學號：T05902136 系級： 資工一 姓名：Jenny Zhang(張臻凝) 母語：English

*all accuracies are from Kaggle private scores

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

答：

The generative model that I chose was gaussian distribution. It is based on the sample code that the TA provided. The Gaussian distribution calculates the probability of a certain class based on the input parameters. Firstly, I separated the data into two classes: >50k and <=50k. After data means and covariances had been calculated, I used the posterior probability formula provided on the pdf to compute the result.

$$z = \underbrace{(\mu^1 - \mu^2)^T \Sigma^{-1}}_{\boldsymbol{w^T}} x \underbrace{- \frac{1}{2}(\mu^1)^T \Sigma^{-1} \mu^1 + \frac{1}{2}(\mu^2)^T \Sigma^{-1} \mu^2 + ln \frac{N_1}{N_2}}_{\text{b}}$$

While looking at the train data, I found that there were missing data for some categories. I added data imputation to the code to reduce problems that missing data might cause, such as bias introduction and efficiency reduction.

With imputation?	Accuracy
Yes	84.67%
No	84.60%

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

答：

The discriminative model that I chose was logistic regression. It measured the relationship between categorical dependent variables and independent variables using a sigmoid function.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

First of all, I picked an initial weight vector and bias. Using gradient descent, I ran iterations with batches. In each epoch, the training data was randomly shuffled. Cross entropy was calculated to see if it had been minimized. To improve my accuracy and reduce unnecessary data columns, I added categorization and data imputation to my code. For example, I grouped all columns where people dropped out before 12th grade as one column and “Local-gov” column and “ State-gov” column as one column. I omitted fnlwgt because the numbers were so big that it skewed the resulting function. The best accuracy that I got was 85.26%.

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：

Feature normalization rescales data to a smaller range, usually [0,1] or [-1,1]. In this case, age, capital gain, capital loss and hours per week were scaled because all other data was in binary. Fnlwgt was omitted due to the reason stated in the previous answer. If we did not use

scaling, they would impact the result more significantly than others and thus giving us a less accurate prediction.

With feature nomalization?	Accuracy
Yes	85.26%
No	84.92%

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：

Regularization was done by adding $\lambda * W$ to the gradient descent vector. Because there were a lot of features in our training data, it was likely that an overfitting would occur. Regularization is an optimization technique that could help take care of this problem. In addition, from the table below, we could see that picking the right lambda would impact our accuracy as well.

With regularization?	Accuracy
Yes (λ = 0.001)	82.82%
Yes (λ = 0.0001)	85.26%
No	85.03%

5.請討論你認為哪個attribute對結果影響最大？

Positive impacts:

-age: Salary usually grows as people get older until they reach their retired age.

-hours per week: The more hours they work, the more likely they will make more money.

-married: people who are married usually have a steady financial income because they have to support the family.

-exec-managerial: from looking at the training data, I see that many people whose occupation is in exec-managerial make over 50k. This makes sense because they play an important role in their company.

-United States: United States is an economic powerhouse.

Negative impacts:

-sex: due to gender inequality, being a female will less likely make over 50k.

-low education level (HS-grad and drop-outs): people with lower educational level usually get a job that pays a lot less because they don't have enough qualifications.