

Module 11: Solutions to Recommended Exercises

TMA4268 Statistical Learning V2020

Martina Hall, Michail Spitieris, Stefanie Muff, Department of Mathematical Sciences, NTNU

April 2, 2020

Problem 1

a)

It is a 4-4-4-3 feedforard neural network with an extra bias node in both the input and the two hidden layers. It can be written in the following form

$$y_c(\mathbf{x}) = \phi_o(\beta_{0c} + \sum_{m=1}^4 \beta_{mc} z_m) = \phi_o(\beta_{0c} + \sum_{m=1}^4 \beta_{mc} \phi_{h*}(\gamma_{0m} + \sum_{l=1}^4 \gamma_{lm} \phi_h(\alpha_{0l} + \sum_{j=1}^4 \alpha_{jl} x_j))).$$

b)

It is not clear wheter the network has 3 input nodes, or 2 input nodes plus one bias node (both would lead to the same representation). The hidden layer has 4 nodes, but no bias node, and the output layer consists of two nodes. This can be used for regression with two responses. If we have a classifiation problem with two classes then we usually use only one output node, but is is possible to use softmax activation for two classes, but that is very uncommon. Remember that for a binary outcome, we would usually only use one output node that encodes for the probability to be in one of the two classes.

c)

When the hidden layer has a linear activation the model is only linear in the original covariates, so adding the extra hidden layer will not add non-linearity to the model. The feedforward model may find latent structure in the data in the hidden layer. In general, however, we would then recommend to directly use logistic regression, because you then end up with a model that is easier to interpret.

d)

This is possible because the neural network is fitted using iterative methods. But, there is not one unique solutions here, and the network will benefit greatly by adding some sort of regulariztion, like weight decay and early stopping.

Problem 2

a)

This is a feedforward network with 10 input nodes plus a bias node, a hidden layer with 5 nodes plus a bias node, and a single node in the output layer. The hidden layer has a ReLU activation function, whereas the output layer has a linear activation function.

The number of the estimated parameters are $(10 + 1) * 5 + (5 + 1) = 61$.

b)

Feedforward network with two hidden layers. Input layer has 4 nodes and no bias term, the first hidden layer has 10 nodes and ReLU activation and a bias node, the second hidden layer has 5 nodes plus a bias node and ReLU activation. One node in output layer with sigmoid activation.

The number of estimated parameters are $4 * 10 + (10 + 1) * 5 + (5 + 1) = 101$.

c)

In module 7 we had an additive model of non-linear function, and interactions would be added manually (i.e., explicitly). Each coefficient estimated would be rather easy to interpret. For neural nets we know that with one hidden layer and squashing type activation we can fit any function (regression), but may need many nodes - and then the interpretation might not be so easy. Interactions are automatically handled with the non-linear function of sums.

Problem 3

a)

```
library(ElemStatLearn)
train_data = zip.train[, -1]
train_labels = factor(zip.train[, 1])
test_data = zip.test[, -1]
test_labels = factor(zip.test[, 1])
mean <- apply(train_data, 2, mean)
std <- apply(train_data, 2, sd)
train_data <- scale(train_data, center = mean, scale = std)
test_data <- scale(test_data, center = mean, scale = std)
```

b)

5 hidden nodes: $257 * 5 + 6 * 10 = 257 * 5 + 6 * 10$ parameters

```
library(nnet)
zipnnnet5 <- nnet(train_labels ~ ., data = train_data, size = 5, MaxNWts = 3000,
  maxit = 5000)
summary(zipnnnet5)
pred = predict(zipnnnet5, newdata = test_data, type = "class")
library(caret)
confusionMatrix(factor(pred), test_labels)
```

The above took some time to run, the results were:

```
> zipnnnet5<- nnet(train_labels~., data=train_data,size=5,MaxNWts=3000,maxit=5000)
iter2960 value 864.566658
final value 864.561810
converged
> summary(zipnnnet5)
a 256-5-10 network with 1345 weights
options were - softmax modelling
  b->h1   i1->h1   i2->h1   i3->h1   i4->h1   i5->h1   i6->h1   i7->h1   i8->h1   i9->h1   i10->h1  i11->h1
-49.27    9.15    1.24   21.03   -2.82   17.97    4.63   11.60   -4.31    2.28   -4.57   -0.00
 i19->h1 i20->h1 i21->h1 i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1 i30->h1
> confusionMatrix(factor(pred),test_labels)
Confusion Matrix and Statistics
```

	Reference									
Prediction	0	1	2	3	4	5	6	7	8	9
0	324	0	6	5	3	9	7	0	3	0
1	1	245	7	0	1	0	0	0	7	5
2	4	7	148	8	12	5	11	0	3	0
3	2	0	6	128	4	10	0	4	5	3
4	4	1	4	0	152	1	1	9	2	4
5	1	0	2	18	1	117	5	0	7	1
6	21	3	11	0	6	1	146	0	3	0
7	0	1	6	4	5	1	0	122	9	5
8	2	2	7	3	6	15	0	1	113	3
9	0	5	1	0	10	1	0	11	14	156

```
Overall Statistics

          Accuracy : 0.8226
          95% CI   : (0.8052, 0.8391)
```

Slightly better and faster with keras, but we see that we probably need more hidden layers since this is a difficult problem! An very short introduction to keras is given later in the second part of module 11.

```
library(keras)
network <- keras_model_sequential() %>% layer_dense(units = 5, activation = "sigmoid",
  input_shape = c(16 * 16)) %>% layer_dense(units = 10, activation = "softmax")
network %>% compile(optimizer = "rmsprop", loss = "categorical_crossentropy",
  metrics = c("accuracy"))
train_data <- array_reshape(train_data, c(7291, 16 * 16))
train_data <- train_data/255
train_labels <- to_categorical(train_labels)

test_data <- array_reshape(test_data, c(2007, 16 * 16))
test_data <- test_data/255
org_test_labels <- test_labels
test_labels <- to_categorical(test_labels)

fitted <- network %>% fit(train_data, train_labels, epochs = 500, batch_size = 128,
  validation_split = 0.2)
library(ggplot2)
plot(fitted) + ggtitle("Fitted model") #same as above
```

```

network %>% evaluate(test_data, test_labels)
# acc : num 0.873
library(caret)
res <- network %>% predict_classes(test_data)
confusionMatrix(factor(as.character(res)), org_test_labels)

```

10 hidden nodes

10 hidden nodes: $257 * 10 + 11 * 10 = 257 * 10 + 11 * 10$ parameters. This is too complex for `nnet` to converge in reasonable time.

```

library(keras)
network <- keras_model_sequential() %>% layer_dense(units = 10, activation = "sigmoid",
  input_shape = c(16 * 16)) %>% layer_dense(units = 10, activation = "softmax")
network %>% compile(optimizer = "rmsprop", loss = "categorical_crossentropy",
  metrics = c("accuracy"))
train_data <- array_reshape(train_data, c(7291, 16 * 16))
train_data <- train_data/255
train_labels <- to_categorical(train_labels)

test_data <- array_reshape(test_data, c(2007, 16 * 16))
test_data <- test_data/255
org_test_labels <- test_labels
test_labels <- to_categorical(test_labels)

fitted <- network %>% fit(train_data, train_labels, epochs = 500, batch_size = 128,
  validation_split = 0.2)
allfitted <- network %>% fit(train_data, train_labels, epochs = 500,
  batch_size = 128)
network %>% evaluate(test_data, test_labels)
# acc : num 0.9133034

```