Homework 2 Report - Income Prediction

學號：b04705026 系級：資管三 姓名：林彥廷

1. (1%) 請比較你實作的generative model、logistic regression的準確率，何者較佳？

Submission and Description	Private Score	Public Score
logistic.csv a day ago by Adam add submission details	0.84424	0.85257
generative.csv a day ago by Adam add submission details	0.81574	0.81695

Ridge Logistic Regression在Public和Private優於Generative Model。

Gaussian Generative Model和Logistic Regression是一種Generative-Discriminative pair，理論上，在這大資料量的問題上，Discriminative Model隨資料大小的增加，收斂的速度會比較慢，但會收斂到比較低的錯誤率*(AY Ng& MI Jordan 2002)*。另外，logistic regression有加上regularization，搜尋在validation set表現最好的lambda，增進generalize的能力、避免overfitting。

2. (1%) 請說明你實作的best model，其訓練方式和準確率為何？

Submission and Description	Private Score	Public Score
svm.csv a day ago by Adam add submission details	0.85517	0.85761

我使用SVM with linear kernel，feature標準化，第一、五個連續變數增加平方項，private和public分數都比logistic和generative來得好。

除了特徵轉換，表現較好的原因我認為是SVM的"Hindge loss"相較於logistic的cross-entropy來的更貼近0-1 Error 也是我們評分的標準 *(Yoonkyung Lee, Yi Lin & Grace Wahba 2004)*。另外也有搜尋在validation set表現最好的C (類似 1/lambda)，來對抗雜訊。

3. (1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關normalization請參考：https://goo.gl/XBM3aE)

Submission and Description	Private Score	Public Score
logistic_normalized.csv a day ago by Adam add submission details	0.84449	0.84938
logistic_UN-normalized.csv a day ago by Adam add submission details	0.79056	0.80159

我使用"Mean normalization"，使用sample mena(trainning set)當作每個feature population mean的estimatation，可以讓每一個feature都在[-1,1]之間。

Normalized Model在private或public都比較好，我認為兩個主因如下。

★ fnlwgt的variance和range和其他變數差異過大，gradient在fnlwgt的方向也會太大，造成其他參數的更新會被忽略，而收斂不到global minimum.

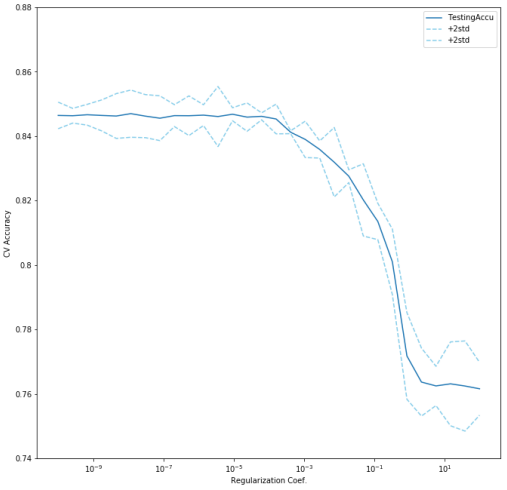
```
In [172]: from scipy.stats import describe
          describe(df['fnlwgt'])

Out[172]: DescribeResult(nobs=32561, minmax=(12285, 1484705), mean=189778.36651208502, variance=11140797791.841892, skewness=1.
          446913435142329, kurtosis=6.217671807559244)
```

★ 如果不進行normalization，training的時候會出現**overflow**，中間幾次更新的參數可能會是錯誤的數值，造成最後結果不好。

```
/home/adam/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:60: RuntimeWarning: overflow encountered in ex
p
/home/adam/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:60: RuntimeWarning: invalid value encountered
in true_divide
/home/adam/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:61: RuntimeWarning: overflow encountered in ex
p
```

4. (1%) 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。(有關regularization請參考：https://goo.gl/SSWGhf P.35)



L2-Regularization可以降低模型複雜度，約束最終參數的norm，通常“Accuracy vs. Lambda”的圖都會有一個“U-shape”，但是這在本次作業並不明顯，即可推測下兩點。

1. 資料雖然有noise，但本次資料量夠大夠多彼此可以相消，學習的時候也不太會過於overfit。

2. 模型的配適度可能藉由更多”特徵轉換”搭配上更強的regularization，也許可以fit更好，同時獲得更好的Testing Accuracy。

5. (1%) 請討論你認為哪個attribute對結果影響最大？

Standardized每個特徵，套入完模型型後，藉由係數的數值、正負，判斷特徵之間的相對大小，完整的正反影響前十排名如下圖。

- 正面影響：
  - Native Country 原生國家：這資料美國(佔70%⬆)，<u>其他國家樣本數很少，無法推斷其他原生國收入之關西。</u>
  - 教育程度：有高等教育程度的人，有高收入，<u>對比負面影響的變數，更可以證實這項推論。</u>
  - 資本收入：大部分人的收入來自於薪資，所以有資本收入的人很可能是有”閒錢”的人。
- 負面影響：
  - 家庭因素：離婚、喪偶都是嚴重的負面影響。
  - 職業：無工作者、藍領階級收入較低。

```
w = model.w_log.reshape((-1,))
w = np.delete(w, 0)
print('Top 10 Positively influential variables:\n', df.columns[np.argsort(w)[::-1]][:10].values)
print('\nTop 10 Negatively influential variables:\n', df.columns[np.argsort(w)[::+1]][:10].values)
```

```
Top 10 Positively influential variables:
 ['native_country_ United-States' 'capital_gain' 'native_country_ Mexico'
 'native_country_ ?' 'marital_status_ Married-civ-spouse'
 'workclass_ Private' 'relationship_ Husband' 'education_ HS-grad'
 'education_ Some-college' 'education_ Bachelors']

Top 10 Negatively influential variables:
 ['education_ Preschool' 'occupation_ Priv-house-serv'
 'workclass_ Without-pay' 'workclass_ Never-worked'
 'occupation_ Armed-Forces' 'marital_status_ Never-married' 'race_ Other'
 'fnlwgt' 'occupation_ Farming-fishing'
 'marital_status_ Married-spouse-absent']
```

References

[1] Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems. 2002.

[2] Lee, Yoonkyung, Yi Lin, and Grace Wahba. "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data." Journal of the American Statistical Association 99.465 (2004): 67-81.