COM 525000 Statistical Learning
Computer Homework

(Due January 18, 2021)

Note: Please complete the following assignments using python3.x (**Without using the linear regression, ridge regression, cross-validation or logistic regression packages**). Also submit your report and code on iLMS. You should name your code like problem1(a), problem1(b), and so on.

1. Linear Regression (8%+10%+2%+2%) In this problem, let us investigate the relationship between the fat intake from different kinds of food and confirmed percentage in each country. The dataset is from (https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset) that is used to build various regression models.
Please use ***covid-19.csv*** to complete the following questions.

(a) In the ***covid-19.csv***, there are 6 predictors including the fat intake from X_1=*seafood*, X_2=*meat*, X_3=*offals*, X_4=*spices*, X_5=*vegetables* and obesity ratio X_6=*obesity*. Please write a program to complete the following table.

	Coefficient	Std. Error	t-statistic	p-value
(intercept)				
X_1				
X_2				
X_3				
X_4				
X_5				
X_6				

(b) Suppose that the data is fitted by using only two of the above predictors. Write a program to find the choice that yields the smallest RSS, and complete the following table.

	Coefficient	Std. Error	t-statistic	p-value
(Feature 1)				
(Feature 2)				

Please specify the feature 1 and feature 2 you found.

(c) Similar to (b), please use **ridge regression** to find the two features that yield the smallest RSS. Note that the tuning parameter λ should be searched by 5-fold cross-validation (CV) where $\lambda \in [10^{-5}, 10^{-4.5}, \ldots, 10^{2.5}, 10^{3}]$. Also, plot the curve of CV error versus λ in log-scale. Please specify the value of λ what you choose, and complete the following table.

	Coefficient
(Feature 1)	
(Feature 2)	

Find the features and compare the RSE and R^2 between (a), (b), and (c).

2. Logistic Regression (10%+8%) In this problem, we use Kannada handwritten characters dataset (https://www.kaggle.com/c/Kannada-MNIST) and we are going to use first three Kannada characters to train our logistic regression model. Each image size is 28×28 pixels, there are total 6000 images of each character for training and 1024 images for testing. We provide the dataset in the attached files 'train.csv' and 'test.csv'. Note that the first column is true labels and the remaining columns denote pixel values of flatten images.

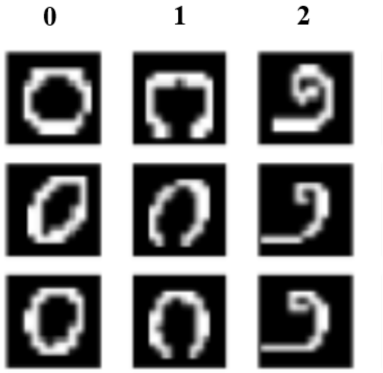

Figure 1: Images from Kannada MNIST Dataset.

(a) Build a logistic regression model to classify the **second** and **third** characters (i.e. digit $1, 2$ in Fig. 1). Please specify your step size, initial value of β and plot the confusion matrix.

Hint : Use gradient descent to compute the coefficients. (Without any packages)

$$\beta(k+1) = \beta(k) - \eta \nabla J(\beta(k))$$

where $\nabla J(\beta(k)) = -\sum_{i=1}^{n} \frac{(2y_i-1)x_i}{1+e^{(2y_i-1)\beta(k)^T x_i}}$.

(b) Use Linear Discriminant Analysis (LDA) to classify **first three** characters as shown in Fig. 1, and plot the confusion matrix.