Assignment 2

Machine Learning Workshop

We will develop a classifier to predict if a passenger from Titanic survived or not. Go to Kaggle web site https://www.kaggle.com/c/titanic/overview and download the training and testing data sets. (Verification: 891 data points for training and 418 data points for testing datasets)

(a.) [50 pts] Preprocess the data, impute missing values as you see fit, and remove features that you see useless.

(b.) [50 pts] Submit your predictions to Kaggle for the test dataset and report your accuracy in your submission. (You will need an account at Kaggle – use a dummy email address to protect your school/work email address, etc.) For Your Info, I achieved 79% using my preprocessing pipeline and a Random Forest classifier. Which is not the best as in Kaggle there are better results. Kaggle also has some results with 100% accuracy which cannot be taken as honest submissions in my opinion.

I used the following code to export the predictions for Kaggle:

```
def save_preds(_fn, _y_pred, _df):
    import csv
    with open(_fn, 'w') as fout:
        writer = csv.writer(fout, delimiter=',', lineterminator='\n')
        writer.writerow(['Survived', 'PassengerId'])
        for y, passengerId in zip(_y_pred, _df['PassengerId']):
            writer.writerow([y, passengerId])

save_preds('predictions_erhan.csv', y_pred, df_test_org)
```

Note that _df has to have the 'PassengerId' which I did not use for the classification.