**Assignment 3**

**N Kapil Chandra**

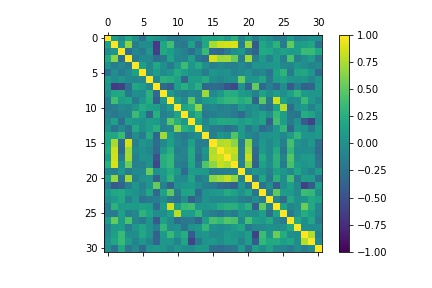
**MM14B022**

**Dataset:**

First let us have a look at the data set. The data set has 30 features and one target feature. In all the features few features have very high values of the order of 10^(5) and few features are less than 1. If we try to train our model on this data set directly it will result in the loss function not decreasing at all, that is, the model won’t train. For this we need to normalize the data set. For doing this I have used scale function in preprocessing module of sklearn library.

**Observation 1:** After normalizing the data the model will train properly with the loss function decreasing gradually. Without normalization the model won’t train.

Now looking at the features of the data, let us construct a correlation heat map to understand how each pairs of variables are correlated.



The above plot shows a heat map of correlation between all the features. If two variables lie either in yellow region or dark blue region means that they are highly positively or negatively correlated, respectively.

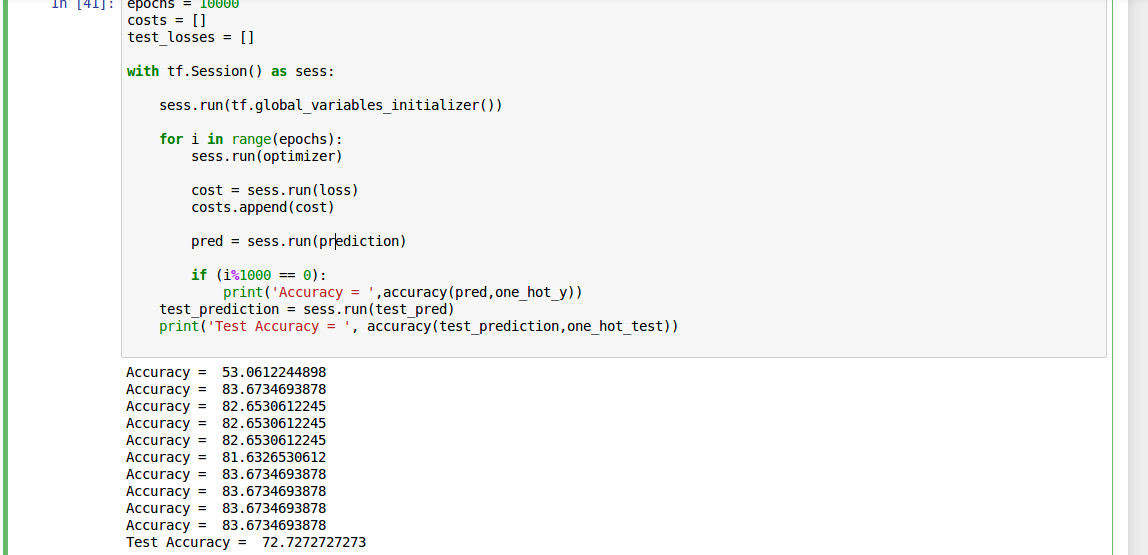
The last feature of the data set is target, to see which variables affect the target feature the most we have to see the last column in the heat map.

**Observation 2:** So the following features are most correlated with overall survival:

0\_original\_shape\_Compactness2, Age and 1\_original\_shape\_Compactness2

And to some extent : 2\_original\_shape\_SurfaceArea, 0\_original\_shape\_Maximum2DDiameterColumn, 0\_original\_firstorder\_Maximum also contribute.

**Binary Classification:**

For binary classification I have used a learning rate of 0.1 and used Adam optimizer in tensorflow. I have used regularization with regularization parameter of 0.01. I have trained the model for 10000 iterations and obtained a training accuracy of 83.673 and a test accuracy of 72.7272.

Sensitivity obtained = 0.789

Specificity obtained = 0.642

The correlation coefficient I got for predicted labels and ground truth labels is 0.437.

**Multiclass classification:**

For multiclass classification I have used a learning rate of 0.1 and used Adagrad optimizer in tensorflow. I have trained the model for 5000 iterations and obtained a training accuracy of 74.489 and a test accuracy of 57.575.



**For class 0:**

Sensitivity obtained = 0.642

Specificity obtained = 0.789

**For class 1:**

Sensitivity obtained = 0.5

Specificity obtained = 0.826

**For class 2:**

Sensitivity obtained = 0.555

Specificity obtained = 0.75

The correlation coefficient I got for predicted labels and ground truth labels is 0.377.

**Note:** I have attached two files ‘multi.py’ and ‘binary.py’ which contain code for both the cases.