```
In [19]: cd
         /Users/jeongmingi
```

```
In [20]: cd Desktop/DataMining2
         /Users/jeongmingi/Desktop/DataMining2
```

```
In [21]: import pickle
```

유저 나이를 구분하는데 있어서 일렬의 나이를 구분하기에는 각 class간의 data가 적기 때문에 classification(c=1.0, train_n = 6000개, 8% 정확도)보다는 regression이 필요하다.
따라서 classification은 10대, 20대, 30대... 을 구분하는데 사용한다.

```
In [22]: from sklearn.datasets import load_svmlight_file
         X, y = load_svmlight_file("data/user_age.dat")
```

```
In [23]: print X.shape, y.shape
         n_samples, n_features = X.shape

         (42934, 49683) (42934,)
```

```
In [57]: ls

         558691_589646044379538_942129678_n-1.jpg
         DataMining/
         DataMining2/
         NPKI/
         Untitled0.ipynb
         Untitled1.ipynb
         Untitled2.ipynb
         Untitled3.ipynb
         adme_analysis_temp.ipynb
         age_my_model_70.pkl
         document_classification.ipynb
         entityid_featureid.pkl
         my_model.pkl
         my_model10000.pkl
         new_clustering.ipynb
         predicted_user_apps.ipynb
         user_app_predicted.df
         user_gender 2.txt
         user_gender 3.txt
         user_gender 4.txt
         user_gender 5.txt
         user_gender.txt
         user_gender7_pred
```

```
In [59]: X_, y_ = load_svmlight_file("user_gender 2.txt")
```

```
In [75]: print X_[1]

           (0, 113)      1.0
```

```
(0, 2216)     1.0
(0, 2850)     1.0
(0, 2853)     1.0
(0, 3544)     1.0
(0, 4605)     1.0
(0, 4609)     1.0
(0, 5656)     1.0
(0, 6189)     3.0
(0, 6491)     1.0
(0, 7383)     2.0
(0, 7402)     1.0
(0, 7834)     1.0
(0, 7903)     1.0
(0, 8007)     1.0
(0, 8026)     1.0
(0, 8165)     2.0
(0, 8679)     5.0
(0, 8680)     1.0
(0, 8749)     1.0
(0, 8973)     1.0
(0, 9114)     11.0
(0, 9178)     2.0
(0, 9756)     1.0
(0, 9769)     1.0
:     :
(0, 39808)    1.0
(0, 40805)    1.0
(0, 40921)    1.0
(0, 41145)    9.0
(0, 42202)    1.0
(0, 42477)    1.0
(0, 42835)    1.0
(0, 43465)    1.0
(0, 43480)    1.0
(0, 45218)    1.0
(0, 46588)    1.0
(0, 46994)    1.0
(0, 47045)    1.0
(0, 47079)    1.0
(0, 47923)    1.0
(0, 48218)    1.0
(0, 48417)    1.0
(0, 48480)    1.0
(0, 48753)    1.0
(0, 48754)    1.0
(0, 48943)    1.0
(0, 49318)    1.0
(0, 49964)    1.0
(0, 50338)    1.0
(0, 51300)    1.0
```

In [24]:

```
from sklearn.svm import LinearSVC
```

In [25]:

```
import numpy as np
```

In [41]:

```
from sklearn.grid_search import GridSearchCV
from sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.5,

Lparam = {
         'C': np.logspace(-5, 5, 4),
}
print(Lparam)

gcv = GridSearchCV(LinearSVC(), Lparam, cv=3, n_jobs=-1)

%time _ = gcv.fit(X_train, y_train)
```

```
{'C': array([  1.00000000e-05,   2.15443469e-02,   4.64158883e+01,
         1.00000000e+05])}
CPU times: user 32.09 s, sys: 0.71 s, total: 32.80 s
Wall time: 127.75 s
```

In [42]:

```
gcv.best_params_, gcv.best_score_
```

Out[42]:

```
({'C': 0.021544346900318846}, 0.69874703408528216)
```

In [43]:

```
L_svc = LinearSVC(C=0.021544).fit(X_train,y_train)
L_svc.score(X_train,y_train), L_svc.score(X_test,y_test)
```

Out[43]:

```
(0.90632133041412399, 0.70219406530954487)
```

In [47]:

```
from sklearn.externals import joblib

filename = "age_my_model_70.pkl"
joblib.dump(L_svc, filename, compress=9)
```

Out[47]:

```
['age_my_model_70.pkl']
```

In [46]:

```
pwd
```

Out[46]:

```
u'/Users/jeongmingi/Desktop/DataMining2'
```

In [ ]: