學號：R07922004 系級：資工碩一 姓名：吳星耀

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

	Public Score	Private Score	Average Score
generative model	0.84729	0.84178	0.84454
logistic regression	0.85184	0.85345	0.85265

logistic regression的結果比generative model好。原因可能是如教授在影片（ML Lecture 5:logistic Regression）中講的，由於generative假設資料是來自於Gaussian的機率模型，而此假設可能在這次作業的資料不成立，因此這個假設可能反而會讓它沒辦法得到更好的function。

2. 請說明你實作的best model，其訓練方式和準確率為何？

在我實作最好的模型中，我取了所有一次項、並加上除了one-hot encoding之外的feature的二次項、除了one-hot encoding與fnlwgt的三次項，之後再做normalization與logistic gradient descent。logistic gradient descent部份我用adagrad實作，並將adagrad learning rate係數調為0.1，iteration次數為3000，weight與bias的初始值皆為0.1，並做了λ=0.1的regularization。

這個model在public score與private score的結果分別為0.85714、0.85737。

在Kaggle期限截止過後我有試過將iteration增加至10000與再增加更高次的feature，我發現如此能再提昇更多準確度。

3. 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

以logistic regression來看：

如果未使用特徵標準化，在使用有adagrad的gradient descent做了3000次更新（iteration=3000）後，在training set的準確度只有0.7798，上傳到kaggle的public、private 準確度更是只有0.26375、0.26630。將iteration提高到10000後，在training set上的準確度為0.7938，kaggle的public、private 準確度各為0.79975、0.79891。而做了normalization後，在iteration=3000的時候training accuracy已經達到0.85332，在kaggle上更是有public: 0.85184、private: 0.84178的準確度。因此可以判斷normalization能夠大幅加速gradient descent的速度。

以generative model來看：

在沒有做normalization的情況下generative model在training set的準確度為0.84236，且在kaggle 的public / private準確度為 0.84582 / 0.84141；而有做normalization後，在training set的準確度為0.84223，在kaggle 的public / private準確度為 0.84729 / 0.84178，準確度為有做normalization的稍微好一點，但差距甚小。由於normalization的目的是為了讓不同feature間的差距幅度減少，以提昇像是gradient descent這種會有『因為其中一個feature過大造成其他feature對更新影響變小』特性的運算的效率。而

generative model在一開始就假設資料是由高斯分布產生，因此在由covariance matrix、mean計算出參數的時候就已經隱含著normalization的特性，因此再做一次normalization的效果就不太顯著。

4. 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。

   logistic regression, 一次項feature, adagrad, iteration = 10000

λ	training accuracy	private / public accuracy
1	0.85326	0.85333 / 0.85184
0.01	0.85329	0.85345 / 0.85184
0.001	0.85323	0.85345 / 0.85184
0	0.85332	0.85345 / 0.85171

   由上表可以發現，regularization在這些feature下的效果也不顯著，但是在λ＝0.01跟0.001時在kaggle上的平均準確度較高，而λ＝0時準確度微微下降，這可能是overfit的影響。而在λ＝1的時候準確度也稍稍下降，這可能是underfit的結果。

5. 請討論你認為哪個attribute 對結果影響最大？

   我以實驗來判斷哪個feature可能比較重要。每次將一種feature刪除，取剩下的feature來train，得到一組training set accuracy，將accuracy依照大小由小到大排列，越小的代表該被刪除的feature對accuracy越重要。得到結果accuracy由小至大排列為：

   capital_gain, hours_per_week, capital_loss, age, fnlwgt, sex, Protective-serv, 7th-8th, Tech-support, 5th-6th, Prof-school, France, Unmarried, Vietnam, Own-child, Wife, England, Germany, South, 11th, 9th, Separated, Columbia, Iran, Ireland, Italy, Federal-gov, 12th, Bachelors, Doctorate, Masters, Never-married, Armed-Forces, Transport-moving, Husband, Other, Dominican-Republic, Greece, Haiti, Local-gov, Never-worked, Private, Self-emp-inc, State-gov,?_workclass, 10th, 1st-4th, Assoc-voc, HS-grad, Some-college, Adm-clerical, Craft-repair, Handlers-cleaners, Machine-op-inspct, ?_occupation, Amer-Indian-Eskimo, Asian-Pac-Islander, Black, White, Cambodia, China, Cuba, Ecuador, El-Salvador, Guatemala, Holand-Netherlands, Honduras, Hong, India, Laos, Mexico, Nicaragua, Peru, Portugal, Puerto-Rico, Scotland, Taiwan, Trinadad&Tobago, United-States, ?_native_country, Self-emp-not-inc, Without-pay,Assoc-acdm, Preschool, Divorced, Married-spouse-absent, Other-relative, Canada, Hungary, Outlying-US(Guam-USVI-etc), Poland, Thailand, Married-AF-spouse, Married-civ-spouse, Exec-managerial, Farming-fishing, Priv-house-serv, Prof-specialty, Sales, Jamaica, Philippines, Yugoslavia, Widowed, Not-in-family, Other-service, Japan

   前三名中，capital_gain與hours_per_week可以聯想到與高收入有較大關聯；capital_loss可能則與較低收入有關連。

   因此我覺得最重要的attribute為capital_gain。