Week 8:

Que Tran

25th April 2017

Time	Topic	Reading
8.15-	Recap of tutorials last week Lecture: Data preprocessing	J. Han and M. Kamber, Chapter 2
9.00-	Exercises on Data preprocessing	
11.45-	Lunch break	
12.15-	Questions and answers	
13.00-	Continue with previous exercises	

Rest of Day Catch up with previous tutorials

This PDF document is available in an HTML version at `http://www.hg.schaathun.net/FPIA/week08.html`.

1 Data preprocessing

At this point, the priority is to complete exercises implementing and using the neural network.

If you have time, the following exercises can give practice on the techniques from today's lecture. They require a certain level of independence and may be challenging.

1.1 Missing values

Consider your own neural network and the Credit Approval data set.

1. How can you encode the data?
2. There are 37 cases having one or more missing values. Try to fill in missing values using techniques from today's lecture. You can try using different techniques and compare the results.

1.2 Normalizing data

We know that feature A14 and A15 are in the range of [0, 2000] and [0, 100000], respectively. Normalize the data. Train and test your network again, then compare the results.

1.3 Feature selection

Try to remove features from the data set and see if it affects the performance. Which features can be removed with no penalty? Can any be removed for benefit?

1.4 Feature extraction (optional)

In this exercise, we will apply PCA to reduce number of dimensions of the data set. You can implement PCA on your own if you are confident, or you can use this package. To install the package, run:

```
cabal update
cabal install hstatistics
```

You can try different thresholds to find the principal components. Train and test your network again with the new data sets.