巨量資料分析第二次作業

繳交期限：2016/05/31

1. 請以XGBoost與Gradient Boosting對於Microsoft Malware 2015 Dataset進行分析，透過參數組合分析，選擇最適合的參數組合。結果請以markdown撰寫分析，程式碼可以ipython notebook格式撰寫，請於5/20前繳交至系統。
   1. 請針對各個參數逐一進行分析，建議先固定某個參數，並逐步選擇最佳評估參數
   2. 可輔以heatmap來呈現grid search的分析結果
   3. 請留意binary與multi-class classification在評估上的差異 (Precision/Recall僅適合binary)
2. 加分題：
   1. 比較standalone與分散式運算的效能
   2. 應用於其他資料上(matrix須大於 10K \* 100)
3. 參考資料：
   1. scipy-2016-sklearn-tutorial, 13. cross-validation: <https://github.com/amueller/scipy-2016-sklearn/tree/master/notebooks>
   2. Complete Guide to Parameter Tuning in XGBoost: <https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/>
   3. Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python: <https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/>