Implement SGD Classifier with Logloss and L2 regularization Using SGD without using sklearn

There will be some functions that start with the word "grader" ex: grader_weights(), grader_sigmoid(), grader_logloss() etc, you should not change those function definition.

Every Grader function has to return True.

Importing packages

In [1]:

```
import numpy as np
import pandas as pd
import math
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import linear_model
```

Creating custom dataset

In [2]:

```
# please don't change random_state
X, y = make_classification(n_samples=50000, n_features=15, n_informative=10, n_redundant=5,
                           n_classes=2, weights=[0.7], class_sep=0.7, random_state=15)
# make_classification is used to create custom dataset
# Please check this link (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_class
ification.html) for more details
```

In [3]:

```
X.shape, y.shape
```

Out[3]:

```
((50000, 15), (50000,))
```

Splitting data into train and test

In [4]:

```
#please don't change random state
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=15)
```

In [5]:

```
# Standardizing the data.
scaler = StandardScaler()
x_train = scaler.fit_transform(X_train)
x_test = scaler.transform(X_test)
```

In [6]:

```
X_train.shape, y_train.shape, X_test.shape, y_test.shape
```

Out[6]:

```
((37500, 15), (37500,), (12500, 15), (12500,))
```

SGD classifier

In [7]:

```
# alpha : float
# Constant that multiplies the regularization term.

# eta0 : double
# The initial learning rate for the 'constant', 'invscaling' or 'adaptive' schedules.

clf = linear_model.SGDClassifier(eta0=0.0001, alpha=0.0001, loss='log', random_state=15, penalty='l2',
tol=1e-3, verbose=2, learning_rate='constant')
clf
# Please check this documentation (https://scikit-learn.org/stable/modules/generated/sklearn.linear_mod
el.SGDClassifier.html)
```

Out[7]:

```
SGDClassifier(alpha=0.0001, average=False, class_weight=None,
              early_stopping=False, epsilon=0.1, eta0=0.0001,
              fit_intercept=True, l1_ratio=0.15, learning_rate='constant',
              loss='log', max_iter=1000, n_iter_no_change=5, n_jobs=None,
              penalty='l2', power_t=0.5, random_state=15, shuffle=True,
              tol=0.001, validation_fraction=0.1, verbose=2, warm_start=False)
```

In [8]:

```
clf.fit(X=X_train, y=y_train) # fitting our model
```

```
-- Epoch 1
Norm: 0.77, NNZs: 15, Bias: -0.316653, T: 37500, Avg. loss: 0.455552
Total training time: 0.01 seconds.
-- Epoch 2
Norm: 0.91, NNZs: 15, Bias: -0.472747, T: 75000, Avg. loss: 0.394686
Total training time: 0.03 seconds.
-- Epoch 3
Norm: 0.98, NNZs: 15, Bias: -0.580082, T: 112500, Avg. loss: 0.385711
Total training time: 0.04 seconds.
-- Epoch 4
Norm: 1.02, NNZs: 15, Bias: -0.658292, T: 150000, Avg. loss: 0.382083
Total training time: 0.05 seconds.
-- Epoch 5
Norm: 1.04, NNZs: 15, Bias: -0.719528, T: 187500, Avg. loss: 0.380486
Total training time: 0.06 seconds.
-- Epoch 6
Norm: 1.05, NNZs: 15, Bias: -0.763409, T: 225000, Avg. loss: 0.379578
Total training time: 0.07 seconds.
-- Epoch 7
Norm: 1.06, NNZs: 15, Bias: -0.795106, T: 262500, Avg. loss: 0.379150
Total training time: 0.09 seconds.
-- Epoch 8
Norm: 1.06, NNZs: 15, Bias: -0.819925, T: 300000, Avg. loss: 0.378856
Total training time: 0.10 seconds.
-- Epoch 9
Norm: 1.07, NNZs: 15, Bias: -0.837805, T: 337500, Avg. loss: 0.378585
Total training time: 0.11 seconds.
-- Epoch 10
Norm: 1.08, NNZs: 15, Bias: -0.853138, T: 375000, Avg. loss: 0.378630
Total training time: 0.12 seconds.
Convergence after 10 epochs took 0.12 seconds
```

Out[8]:

```
SGDClassifier(alpha=0.0001, average=False, class_weight=None,
              early_stopping=False, epsilon=0.1, eta0=0.0001,
              fit_intercept=True, l1_ratio=0.15, learning_rate='constant',
              loss='log', max_iter=1000, n_iter_no_change=5, n_jobs=None,
              penalty='l2', power_t=0.5, random_state=15, shuffle=True,
              tol=0.001, validation_fraction=0.1, verbose=2, warm_start=False)
```

In [9]:

```
clf.coef_, clf.coef_.shape, clf.intercept_
#clf.coef_ will return the weights
#clf.coef_.shape will return the shape of weights
#clf.intercept_ will return the intercept term
```

Out[9]:

```
(array([[-0.42336692,  0.18547565, -0.14859036,  0.34144407, -0.2081867 ,
          0.56016579, -0.45242483, -0.09408813,  0.2092732 ,  0.18084126,
          0.19705191,  0.00421916, -0.0796037 ,  0.33852802,  0.02266721]]),
 (1, 15),
 array([-0.8531383]))
```

```
# This is formatted as code
```

Implement Logistic Regression with L2 regularization Using SGD: without using sklearn

1. We will be giving you some functions, please write code in that functions only.
2. After every function, we will be giving you expected output, please make sure that you get that output.

- Initialize the weight_vector and intercept term to zeros (Write your code in def initialize_weights())
- Create a loss function (Write your code in def logloss())

  $log loss = -1*\frac{1}{n}\Sigma_{for each Yt,Y_{pred}}(Ytlog10(Y_{pred})+(1-Yt)log10(1-Y_{pred}))$
- for each epoch:
  - for each batch of data points in train: (keep batch size=1)
    - calculate the gradient of loss function w.r.t each weight in weight vector (write your code in def gradient_dw())

      $dw^{(t)} = x_n(y_n − σ((w^{(t)})^{T} x_n+b^{t}))- \frac{λ}{N}w^{(t)})$
    - Calculate the gradient of the intercept (write your code in def gradient_db()) check this

      $ db^{(t)} = y_n- σ((w^{(t)})^{T} x_n+b^{t}))$
    - Update weights and intercept (check the equation number 32 in the above mentioned pdf):
      $w^{(t+1)}← w^{(t)}+α(dw^{(t)}) $

      $b^{(t+1)}←b^{(t)}+α(db^{(t)}) $
  - calculate the log loss for train and test with the updated weights (you can check the python assignment 10th question)
  - And if you wish, you can compare the previous loss and the current loss, if it is not updating, then you can stop the training
  - append this loss in the list (this will be used to see how loss is changing for each epoch after the training is over)

Initialize weights

In [10]:

```
def initialize_weights(dim):
    ''' In this function, we will initialize our weights and bias'''
    #initialize the weights to zeros array of (1,dim) dimensions
    #you use zeros_like function to initialize zero, check this link https://docs.scipy.org/doc/numpy/r
eference/generated/numpy.zeros_like.html
    w = np.zeros_like(dim)
    #initialize bias to zero
    b = 0
    return w,b
```

Grader function - 1

In [11]:

```
dim=X_train[0]
w,b = initialize_weights(dim)
def grader_weights(w,b):
  assert((len(w)==len(dim)) and b==0 and np.sum(w)==0.0)
  return True
grader_weights(w,b)
```

Out[11]:

True

Compute sigmoid

$sigmoid(z)= 1/(1+exp(-z))$

In [12]:

```
def sigmoid(z):
    ''' In this function, we will return sigmoid of z'''
    # compute sigmoid(z) and return
    val = 1/(1+np.exp(-z))

    return val
```

Grader function - 2

In [13]:

```
def grader_sigmoid(z):
  val=sigmoid(z)
  assert(val==0.8807970779778823)
  return True
grader_sigmoid(2)
```

Out[13]:

True

Compute loss

$log loss = -1*\frac{1}{n}\Sigma_{for each Yt,Y_{pred}}(Ytlog10(Y_{pred})+(1-Yt)log10(1-Y_{pred}))$

In [14]:

```
def logloss(y_true,y_pred):
    '''In this function, we will compute log loss '''
    length = len(y_true)
    loss = 0.0
    for i in range(length):
      loss += y_true[i]*(np.log10(y_pred[i]))+(1-y_true[i])*(np.log10(1-y_pred[i]))
    loss = -(loss/length)
    return loss
```

Grader function - 3

In [15]:

```
def grader_logloss(true,pred):
  loss=logloss(true,pred)
  assert(loss==0.07644900402910389)
  return True
true=[1,1,0,1,0]
pred=[0.9,0.8,0.1,0.8,0.2]
grader_logloss(true,pred)
```

Out[15]:

True

Compute gradient w.r.to 'w'

$dw^{(t)} = x_n(y_n − σ((w^{(t)})^{T} x_n+b^{t}))- \frac{λ}{N}w^{(t)}$

In [16]:

```
def gradient_dw(x,y,w,b,alpha,N):
    '''In this function, we will compute the gardient w.r.to w '''
    dw = x*(y-sigmoid(np.dot(w.T,x)+b))-(alpha/N)*w
    return dw
```

Grader function - 4

In [17]:

```
def grader_dw(x,y,w,b,alpha,N):
  grad_dw=gradient_dw(x,y,w,b,alpha,N)
  assert(np.sum(grad_dw)==2.613689585)
  return True
grad_x=np.array([-2.07864835,  3.31604252, -0.79104357, -3.87045546, -1.14783286,
       -2.81434437, -0.86771071, -0.04073287,  0.84827878,  1.99451725,
        3.67152472,  0.01451875,  2.01062888,  0.07373904, -5.54586092])
grad_y=0
grad_w,grad_b=initialize_weights(grad_x)
alpha=0.0001
N=len(X_train)
grader_dw(grad_x,grad_y,grad_w,grad_b,alpha,N)
```

Out[17]:

True

Compute gradient w.r.to 'b'

$ db^{(t)} = y_n- σ((w^{(t)})^{T} x_n+b^{t})$

In [18]:

```
 def gradient_db(x,y,w,b):
     '''In this function, we will compute gradient w.r.to b '''
     db = y - sigmoid(np.dot(w.T,x)+b)
     return db
```

Grader function - 5

In [19]:

```
def grader_db(x,y,w,b):
  grad_db=gradient_db(x,y,w,b)
  assert(grad_db==-0.5)
  return True
grad_x=np.array([-2.07864835,  3.31604252, -0.79104357, -3.87045546, -1.14783286,
       -2.81434437, -0.86771071, -0.04073287,  0.84827878,  1.99451725,
        3.67152472,  0.01451875,  2.01062888,  0.07373904, -5.54586092])
grad_y=0
grad_w,grad_b=initialize_weights(grad_x)
alpha=0.0001
N=len(X_train)
grader_db(grad_x,grad_y,grad_w,grad_b)
```

Out[19]:

```
True
```

Implementing logistic regression

In [20]:

```python
def train(X_train,y_train,X_test,y_test,epochs,alpha,eta0):
    ''' In this function, we will implement logistic regression'''
    #Here eta0 is learning rate
    w,b = initialize_weights(X_train[0])
    N = len(X_train)
    # to store train_loss and test_loss for all epochs
    train_loss = []
    test_loss  = []

    # to store previos train_loss
    previous = 0

    # for every epoch
    for epoch in range(epochs):

        #  for every data point(X_train,y_train)
        for each in range(N):
           #compute gradient w.r.to w (call the gradient_dw() function)
           grad_dw = gradient_dw(X_train[each],y_train[each],w,b,alpha,N)
           #compute gradient w.r.to b (call the gradient_db() function)
           grad_db = gradient_db(X_train[each],y_train[each],w,b)
           #update w, b
           w = w + eta0*grad_dw
           b = b + eta0*grad_db

        # predict the output of x_train[for all data points in X_train] using w,b
        y_train_predicted = [sigmoid(np.dot(w.T,x)+b)  for x in X_train]
        #compute the loss between predicted and actual values (call the loss function)
        tr_loss = logloss(y_train,y_train_predicted)
        # store all the train loss values in a list
        train_loss.append(tr_loss)

        # predict the output of x_test[for all data points in X_test] using w,b
        y_test_predicted = [sigmoid(np.dot(w.T,x)+b)  for x in X_test]
        #compute the loss between predicted and actual values (call the loss function)
        te_loss = logloss(y_test,y_test_predicted)
        # store all the test loss values in a list
        test_loss.append(te_loss)

        # you can also compare previous loss and current loss, if loss is not updating then stop the pr
ocess and return w,b
        current = tr_loss
        if (abs(previous-current) < 0.00001):
          return w,b,train_loss,test_loss,epoch
        previous = current
    return w,b,train_loss,test_loss,epoch
```

In [21]:

```python
alpha  = 0.0001  # here alpha is the term used for regularizer.
eta0   = 0.0001  # here eta0 is the term used for learning rate.
N      =  len(X_train)
epochs = 50
# finding best w,b of logistic regression using sgd
w,b,train_loss,test_loss,last_epoch   = train(X_train,y_train,X_test,y_test,epochs,alpha,eta0)
```

In [22]:

```python
# train_loss for all epochs upto end of training
total_epochs = last_epoch+1
for i in range(total_epochs):
  print("Epoch : "+str(i)+" Train_loss "+str(train_loss[i]))
```

```
Epoch : 0 Train_loss 0.1754574844285461
```

```
Epoch : 1 Train_loss 0.16867157050333045
Epoch : 2 Train_loss 0.1663916799246292
Epoch : 3 Train_loss 0.16536827537403162
Epoch : 4 Train_loss 0.16485707459547086
Epoch : 5 Train_loss 0.16458820012928274
Epoch : 6 Train_loss 0.16444271323364384
Epoch : 7 Train_loss 0.16436263615826988
Epoch : 8 Train_loss 0.1643180694666775
Epoch : 9 Train_loss 0.16429307374132515
Epoch : 10 Train_loss 0.1642789743093407
Epoch : 11 Train_loss 0.16427098545835503
```

Goal of assignment

Compare your implementation and SGDClassifier's the weights and intercept, make sure they are as close as possible i.e difference should be in terms of 10^-3

In [23]:

```
# these are the results we got after we implemented sgd and found the optimal weights and intercept
w-clf.coef_, b-clf.intercept_
```

Out[23]:

```
(array([[-0.00268543,  0.00639463,  0.00155456, -0.00331533, -0.00787491,
          0.0071592 ,  0.00715886,  0.00317169,  0.01037639, -0.00928916,
         -0.00028492, -0.00318369,  0.00024727,  0.00040174, -0.00015886]]),
 array([-0.01543912]))
```

Plot epoch number vs train , test loss

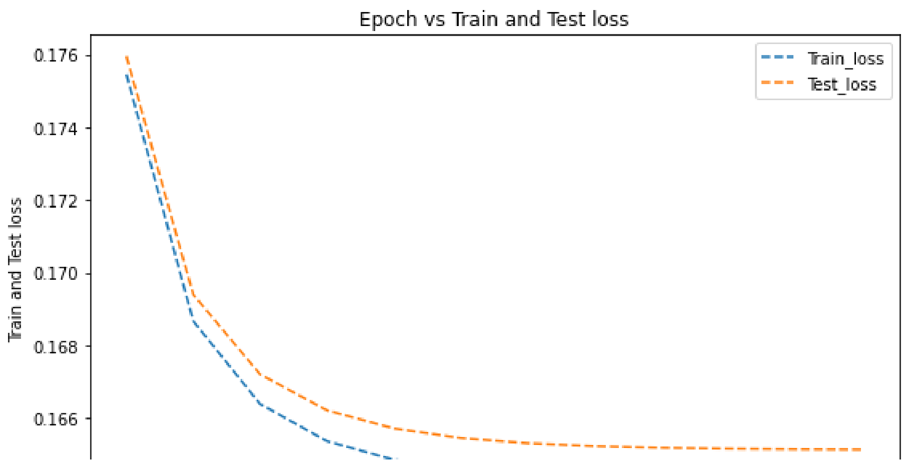
- epoch number on X-axis
- loss on Y-axis

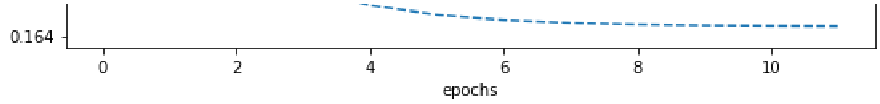
In [24]:

```
import matplotlib.pyplot as plt
```

In [25]:

```
# plot for each epoch vs Train and Test loss
no_of_epochs = range(total_epochs)
plt.figure(figsize=(8,5))
plt.plot(no_of_epochs,train_loss,label="Train_loss",linestyle='dashed',)
plt.plot(no_of_epochs,test_loss,label="Test_loss",linestyle='dashed')
plt.title("Epoch vs Train and Test loss")
plt.xlabel("epochs")
plt.ylabel("Train and Test loss")
plt.legend()
plt.tight_layout()
plt.show()
```





In [26]:

```
def pred(w,b, X):
    N = len(X)
    predict = []
    for i in range(N):
        z=np.dot(w,X[i])+b
        if sigmoid(z) >= 0.5: # sigmoid(w,x,b) returns 1/(1+exp(-(dot(x,w)+b)))
            predict.append(1)
        else:
            predict.append(0)
    return np.array(predict)
print(1-np.sum(y_train - pred(w,b,X_train))/len(X_train))
print(1-np.sum(y_test  - pred(w,b,X_test))/len(X_test))
```

```
0.9542933333333333
0.95192
```

In [26]: