Big_data_data_chunks

March 16, 2019

In [6]:
```
from sklearn.linear_model import SGDClassifier
from sklearn.preprocessing import MinMaxScaler
import pandas as pd
import numpy as np
streaming = pd.read_csv('large_dataset_10_7.csv', header=None, chunksize=10000)
learner = SGDClassifier(loss='log')
minmax_scaller = MinMaxScaler(feature_range=(0, 1))
cumulative_accuracy = list()
for n, chunk in enumerate(streaming):
    if  n == 0:
        minmax_scaller.fit(chunk.iloc[:,1:].values)
    X = minmax_scaller.transform(chunk.iloc[:,1:].values)
    X[X>1] = 1
    X[X<0] = 0
    y = chunk.iloc[:,0]
    if n > 8:
        cumulative_accuracy.append(learner.score(X, y))
    learner.partial_fit(X, y, classes=np.unique(y))
print("Mean accuracy in progressive validation %f" % np.mean(cumulative_accuracy))
```

```
Mean accuracy in progressive validation 0.707814
```

In [ ]: