cox model. The one piece of input which is unusual is the eppr=-2. The eppr term has no influence on the model but instructs the program to send some fit information to the R console for monitoring the model fitting process. Here we see the model loss and concordance calculated using the training data. From the output generated during model fit we see that for the "starting point" of the model fit the loss function is about 6.2 and concordance 0.5. The concordance of about 0.5 is what we expect by chance alone, consistent with the fact that when we begin the model fit we are taking a model chosen at random. After the gradient descent goes through 42 iterations, as suggested by cross validation, the loss based upon the training data goes down to about 5.2 and the concordance increases to about 0.88, a marked improvement. This minimal amount of information sent to the R console is achieved by setting eppr=-2. To avoid any information being sent to the console one may set eppr = -3, or any number less than -2. Numbers larger than -2 will provide more updates during the model fitting process.

To see more information on the model fit we submit

```
## simple model summary
ann_tab_cox_ex1$modelsum
```

```
##        n folds         epochs      length Z1      length Z2           actv
##     10.0000000    200.0000000     16.0000000      8.0000000      1.0000000
##          drpot           mylr             wd             l1          lasso
##      0.0000000      0.0050000      0.0000000      0.0000000      0.0000000
##         lscale          scale     which loss    which agree        CV loss
##      5.0000000      1.0000000     42.0000000     47.0000000      3.4905488
##       CV Agree    CV accuracy     naive loss    naive agree naive accuracy
##      0.7965589      0.0000000      5.1988721      0.8798803      0.0000000
##     agree i_=0
##      0.5323026
```

Here we see that the cross validation is based upon a 10 fold split of the data, and the gradient descent algorithm goes through 200 iterations or epochs for each fold of the data. For this simple model fit we are tuning on number of iterations in the gradient descent fitting. This is not necessarily the most robust way to fit a model but this simple method is common suggesting its usefulness and it will serve to evaluate how NN model fits might be impacted by using a transfer learning form a linear model to the NN (and discuss below). Next we see the first hidden layer is a vector of length 16, and the second hidden layer a vector of length 8. The actv of 1 specifies a ReLU activation function was used, drpot of 0 that no drop out was used, mylr of 0.005 indicates the learning rate used in the optimization. wd and l1 indicate the L2 and L1 penalties used when model fitting, corresponding to ridge and lasso regression. Both are 0 here indicating these penalties were not employed. The "which loss" and "which agree" values indicate the number of epochs when loss is minimized and agreement (concordance or R-square) is maximized in the cross validation and inform the number of epochs to be used in the final model fit using the whole data set. The CV loss and CV concordance based upon these numbers of fits were about 3.49 and 0.8, while the "naive" loss and concordances calculated using training data were much greater at about 5.2 and 0.88. This is as we expect for concordance as the values based upon the training data are associated with a lot of over fitting due to the number of free parameters in the NN model. The naive deviance of 5.2 being larger than the cv deviance of 3.49 may derive from the larger sample size when using the whole data set for calculations. CV accuracy, the fraction of "correctly classified" is not calculated here and 0 is displayed. The concordance for the random starting model was about 0.5, as sent to the R console in the example above.

We can get more information on the NN model, for example with the command

```
## This shows the tensor (matrix) structure used in the model
str(ann_tab_cox_ex1$model$parameters)
```

```
## List of 6
##  $ 0.weight:Float [1:16, 1:100]
##  $ 0.bias  :Float [1:16]
##  $ 3.weight:Float [1:8, 1:16]
##  $ 3.bias  :Float [1:8]
##  $ 6.weight:Float [1:1, 1:8]
##  $ 6.bias  :Float [1:1]
```

The first item here, 0.wieght, describes the dimensions of the tensor (matrix) used in transforming the original data to the first hidden layer, here 100 and 16. While we would usually expect this tensor to be 100 rows tall and 16 columns wide, tensors are often transposed to take greater advantage of machine architecture to speed calculations. Next we see 0.bias, a tensor with 1 dimension and thus essentially a vector, which contains the intercept terms when transforming from the input data to the first hidden layer. The 3.weight and 3.bias terms determine the transformation from the first hidden to the second hidden layer. Here they transform a vector with 16 terms to a vector or 8 terms. Finally the 6.weight and 6.bias terms determine the transformation from the second hidden layer to the model output of 1 dimension. To get more model information we can use the command

```
## This shows the general structure of the neural network model
ann_tab_cox_ex1$model$modules[[1]]
```

```
## An `nn_module` containing 1,761 parameters.
##
## -- Modules ---------------------------------------------------------------
## * 0: <nn_module> #1,616 parameters
## * 1: <nn_relu> #0 parameters
## * 2: <nn_dropout> #0 parameters
## * 3: <nn_module> #136 parameters
```

```
## * 4: <nn_relu> #0 parameters
## * 5: <nn_dropout> #0 parameters
## * 6: <nn_module> #9 parameters
## * 7: <nn_identity> #0 parameters
```

This shows the model structure. The first term nn_module, involves a linear transformation from the inputs to a hidden layer. This module is based upon the "nn_linear" tensor module which we modified to allow us to more easily set or update model weights and biases. This is followed by a ReLU activation and then a "dropout" with probability 0. The activation function is a nonlinear transformation that allows the NN to fit more general response surfaces than linear models. Without this the NN would simply involve matrix multiplication which would result in another matrix, and reduce to a linear model. The dropout can be used to randomly set terms to 0, i.e. to drop them out, with a specified probability. This can sometimes improve model fit. Here though we set this probability to 0 so this is not impacting our model fit. The final element "nn_identity" unsurprisingly means to apply the identity function. This could be set to "sigmoid", i.e. (1/(1+exp(-x)), to transform the input values to the interval (0,1), as can be done when fitting generalizations of the logistic regression model.

See here how we have been inspecting elements of lists of lists. The basic flow of model fitting using R torch is to first define a model object and then use numerical routines to update the contents of this model object using numerical algorithms. Here we combine this model object with other information in a list to organize in one place the model, information about the model fit and how the model was derived.

Though we can, we do not typically inspect individual parameters (weights and biases) in a NN. The value of NN models lies instead in their ability to predict. In general predicted values can be gotten in R torch by a command like model(newdata). We can get predicteds form an ann_tab_cv() output object, for example, with a command like

```
# ann_tab_cv() predicteds in torch tensor format
preds = ann_tab_cox_ex1$model(xs)
preds[1:8]
```

```
## torch_tensor
## -1.2694
## -2.9719
##  2.3675
## -0.3093
## -0.3961
##  0.6866
## -2.3836
## -1.3124
## [ CPUFloatType{8,1} ][ grad_fn = <SliceBackward0> ]
```

Since R torch models work with tensors the model predicteds, too, are put in a torch tensor format. We can easily change the output to a usual R numerical format by wrapping the output in the as.numeric() function, for example, as in

```
# ann_tab_cv() predicteds in standard R numerical format
preds = as.numeric( ann_tab_cox_ex1$model(xs) )
preds[1:8]
```

```
## [1] -1.2693577 -2.9718771  2.3675289 -0.3093177 -0.3960863  0.6865525 -2.3835573
## [8] -1.3124293
```

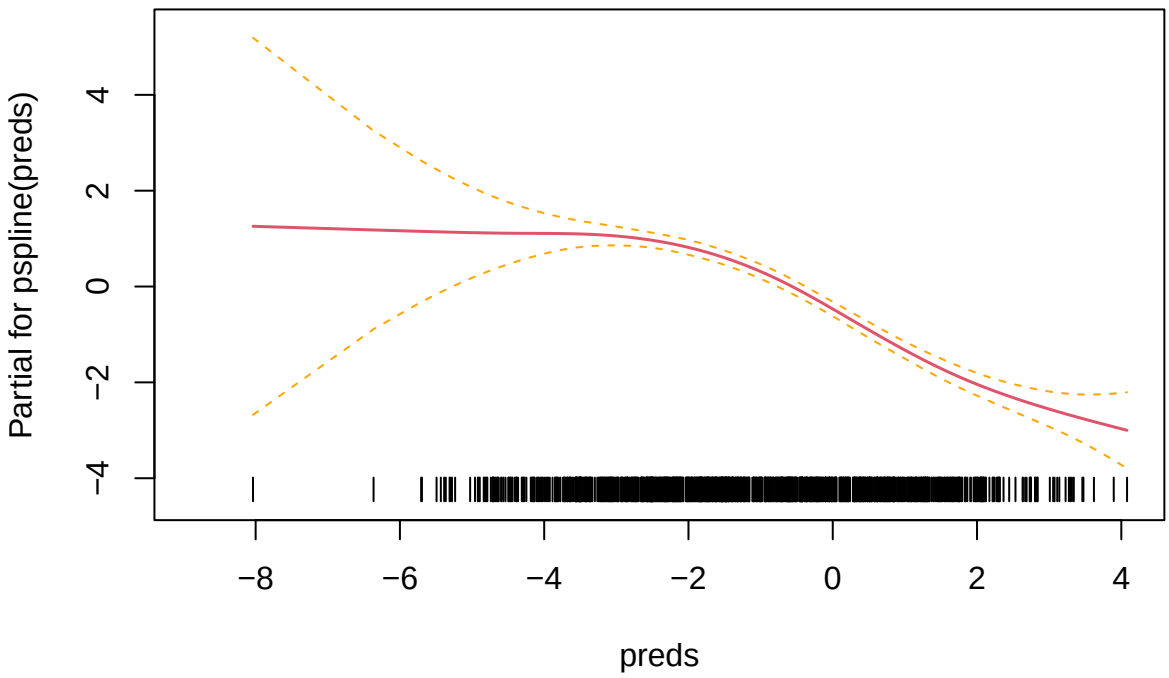
We can then use the predicteds in this R numerical format for evaluation using other R functions and packages. For example for calibration of the NN model one could begin with the code

```
# ann_tab_cv() lienar calibration
cox_fit1 = coxph(Surv(yt, event) ~ preds)
summary(cox_fit1)
```

```
## Call:
## coxph(formula = Surv(yt, event) ~ preds)
##
##   n= 1000, number of events= 699
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## preds 1.2725    3.5696   0.0383 33.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## preds      3.57     0.2801     3.311     3.848
##
## Concordance= 0.883  (se = 0.005 )
## Likelihood ratio test= 1512  on 1 df,   p=<2e-16
## Wald test            = 1104  on 1 df,   p=<2e-16
## Score (logrank) test = 1252  on 1 df,   p=<2e-16
```

Here the coefficient of about 1.27 being greater than 1 suggests the NN model may be underestimating the hazard ratios between sample elements. To further evaluate the need for calibration we can fit a spline on the NN predicteds as in

```
# Fit a spline to preds using coxph, and plot
cox_fit2 = coxph(Surv(y_, event) ~ pspline(preds))
termplot(cox_fit2,term=1,se=T,rug=T)
```



The spline fit depicted in this graph does not appear to be consistent with a straight line suggesting the NN is not, in its current form, well calibrated.

So, why the ann in ann_tab_cv? Many of the torch functions begin with nn_. To not confuse our function with a native torch function we begin the name with a different, yet we hope recognizable, letter sequence. The ann can be thought of as denoting "artificial NN".

Transfer learning from linear models to neural network models

It is very common to start model fits informed by previously fit NNs. Recognizing that NNs are a generalization of linear models one can start an NN model fit informed by a linear model. Basically one can chose some of the weights (like betas in a linear regression) and biases (like intercepts) to replicate the informing linear model, and let the other weights and biases be chosen at random as is common, in many contexts usual, when fitting a NN.

Fitting a neural network model to "tabular" data informed by a linear model using nested.glmnetr()

To demonstrate such a NN model fit informed by a linear model we use the nested.glmnetr() function. This is out of convenience in that the nested.glmenter() function already fits (linear) lasso and NN models and so contains the pieces to combine the two for transfer learning.

An informed NN based upon the Cox regression partial likelihood

A NN for "tabular" data informed by a linear model can be fit using nested.glmnetr()

```
## Fit a NN informed with starting point for iterative fit by a lasso fit
time_start = diff_time()
```

```
## Start at Sys.time = 2023-07-28 19:33:18.654072
```

```
nested_cox_fit_ex2 = nested.glmnetr(xs=xs, y_=yt, event=event, family="cox",
                    dolasso=1, doann=1, ensemble=c(0,0,0,1), folds_n=c(5,0),
                    seed=c(101844880,882560297),track=0)
time_last = diff_time(time_start)
```

```
## Sys.time = 2023-07-28 19:33:59.064862,  elapsed time = 0:0:40 h:m:s
```

```
nested_cox_fit_ex2$tuning
```

```
##  steps_n  folds_n   method  dolasso   doann_   doxgb_  dorpart   dostep
##    "100"      "5" "loglik"      "1"      "1"      "0"      "0"      "0"
##    doaic     ties    limit
##      "0"  "efron"      "1"
```

Here we see that to fit a NN informed by a lasso model the user need do very little beyond specifying the data for the fit. In addition to providing the usual data for a Cox model, one specifies doann=1 to "do an ann" model, and ensemble=c(0,0,0,1) where the ensemble[4] = 1 indicates the NN model is to be fit informed by the lasso model. The folds_n=c(10,0) specifies there are to be 10 folds for CV, and the 0 indicates to not do a nested CV, that is to just do the one layer of CV. If unspecified, e.g. as in folds_n=10, a nested CV will be performed allowing one to assess and compare model performances. How to specify these terms is described in the reference manual. Here we used diff_time() to show how long the program took (Apple M1) to fit the NN model.

One can get model information about the model fit similar to when using the ann_tab_cv() function, for example

```
summary(nested_cox_fit_ex2)
```

```
##
##  Sample information including number of records, events, number of columns in
##  design (predictor, X) matrix, and df (rank) of design matrix:
##          family               n         nevents      xs.columns           xs.df
##           "cox"          "1000"           "699"           "100"            "94"
## null.dev/events
##        "12.418"
##
##  Tuning parameters for models :
## folds_n
##     "5"
##
##  Tuning parameters for  lasso update  ANN model :
##   n folds    epochs length Z1 length Z2      actv     drpot      mylr        wd
```

```
##     5.000   200.000    18.000    10.000     1.000     0.000     0.001     0.000
##        l1    lscale     scale
##     0.000     5.000     1.000
##
##  Naive agreement for cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   0.833   0.845   0.832   0.838   0.840   0.849   0.850
##
##  Number of non-zero terms in cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##      10      39       8       9       6       9      99
##
##  Cross validation informed neural network :
##
##       naive agreement :
## lasso update
##        0.838
```

```
nested_cox_fit_ex2
```

```
##
##  Sample information including number of records, events, number of columns in
##  design (predictor, X) matrix, and df (rank) of design matrix:
##          family               n         nevents      xs.columns           xs.df
##           "cox"          "1000"           "699"           "100"            "94"
## null.dev/events
##        "12.418"
##
##  Tuning parameters for models :
## folds_n
##     "5"
##
##  Tuning parameters for  lasso update  ANN model :
##   n folds    epochs length Z1 length Z2      actv     drpot      mylr        wd
##     5.000   200.000    18.000    10.000     1.000     0.000     0.001     0.000
##        l1    lscale     scale
##     0.000     5.000     1.000
##
##  Naive agreement for cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   0.833   0.845   0.832   0.838   0.840   0.849   0.850
##
##  Number of non-zero terms in cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##      10      39       8       9       6       9      99
##
##  Cross validation informed neural network :
##
##       naive agreement :
## lasso update
##        0.838
```

```
## The lasso informed model fit is saved to "object"$ann_fit_4
## The object that provides the lasso information to ann_fit_4 is "object"$cv_glment_fit
nested_cox_fit_ex2$ann_fit_4$modelsum
```

```
##        n folds         epochs      length Z1      length Z2           actv
##      5.0000000    200.0000000     18.0000000     10.0000000      1.0000000
##          drpot           mylr             wd             l1          lasso
##      0.0000000      0.0010000      0.0000000      0.0000000      1.0000000
##         lscale          scale     which loss    which agree        CV loss
##      5.0000000      1.0000000      1.0000000      1.0000000      3.9233511
##       CV Agree    CV accuracy     naive loss    naive agree naive accuracy
##      0.8366672      0.0000000      5.5189328      0.8377749      0.0000000
##     agree i_=0         bestof
##      0.8377749      1.0000000
```

```
## This shows the number of terms, weights and biases, for the the two hidden layers
str(nested_cox_fit_ex2$ann_fit_4$model$parameters)
```

```
## List of 6
##  $ 0.weight:Float [1:18, 1:100]
##  $ 0.bias  :Float [1:18]
##  $ 3.weight:Float [1:10, 1:18]
##  $ 3.bias  :Float [1:10]
##  $ 6.weight:Float [1:1, 1:10]
##  $ 6.bias  :Float [1:1]
```

From this we also see the number of elements in the two hidden layers are 18 and 10. We can inspect the model structure as in

```
# view more information on the NN model structure
nested_cox_fit_ex2$ann_fit_4$model$modules[[1]]
```

```
## An ‘nn_module‘ containing 2,019 parameters.
##
## -- Modules ---------------------------------------------------------------------
## * 0: <nn_module> #1,818 parameters
## * 1: <nn_relu> #0 parameters
## * 2: <nn_dropout> #0 parameters
## * 3: <nn_module> #190 parameters
## * 4: <nn_relu> #0 parameters
## * 5: <nn_dropout> #0 parameters
## * 6: <nn_module> #11 parameters
## * 7: <nn_identity> #0 parameters
```

We can obtain predicteds from the lasso informed NN models. Since this involves combining information from the lasso and NN fits we wrap this in the function predict_ann_tab() with an example usage of

```
#print(ann_cv_lin_fit1$model)
preds = predict_ann_tab(nested_cox_fit_ex2, xs, modl=4)
preds[1:10]
```

```
##  [1] -0.24385998 -2.63770723  2.40171099  0.71735048 -0.50466347  0.59652889
##  [7] -3.04103565  0.08762515  3.68572569  0.64531612
```

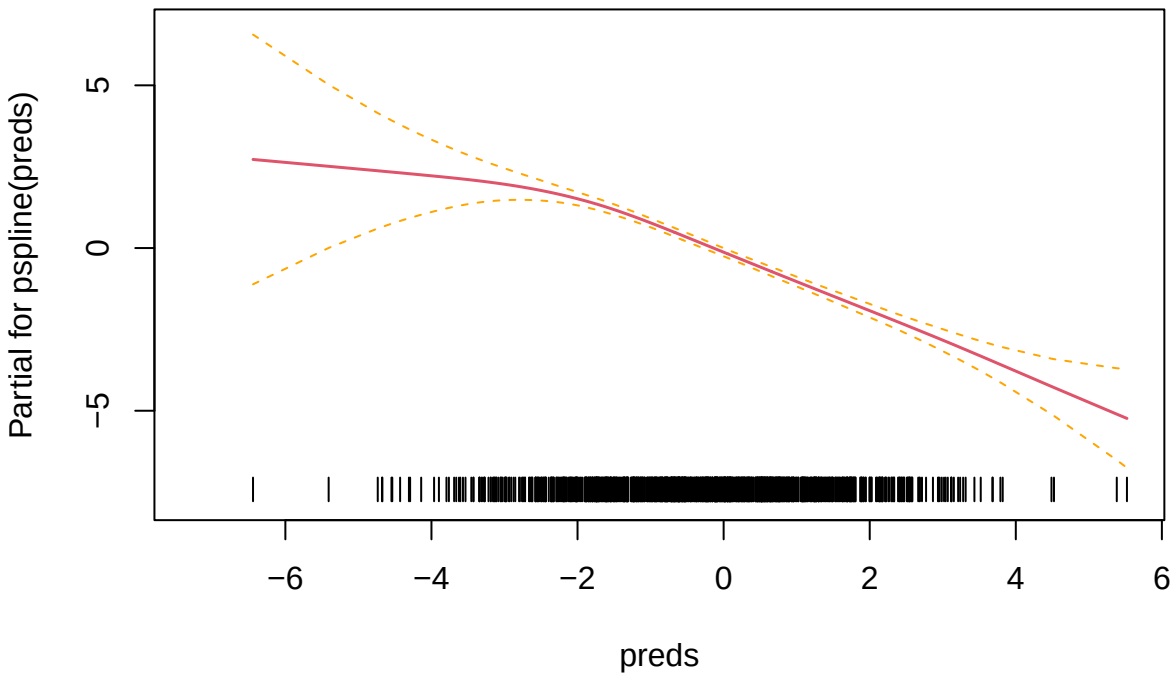
Here the first input is the nested.glmnetr() output object, the second the input the data for which we want the predicteds and finally modl=4 means we want the predicted from the model indicated by ensemble[4] = 1 in the call to nested.glmentr().

```
# ann_tab_cv() lienar calibration
cox_fit3 = coxph(Surv(yt, event) ~ preds)
summary(cox_fit3)
```

```
## Call:
## coxph(formula = Surv(yt, event) ~ preds)
##
##   n= 1000, number of events= 699
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## preds 1.0052    2.7324   0.0337 29.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## preds     2.732      0.366     2.558     2.919
##
## Concordance= 0.838  (se = 0.007 )
## Likelihood ratio test= 967.6  on 1 df,   p=<2e-16
## Wald test            = 889.9  on 1 df,   p=<2e-16
## Score (logrank) test = 866.8  on 1 df,   p=<2e-16
```

Here the liner calibration coefficient is very near to 1, much nearer than for the model not informed by the lasso fit.

```
# Fit a spline to preds using coxph, and plot
cox_fit4 = coxph(Surv(y_, event) ~ pspline(preds))
termplot(cox_fit4,term=1,se=T,rug=T)
```



The spline fit is nearer a straight line than was the case for the uninformed NN fit, but still not as straight as we might want.

An informed NN based upon least squares

This model is essentialy a generalization of linear regression and can be fit, for example, by

```
nested_nrm_fit_ex2 = nested.glmnetr(xs=xs, y_=y_, family="gaussian", dolasso=1,
        seed=c(17820414,95337508), doann=1, ensemble=c(1,0,0,1), folds_n=c(10,0), track=-1)
```

We can inspect numerically the fit by

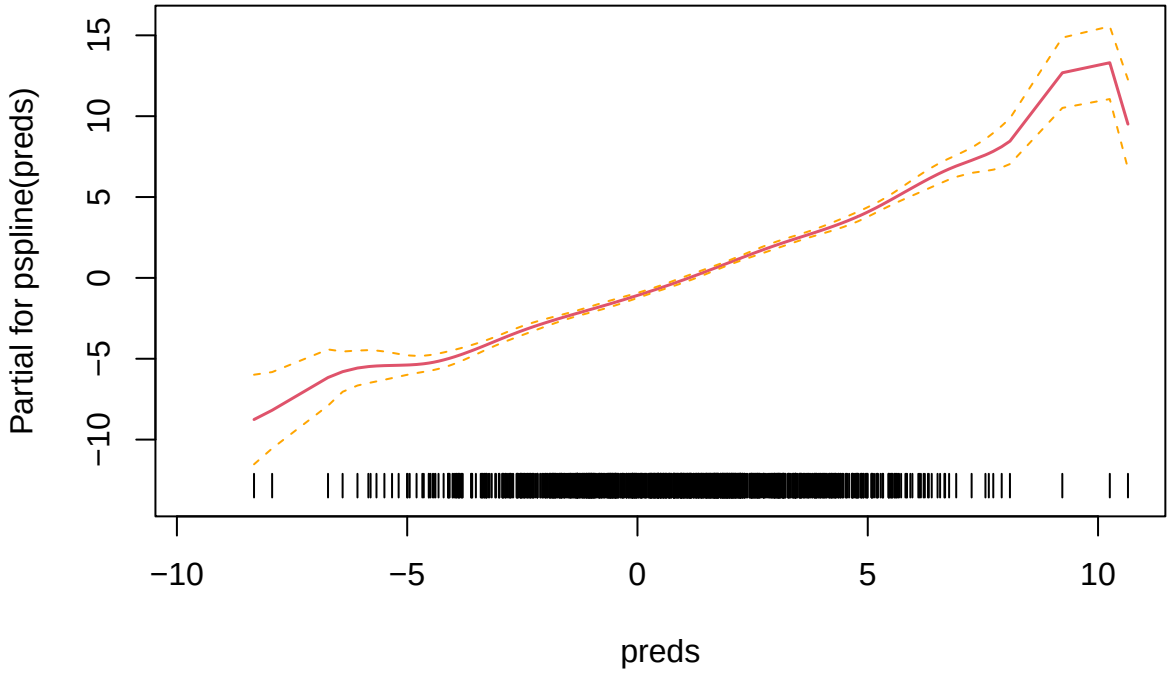
```
preds = predict_ann_tab(nested_nrm_fit_ex2, xs, 4)
glm_fit1 <- glm(y_ ~ preds, family="gaussian")
summary(glm_fit1)
```

```
##
## Call:
## glm(formula = y_ ~ preds, family = "gaussian")
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08032    0.04914  -1.634    0.102
## preds        0.99952    0.01731  57.732   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.148111)
##
##     Null deviance: 9303.4  on 999  degrees of freedom
## Residual deviance: 2143.8  on 998  degrees of freedom
## AIC: 3606.5
##
## Number of Fisher Scoring iterations: 2
```

and graphically by

```
glm_fit2 <- glm(y_ ~ pspline(preds), family="gaussian")
termplot(glm_fit2, rug=T, se=T)
```



Here the linear calibration coefficient is about 1 and the spline fit on the predicteds about linear with wiggle in the extremes where there are few data.

An informed NN based upon logistic regression

An example NN fit based upon the logistic model frame work is

```
## fit a neural network based upon logistic regression framework
set.seed(  4695289  )
torch_manual_seed(  24260321  )
```

```
nested_bin_fit_ex4 = suppressWarnings( nested.glmnetr(xs=xs, y_=yb, family="binomial",
                     dolasso=1, doann=1, ensemble=c(1,0,0,1), folds_n=c(5,0),
                     track=0) )
## print a short summary
nested_bin_fit_ex4
```

```
##
##  Sample information including number of records, events, number of columns in
##  design (predictor, X) matrix, and df (rank) of design matrix:
##       family            n      nevents   xs.columns        xs.df null.dev/obs
##   "binomial"       "1000"        "609"        "100"         "94"      "1.338"
##
##  Tuning parameters for models :
## folds_n
##     "5"
##
##  Tuning parameters for  lasso update  ANN model :
##    n folds    epochs length Z1 length Z2      actv     drpot      mylr        wd
##    5.000   200.000    18.000    10.000     1.000     0.000     0.001     0.000
##       l1    lscale     scale
##    0.000     5.000     1.000
##
##  Naive agreement for cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   0.896   0.915   0.894   0.894   0.905   0.924   0.926
##
##  Number of non-zero terms in cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##      14      45       8       8       8       8      99
##
##  Cross validation informed neural network :
##
##       naive agreement :
##   Uninformed lasso update
##         0.991        0.894
```

Comparison of lasso and neural network models

As the nested.glmnetr() function performs nested cross validation of the lasso and NN models we can use it to compare performances between these models. An example is

```
# Fit a neural network model informed by cross validation
set.seed( 17820414 )
torch_manual_seed( 95337508 )
simdata=glmnetr.simdata(nrows=1011, ncols=100, beta=NULL, intr=c(1,0,0,1))
ensemble=c(1,0,0,1) ;
doann=list(epochs=200, epochs2=400, mylr=0.002, mylr2=0.001, eppr=-3, eppr2=-3,
          lenz1=16, lenz2=8, actv=1, drpot=0, wd=0, wd2=0, L1=0, L12=0,
          fold_n=5, minloss=0, gotoend=0)
nested_nrm_fit_ex5 = nested.glmnetr(xs=simdata$xs,y_=simdata$y_,
                     family="gaussian",dolasso=1,doann=doann,
```

```
                    ensemble=ensemble,folds_n=c(5,1),steps_n=40,track=0)
## print a short summary
nested_nrm_fit_ex5
```

```
##
##  Sample information including number of records, number of columns in
##  design (predictor, X) matrix, and df (rank) of design matrix:
##       family            n   xs.columns        xs.df null.dev/obs
##   "gaussian"       "1011"        "100"         "94"     "14.936"
##
##  Tuning parameters for models :
## folds_n
##     "5"
##
##  Tuning parameters for  lasso update  ANN model :
##   n folds    epochs length Z1 length Z2      actv     drpot      mylr        wd
##     5.000   400.000    18.000    10.000     1.000     0.000     0.001     0.000
##        l1    lscale     scale
##     0.000     5.000     1.000
##
##  Nested cross validation agreement (r-square) for cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   0.838   0.841   0.840   0.840   0.678   0.840   0.831
##
##  Nested cross validation agreement (r-square) for cross validation informed Neural Network :
##   Uninformed lasso update
##        0.721        0.840
```

where the NN model with an R-square of 0.84 seems to perform similar to the relaxed lasso models with an R-squares of about 0.84. Further, the NN fit without this transfer of information from the lasso model with it's R-square of 0.72, did not perform nearly as well as the lasso informed NN or the lasso model itself. This shows the direct benefit of this transfer learning of the information from the linear model when fitting NN models. It is not the NN itself that it is performing competitivley with the lasso model but the NN model fit in conjunction with the lasso model information. We can further inspect performances with the summary function as in

```
# print nested CV  concordances
summary( nested_nrm_fit_ex5 )
```

```
##
##  Sample information including number of records, number of columns in
##  design (predictor, X) matrix, and df (rank) of design matrix:
##       family            n   xs.columns        xs.df null.dev/obs
##   "gaussian"       "1011"        "100"         "94"     "14.936"
##
##  Tuning parameters for models :
## folds_n
##     "5"
##
##  Tuning parameters for  lasso update  ANN model :
##   n folds    epochs length Z1 length Z2      actv     drpot      mylr        wd
##     5.000   400.000    18.000    10.000     1.000     0.000     0.001     0.000
```

```
##        l1    lscale     scale
##     0.000     5.000     1.000
##
##
##  Nested Cross Validation averages for LASSO (1se and min), Relaxed LASSO, and gamma=0 LASSO :
##
##       deviance per record :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   2.517   2.389   2.398   2.375   4.676   2.375   2.615
##
##       deviance per record (linerly calibrated) :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   2.383   2.342   2.357   2.354   4.619   2.354   2.498
##
##       number of nonzero model terms :
##     1se     min    1seR    minR 1seR.G0 minR.G0
##    20.6    38.8    15.4    17.2    12.4    17.2
##
##       linear calibration coefficient :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   1.106   1.053   1.041   0.995      NA   0.995   1.099
##
##       agreement (r-square) :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   0.838   0.841   0.840   0.840   0.678   0.840   0.831
##
##  Naive agreement for cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##   0.849   0.855   0.849   0.851   0.855   0.860   0.863
##
##  Number of non-zero terms in cross validation informed LASSO :
##     1se     min    1seR    minR 1seR.G0 minR.G0   ridge
##      24      38      17      18      16      18      99
##
##
##  Nested Cross Validation averages for neural network :
##
##       deviance per record :
##    Uninformed lasso update
##         4.149        2.405
##
##       linear calibration coefficient :
##    Uninformed lasso update
##         0.988        0.994
##
##       agreement (r-square) :
##    Uninformed lasso update
##         0.721        0.840
##
##  Cross validation informed neural network :
##
##       naive agreement :
##    Uninformed lasso update
##         0.948        0.852
```

Here the cross validated linear calibration coefficients are about 1 for the lasso informed NN and the tuned relaxed lasso model suggesting these models may be reasonably well calibrated. The fully penalized lasso and the ridge regression model, with cross validated linear calibration coefficients deviating more form 1, are less well calibrated.

For this example we deliberately simulated data where there are interaction or product terms in the analysis data set. This we did by setting intr=c(1,0,1,0) in the call to glmnetr.simdata(). NNs can, if there is sufficient information in the data, pick up non-linear and interaction (or product) terms. In the absence of non-linearity or interactions a strict linear model should out perform a NN model because it more parsimonious uses model parameters.

We do not show more simulation results recognizing that one can typically show one model to be better than the others by simulating the right data set. Instead others can run the nested.glmnetr() function on their own data sets, potentially historical data sets if at hand, and see which models tend to perform better in their setting.

Internal implementaiton of the "transfer learning"

The NN model informed by the relaxed lasso model when ensemble[4]=1 adds a column of predicteds to the input matrix in the nested.glmnetr() call, and then extends the hidden layers to carry the positive and negative part of the lasso predicteds through to the final model output. It also treats the weights and biases for this new column so that they remain the same while the ohter weights and biases are updated according to the gradient descent algorithm. A second option set by ensemble[2] = 1 appends the lasso predicteds to the xs predictor matrix and treats this similar to the other features or input variables. The NN model informed by the relaxed lasso model when ensemble[3]=1 fits like with ensemble[4] = 1 but alows the weights and biases to update during the iterative fit. The models fit for ensemble[i+4]=1 are like those fit for ensemble[i]=1 except only those variables with non-zero coefficients in the lasso model are included in the input data set.

Before fitting the NN models the predictor variables are standardized to have mean 0 and standard deviation of 1. This is accounted for when deriving predicteds. This is important when using L1 (lasso) or L2 (ridge or weight decay) penalties to assure the models are not scale dependent. This is also done in the *glmnet* package functions.

Further extensions of "transfer learning"

Just as one can assign a subset of the bias and weight (initial) values for the NN to replicate a linear model, and thereby improve model fit, one can also define another subset of the bias and weight values to replicate linear splines. One attractive feature of the NN numerical optimization is that the bias terms dictating where the spline "knots" are updated during model fit to improve fit, that is they need not be fixed in advance of the fitting.

Transfer learning and the gradient boosting machine

Just as we have informed the NN model with the results from a relaxed lasso model fit we an do the same with gradient boosting machines (GBM) by either adding the lasso predicted as an additional feature (predictor) or including it as an offset. A GBM is fit by setting doxgb=1 in the nested.glmnetr() call, and the tranfer lerning model governed by using the ensemble parameter like wiht the NN. Results are similar but different from those for the NN models. For our data sets the GBMs took much longer to run because of our search algorithm to find a better hyperparameter set.