Statistical Learning Theory,
Assignment 1 – Logistic Regression

Due: 11:59PM, October 19, 2020
Issued: October 05, 2019

Instructions:

- Implement "Regularized Logistic Regression" and apply it to the dataset "Wisconsin Breast Cancer Database" described below.
- Be sure to train your regression model using "Train_Data.txt," and evaluate your model's accuracy using "Test_Data.txt."
- Feel free to use the ipynb codes distributed during the class, and modify them to complete this assignment.
- You may **not** use any other languages other than Python to accomplish your task.

Submission:

- Write the following items in a pdf document and submit the pdf file to the Cyber.ewha.ac.kr assignment dropbox by the deadline.
  ①. Define what your features/attributes are in your model. Appropriate features should be selected (*e.g.*, [x1, x2, x7] or all of them) or defined (*e.g.*, x1•x2, $x1^2$]) to improve the model's performance.
  ②. Report your logistic regression model's train accuracy (%), test accuracy (%), and show a plot of loss log over epoch during training.
  ③. Select the appropriate types (L1, L2, or L1+L2) of regularization parameter λ and its value (e.g., λ=10), and justify your choice of λ.
  ④. Report your ***regularized*** logistic regression model's train accuracy (%), test accuracy (%), and show a plot of loss log over epoch during training.
  ⑤. What efforts have you made to improve accuracy of your final model? (e.g., I defined new feature using the given features or chose … attributes as a feature for my model as follows…) Report your final best model's test accuracy (%).
- Submit two files including (1) your source codes (format: ipynb) and (2) your answers (format: pdf). Please upload only the two files.

Grading:

- We will review and comment on your submission regarding the style of your Python/PyTorch code. You must attempt every question in order to receive credit.
- If your final best model's performance (based on test accuracy [%]) is in the top 20, you will be given a 0.5 point bonus (out of 5).

Note:

- If you submit the assignment late, we will deduct the assignment score by 20% per day and will not accept submissions after the solution has been distributed. The solution will

be uploaded 2 days after the deadline.

Dataset information:

- There are 569 items (patients). There is an ID followed by 10 predictors variables (thickness, cell size uniformity, etc.) The variable to predict is encoded as 2 (benign) or 4 (malignant). Sample:

```
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
. . .
1017122,8,10,10,8,7,10,9,7,1,4
1018099,1,1,1,1,2,10,3,1,1,2
. . .
```

1. Title: Wisconsin Breast Cancer Database (January 8, 1991)
2. Number of Instances: 699 (as of 15 July 1992)
3. Number of Attributes: 10 plus the class attribute
4. Attribute Information: (the last column is class attributes)

#	Attribute	Domain
1)	Sample code number	id number
2)	Clump Thickness	1 - 10
3)	Uniformity of Cell Size	1 - 10
4)	Uniformity of Cell Shape	1 - 10
5)	Marginal Adhesion	1 - 10
6)	Single Epithelial Cell Size	1 - 10
7)	Bare Nuclei	1 - 10
8)	Bland Chromatin	1 - 10
9)	Normal Nucleoli	1 - 10
10)	Mitoses	1 - 10
11)	Class:	(2 for benign, 4 for malignant)

8. Missing attribute values: 16
   There are 16 instances in the datasets that contain a single missing (i.e., unavailable) attribute value, now denoted by "?".

9. Class distribution:
   Benign: 458 (65.5%)
   Malignant: 241 (34.5%)