1. 比较你 handcraft 的 generative model、logistic regression 的准确率，哪一个好？

使用 generative model 的结果：

```
在训练集的验证集上验证结果：
generation model 准确率 = 0.768428

/home/ddrh/Software/python_venv/general/li
es/ipykernel_launcher.py:2: RuntimeWarning
in exp

在测试集上验证结果：
测试集准确率 = 0.236226
```

使用 logistic regression 的结果：

```
在训练集的验证集上验证结果：
logistic regression model 准确率 = 0.243857

在测试集上验证结果：
logistic regression model 准确率 = 0.785517
```

从结果来看 logistic regression 效果更好。

2. 请說明你 handcraft 的 best model 的训练方式和准确率是怎样的？

参数 logistics regression：

采用 SGD 随机梯度下降，设置 batch_size = 32，学习率 0.001，一共训练了 300 轮

3. handcraft 输入特征标准化(feature normalization)并讨论其对你的模型准确率的影响

在对输入特征进行标准化后，效果如下：

generative model：

```
在训练集的验证集上验证结果：
generation model 准确率 = 0.767199
```

```
/home/ddrh/Software/python_venv/ger
es/ipykernel_launcher.py:2: Runtime
in exp
```

```
在测试集上验证结果：
测试集准确率 = 0.763774
```

logistic regression：

```
在训练集的验证集上验证结果：
logistic regression model 准确率 = 0.845209
```

```
在测试集上验证结果：
logistic regression model 准确率 = 0.853817
```

对比 2 可知，标准化对预测结果影响很大。

4. 请实作 logistic regression 的正则化(regularization)，并讨论其对你的模型准确率的影响。

参数设置上，正则化参数设置为 10，得到的结果如下：

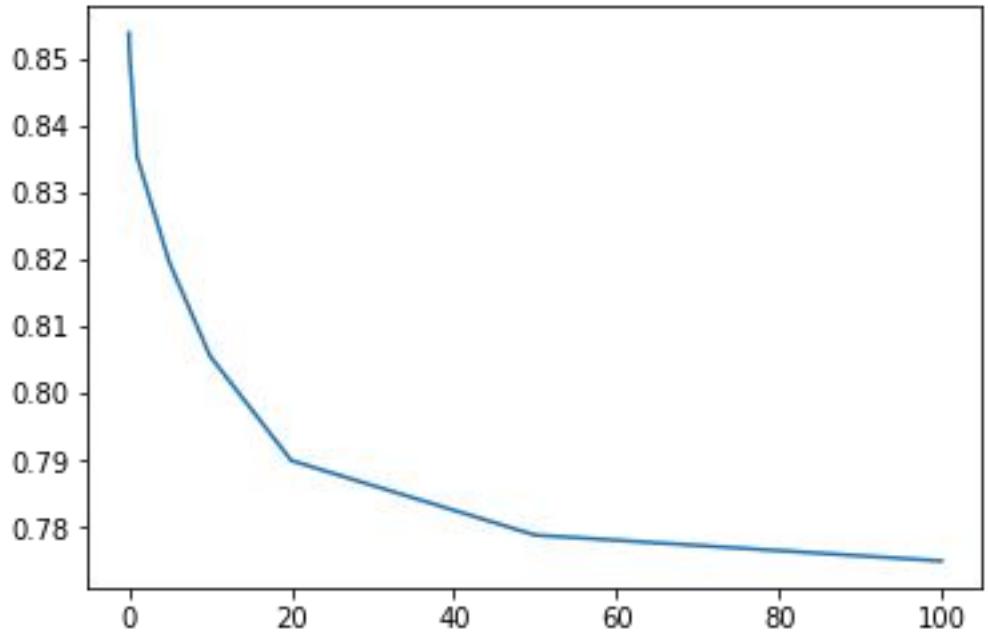
```
在训练集的验证集上验证结果：
logistic regression model 准确率 = 0.791155
```

```
在测试集上验证结果：
logistic regression model 准确率 = 0.805049
```

设置多个正则化参数：

regs = [0.0001, 0.01, 0.1, 1, 5, 10, 20, 50, 100]

得到准确率变化曲线：



可以看出，模型预测的准确率下降了。
随着正则化参数提高而下降，此数据集不适合采用正则化。

5. 请讨论你觉得哪个 attribute 对結果影响最大？

我觉的职业（workclass）对结果影响最大
收入多少主要还是取决于工作的类型。
就像蓝领和白领存在工资差一样。

6. 标明最终 testing set (public/private) 的 accuracy.

```
在训练集的验证集上验证结果：
logistic regression model 准确率 = 0.845209
在测试集上验证结果：
logistic regression model 准确率 = 0.853142
```