學號：R05922085 系級： 資工碩一 姓名：周宗鴻

1. 請說明你實作的 generative model，其訓練方式和準確率為何？

答：

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

我使用的機率模型為老師上課所使用的 Gaussian Distribution，covariance matrix 為共用。首先將 training set 分成兩組，一組為 class 1、一組為 class 2，接著分別求他們最佳的 mean 以及 covariance matrix，然後將 covariance matrix 依照 class 1 與 class 2 比例做平均，最後 2 組 mean 以及 covariance matrix 即為訓練完得到的結果。透過這種訓練方式能夠在 Kaggle Public Score 上得到的分數為 0.83452。

2.請說明你實作的 discriminative model，其訓練方式和準確率為何？

答：

首先新增兩組 feature，此兩組 features 為 training set 的 age 與 hours_per_week 取平方。接著對所有 feature 做標準化，接著實作 adagrad。透過這種訓練方式能夠在 Kaggle Public Score 上得到的分數為 0.85516。

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。

答：

```
def normalize(data, mean = None, sigma = None):
    if mean == None and sigma == None:
        mean = numpy.mean(data)
        sigma = numpy.std(data, ddof = 1)
    return numpy.divide(numpy.add(data, -1 * mean), sigma), mean, sigma
```

在 discriminative model 下 feature 做 normalization 讓收斂速度大幅度的提高，因此準確率在相同訓練時間下也會較高。

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。

答：

```
w = -1 * numpy.dot(features.T, differences) + regularization_rate * weights
```

使用正規化可以避免 overfitting 的問題，但在本次作業使用 regularization 無法提升準確率(降低一點)，且當 regularization rate 設大反而還會造成準確率嚴重下降的情形。

5. 請討論你認為哪個 attribute 對結果影響最大？

答：

我將 feature 逐一移除測試，經過測試後認為 Local-gov 對結果影響較大。