Dale Luginbuhl
M06202420
CS6073

Homework 1: Linear Regression and Neural Network Regression

Step 1: Data

1) How many data samples are included in the dataset?

```
Number of Samples: 3047
```

2) Which problem will this dataset try to address?

```
Predicting the cancer mortality rate of some population with a given
set of characteristics. The population is defined by the features
available in the data set.
```

3) What is the minimum value and the maximum value in the dataset?

```
avgAnnCount
Max: 38150.0
Min: 6.0

avgDeathsPerYear
Max: 14010
Min: 3

TARGET_deathRate
Max: 362.8
Min: 59.7

incidenceRate
Max: 1206.9
Min: 201.3

medIncome
Max: 125635
Min: 22640

popEst2015
Max: 10170292
Min: 827

povertyPercent
```

```
Max: 47.4
Min: 3.2

studyPerCap
Max: 9762.308998
Min: 0.0

MedianAge
Max: 624.0
Min: 22.3

MedianAgeMale
Max: 64.7
Min: 22.4

MedianAgeFemale
Max: 65.7
Min: 22.3

AvgHouseholdSize
Max: 3.97
Min: 0.0221

PercentMarried
Max: 72.5
Min: 23.1

PctNoHS18_24
Max: 64.1
Min: 0.0

PctHS18_24
Max: 72.5
Min: 0.0

PctSomeCol18_24
Max: 79.0
Min: 7.1

PctBachDeg18_24
Max: 51.8
Min: 0.0

PctHS25_Over
Max: 54.8
```

```
Min: 7.5

PctBachDeg25_Over
Max: 42.2
Min: 2.5

PctEmployed16_Over
Max: 80.1
Min: 17.6

PctUnemployed16_Over
Max: 29.4
Min: 0.4

PctPrivateCoverage
Max: 92.3
Min: 22.3

PctPrivateCoverageAlone
Max: 78.9
Min: 15.7

PctEmpPrivCoverage
Max: 70.7
Min: 13.5

PctPublicCoverage
Max: 65.1
Min: 11.2

PctPublicCoverageAlone
Max: 46.6
Min: 2.6

PctWhite
Max: 100.0
Min: 10.1991551

PctBlack
Max: 85.94779858
Min: 0.0

PctAsian
Max: 42.61942454
Min: 0.0
```

```
PctOtherRace
Max: 41.93025142
Min: 0.0

PctMarriedHouseholds
Max: 78.07539683
Min: 22.99248989

BirthRate
Max: 21.32616487
Min: 0.0
```

4) How many features are in each data sample?

```
Number of Features per Sample: 34
```

5) Does the dataset have any missing information? E.g., missing features.

```
PctSomeCol18_24
Samples w/ data: 762
Samples w/o data: 2285

PctEmployed16_Over
Samples w/ data: 2895
Samples w/o data: 152

PctPrivateCoverageAlone
Samples w/ data: 2438
Samples w/o data: 609
```

6) What is the label of this dataset?

```
TARGET_deathRate
```

7) How many percent of data will you use for training, validation and testing?

90/10/10 for training, validation, and testing respectively.

8) What kind of data pre-processing will you use for your training dataset?

1. Remove the ‘Geography’ column
2. Split the ‘binnedInc’ column into ‘binnedIncLow’ and ‘binnedIncHigh’
3. Deduplicate the training samples

4. Remove features that have missing data
5. min-max normalize each feature

Step 2: Model

Model	MSE
Linear Regression	0.00608
ANN-oneL-16	0.00547
ANN-twoL-32-8	0.00587
ANN-threeL-32-16-8	0.00686
ANN-fourL-32-16-8-4	0.00614

1. Analyze the hypothesis you learn in terms of bias and variance. Which model underfitted? Which model overfitted?

The more complex models, like ANN-twoL-32-8, ANN-threeL-32-16-8, and ANN-fourL-32-16-8-4, began to overfit, especially at higher learning rates. These models achieved their best performance at lower learning rates which prevented overfitting (and reduced variance) at the expense of a higher loss (and increased bias).

The simpler models, like Linear Regression and ANN-oneL-16, seemed to better fit the data

No models seemed to be underfitting.

Step 3: Objective

Step 4: Optimization

Step 5: Model Selection

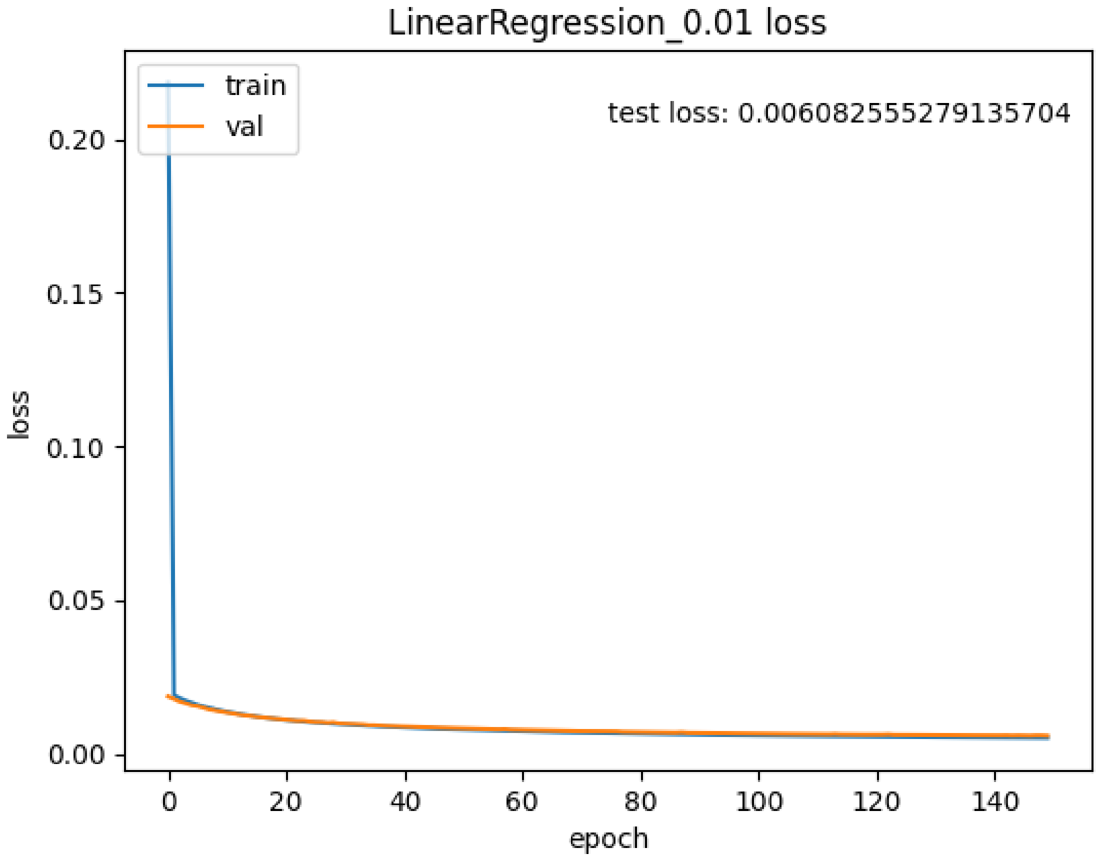
Model	LR: 0.1	LR: 0.01	LR: 0.001	LR: 0.0001
ANN-oneL-16	0.00504	0.00547	0.00736	0.02400
ANN-twoL-32-8	0.02103	0.00587	0.00737	0.01045
ANN-threeL-32-16-8	0.00490	0.00581	0.00686	0.01381
ANN-fourL-32-16-8-4	0.00459	0.00614	0.00649	0.01064

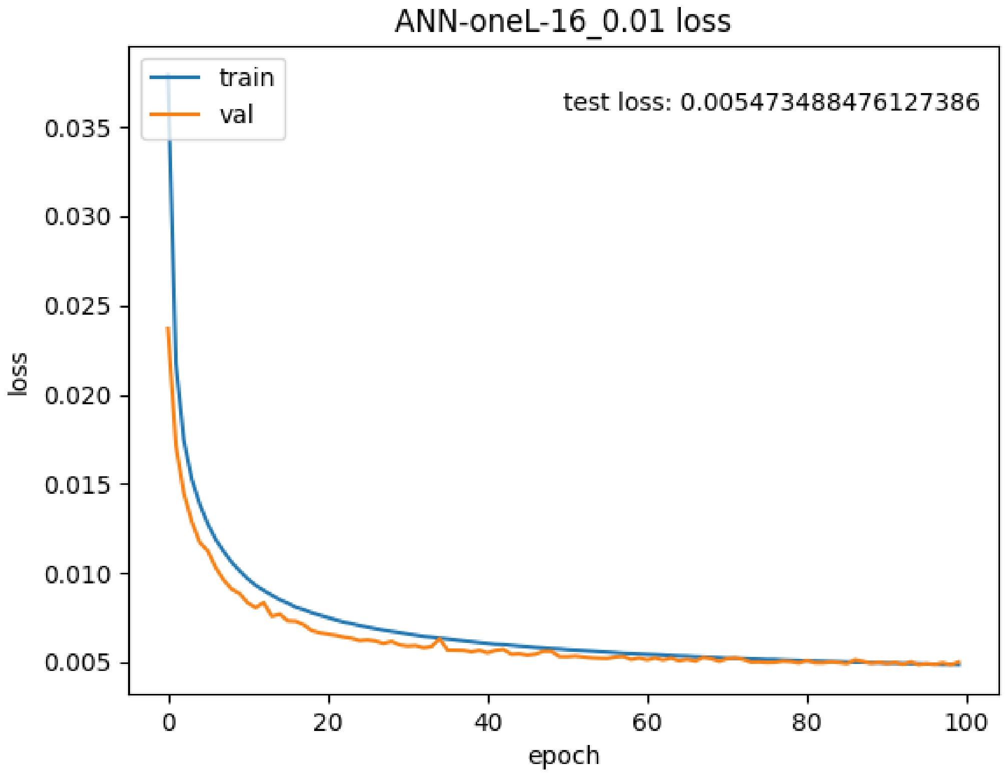
1. Why is the learning rate impacting the model performance? Can you find the best learning rate?

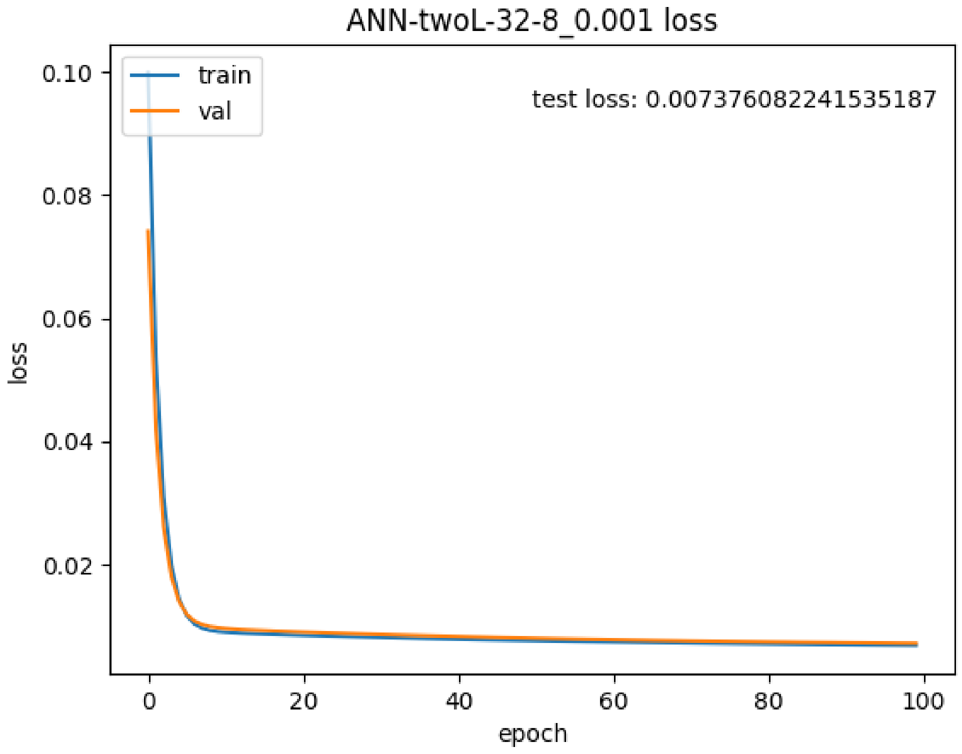
For the smaller learning rates, 0.0001 and even 0.001, the model performance isn't as good because we likely still haven't converged to a minimal loss value. This is because the learning rate is too low and the steps by which we update our weights are too small. Presumably we would converge to a minimal value with more training iterations.
The largest learning rate of 0.1 tended to result in the most minimal loss on the test data (by a small amount). Even though the loss was the lowest, the validation loss tended to be higher and more erratic than the training loss. This leads me to believe that the larger learning rate was causing the model to overfit. If we had done fewer training iterations at this learning rate we might have achieved a better model.

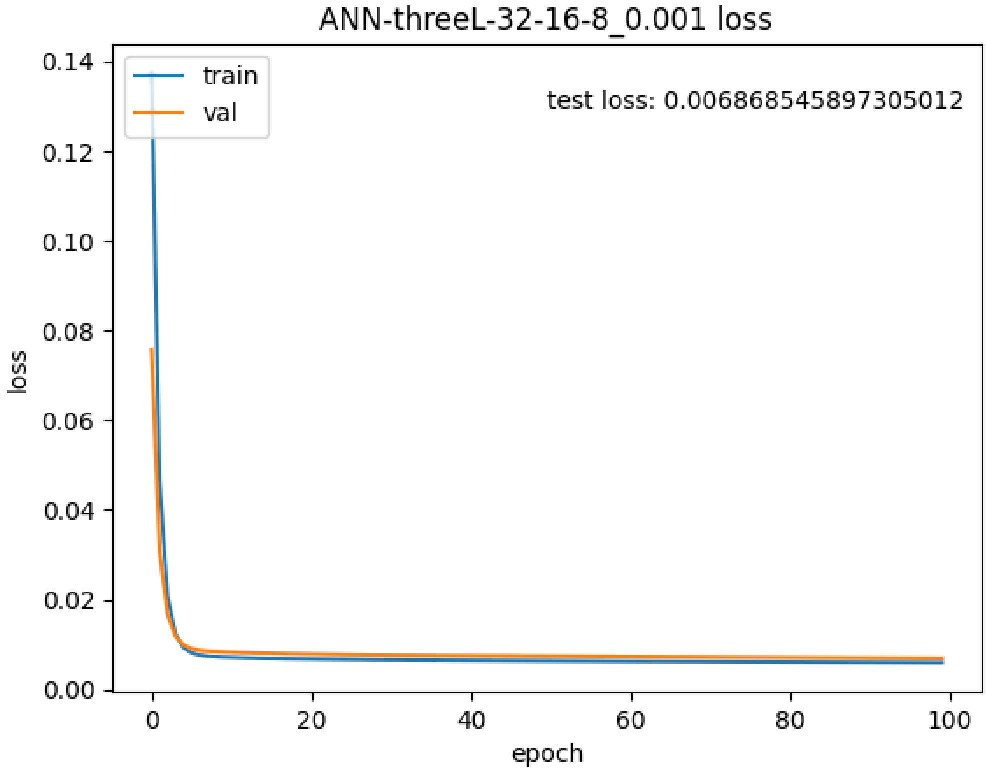
The best learning rate seems to be between 0.01 and 0.001 for our selected models.

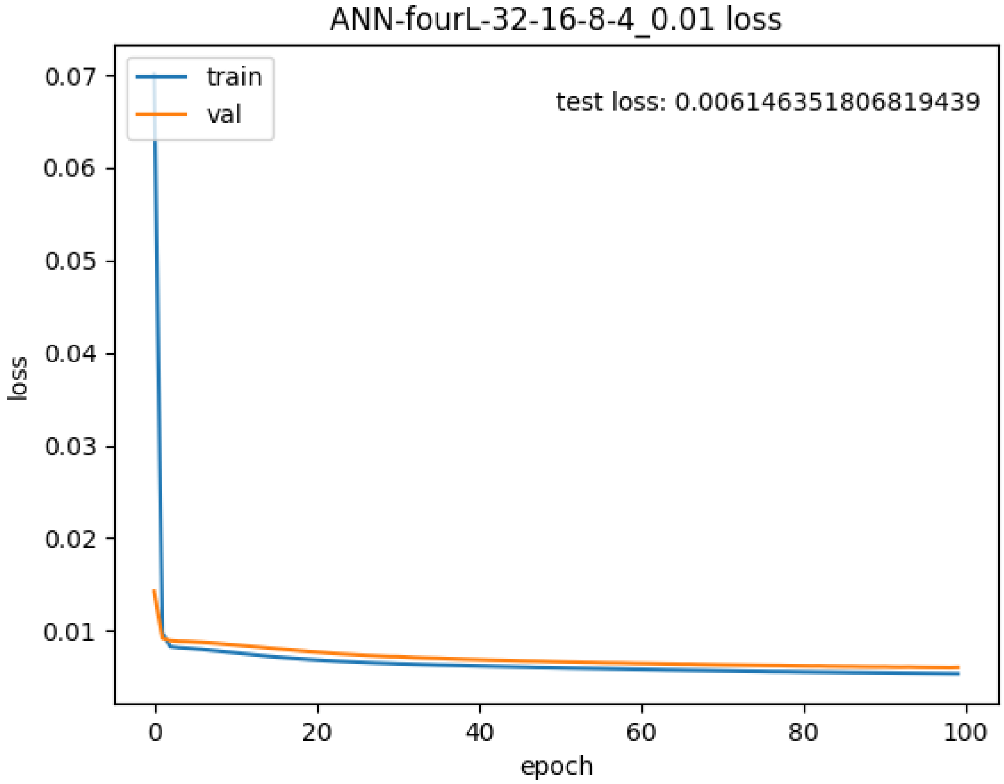
Step 6: Model Performance




ANN-oneL-16_0.01 loss
train
val
test loss: 0.005473488476127386
0.035
0.030
0.025
0.020
0.015
0.010
0.005
loss
0
20
40
60
80
100
epoch


ANN-twoL-32-8_0.001 loss
train
val
test loss: 0.007376082241535187
0.10
0.08
0.06
0.04
0.02
loss
0
20
40
60
80
100
epoch


ANN-threeL-32-16-8_0.001 loss
train
val
test loss: 0.006868545897305012
loss
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
epoch


ANN-fourL-32-16-8-4_0.01 loss
train
val
test loss: 0.006146351806819439
loss
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
20
40
60
80
100
epoch