ECE C147/C247, Winter 2023 HW#2

UCLA ECE

1) Noisy Linear Regression

$D = \{(x^{(1)}, y^{(1)}), \dots, y^{(N)}, y^{(N)}$

size of $D = N \times (d+1)$, +1 for labels

$x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$ where i is the ith house.

=> small weights

Recall: $$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (x^{(i)})^T\theta)^2 = \frac{1}{N}(Y - X\theta)^T(Y - X\theta)$$

Add noise:

$$\tilde{\mathcal{L}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (x^{(i)} + \delta^{(i)})^T\theta)^2$$

$\delta^{(i)}$, i.i.d and $\sim N(0, \sigma^2 I) \in \mathbb{R}^d$

a) $$E[\tilde{\mathcal{L}}(\theta)] = E\left[\frac{1}{N}\sum_{i=1}^{n}(\underbrace{y^{(i)} - (x^{(i)})^T\theta}_{a \in \mathbb{R}} - \delta^{(i)T}\theta)^2\right]$$

$$E[\tilde{\mathcal{L}}(\theta)] = E\left[\frac{1}{N}\sum_{i=1}^{n}(a^2 - 2a\,\delta^{(i)T}\theta + \theta^T\delta^{(i)}\delta^{(i)T}\theta)\right.$$

$$E[\tilde{\mathcal{L}}(\theta)] = \frac{1}{N}\sum_{i=1}^{n} E\left[(y^{(i)} - (x^{(i)})^T\theta)^2 - 2(y^{(i)} - (x^{(i)})^T\theta)\delta^{(i)T}\theta + \theta^T\delta^{(i)}\delta^{(i)T}\theta)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n}\Big[(y^{(i)}-(x^{(i)})^T\theta)^2 - 2(y^{(i)}-(x^{(i)})^T\theta)\,\underbrace{\underbrace{E[\delta^{(i)}]}_{\vec{0}}{}^T\theta}_{0} + \theta^T\underbrace{E[\delta^{(i)}\delta^{(i)T}]}_{\sigma^2 I_{d\times d}}\theta\Big]$$

$$= J(\theta) - 0 + \frac{1}{N}\sum_{i=1}^{n}\sigma^2\theta^T\theta$$

$$= J(\theta) + \sigma^2\theta^T\theta$$

$$\boxed{= J(\theta) + \sigma^2\|\theta\|_2^2} \rightarrow \text{Ridge regression} \rightarrow \text{smaller weights}$$

b) Additional term will have the same effect of Ridge regression with $\lambda = \sigma^2 > 0$, Ridge regression is mainly used to shrink the coefficients and avoid overfitting.

c) $\sigma = 0$, is regular linear regression with $\tilde{J}(\theta) = J(\theta)$ Therefore, there will be no regularization.

d) $\sigma = \infty$, all coefficients should be zero to minimize the cost function

knn_nosol

January 28, 2023

0.1 This is the k-nearest neighbors workbook for ECE C147/C247 Assignment #2

Please follow the notebook linearly to implement k-nearest neighbors.

Please print out the workbook entirely when completed.

The goal of this workbook is to give you experience with the data, training and evaluating a simple classifier, k-fold cross validation, and as a Python refresher.

0.2 Import the appropriate libraries

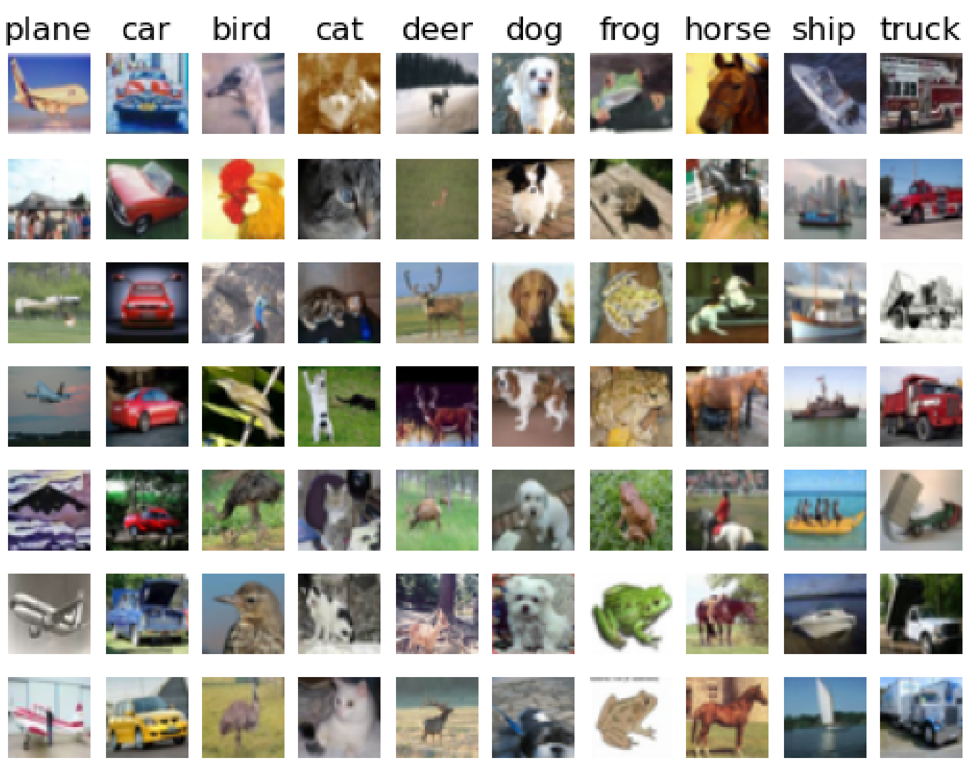
[1]:
```
import numpy as np # for doing most of our calculations
import matplotlib.pyplot as plt# for plotting
from utils.data_utils import load_CIFAR10 # function to load the CIFAR-10␣
↪dataset.
import os
# Load matplotlib images inline
%matplotlib inline

# These are important for reloading any code you write in external .py files.
# see http://stackoverflow.com/questions/1907993/
↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2
```

[2]:
```
# Set the path to the CIFAR-10 data
cifar10_dir = 'cifar-10-batches-py' # You need to update this line
X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

# As a sanity check, we print out the size of the training and test data.
print('Training data shape: ', X_train.shape)
print('Training labels shape: ', y_train.shape)
print('Test data shape: ', X_test.shape)
print('Test labels shape: ', y_test.shape)
```

```
Training data shape:  (50000, 32, 32, 3)
Training labels shape:  (50000,)
Test data shape:  (10000, 32, 32, 3)
Test labels shape:  (10000,)
```

[3]:

```
# Visualize some examples from the dataset.
# We show a few examples of training images from each class.
classes = ['plane', 'car', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse',␣
↪'ship', 'truck']
num_classes = len(classes)
samples_per_class = 7
for y, cls in enumerate(classes):
    idxs = np.flatnonzero(y_train == y)
    idxs = np.random.choice(idxs, samples_per_class, replace=False)
    for i, idx in enumerate(idxs):
        plt_idx = i * num_classes + y + 1
        plt.subplot(samples_per_class, num_classes, plt_idx)
        plt.imshow(X_train[idx].astype('uint8'))
        plt.axis('off')
        if i == 0:
            plt.title(cls)
plt.show()
```



[4]:

```
# Subsample the data for more efficient code execution in this exercise
num_training = 5000
mask = list(range(num_training))
X_train = X_train[mask]
```

```
y_train = y_train[mask]

num_test = 500
mask = list(range(num_test))
X_test = X_test[mask]
y_test = y_test[mask]

# Reshape the image data into rows
X_train = np.reshape(X_train, (X_train.shape[0], -1))
X_test = np.reshape(X_test, (X_test.shape[0], -1))
print(X_train.shape, X_test.shape)
```

```
(5000, 3072) (500, 3072)
```

1 K-nearest neighbors

In the following cells, you will build a KNN classifier and choose hyperparameters via k-fold cross-validation.

[5]:
```
# Import the KNN class

from nndl import KNN
```

[6]:
```
# Declare an instance of the knn class.
knn = KNN()

# Train the classifier.
#   We have implemented the training of the KNN classifier.
#   Look at the train function in the KNN class to see what this does.
knn.train(X=X_train, y=y_train)
```

1.1 Questions

(1) Describe what is going on in the function knn.train().

(2) What are the pros and cons of this training step?

1.2 Answers

(1) Saves all data which includes X_train and y_train. Therefore, knn object can use that data to test model performance.

(2) Pros: Easy and fast, Cons: program holds size(X_train) + size(y_train) memory which can be exhaustive if the dimensions are large.

1.3 KNN prediction

In the following sections, you will implement the functions to calculate the distances of test points to training points, and from this information, predict the class of the KNN.

```
[7]: # Implement the function compute_distances() in the KNN class.
# Do not worry about the input 'norm' for now; use the default definition of␣
↪the norm
# in the code, which is the 2-norm.
# You should only have to fill out the clearly marked sections.

import time
time_start =time.time()

dists_L2 = knn.compute_distances(X=X_test)

print('Time to run code: {}'.format(time.time()-time_start))
print('Frobenius norm of L2 distances: {}'.format(np.linalg.norm(dists_L2,␣
↪'fro')))
```

```
Time to run code: 21.368305921554565
Frobenius norm of L2 distances: 7906696.077040902
```

Really slow code Note: This probably took a while. This is because we use two for loops. We could increase the speed via vectorization, removing the for loops.

If you implemented this correctly, evaluating np.linalg.norm(dists_L2, 'fro') should return: ~7906696

1.3.1 KNN vectorization

The above code took far too long to run. If we wanted to optimize hyperparameters, it would be time-expensive. Thus, we will speed up the code by vectorizing it, removing the for loops.

```
[8]: # Implement the function compute_L2_distances_vectorized() in the KNN class.
# In this function, you ought to achieve the same L2 distance but WITHOUT any␣
↪for loops.
# Note, this is SPECIFIC for the L2 norm.

time_start =time.time()
dists_L2_vectorized = knn.compute_L2_distances_vectorized(X=X_test)
print('Time to run code: {}'.format(time.time()-time_start))
print('Difference in L2 distances between your KNN implementations (should be␣
↪0): {}'.format(np.linalg.norm(dists_L2 - dists_L2_vectorized, 'fro')))
```

```
Time to run code: 0.14246535301208496
Difference in L2 distances between your KNN implementations (should be 0): 0.0
```

Speedup Depending on your computer speed, you should see a 10-100x speed up from vectorization. On our computer, the vectorized form took 0.36 seconds while the naive implementation took 38.3 seconds.

1.3.2 Implementing the prediction

Now that we have functions to calculate the distances from a test point to given training points, we now implement the function that will predict the test point labels.

[9]:
```
# Implement the function predict_labels in the KNN class.
# Calculate the training error (num_incorrect / total_samples)
#   from running knn.predict_labels with k=1

error = 1

# ================================================================ #
# YOUR CODE HERE:
#   Calculate the error rate by calling predict_labels on the test
#   data with k = 1.  Store the error rate in the variable error.
# ================================================================ #
y_pred = knn.predict_labels(dists_L2_vectorized,k = 1)
error = np.sum(y_pred != y_test)/y_pred.shape[0]
# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #

print(error)
```

```
0.726
```

If you implemented this correctly, the error should be: 0.726.

This means that the k-nearest neighbors classifier is right 27.4% of the time, which is not great, considering that chance levels are 10%.

2 Optimizing KNN hyperparameters

In this section, we'll take the KNN classifier that you have constructed and perform cross-validation to choose a best value of k, as well as a best choice of norm.

2.0.1 Create training and validation folds

First, we will create the training and validation folds for use in k-fold cross validation.

[10]:
```
# Create the dataset folds for cross-valdiation.
num_folds = 5

X_train_folds = []
y_train_folds =  []

# ================================================================ #
# YOUR CODE HERE:
#   Split the training data into num_folds (i.e., 5) folds.
#   X_train_folds is a list, where X_train_folds[i] contains the
```

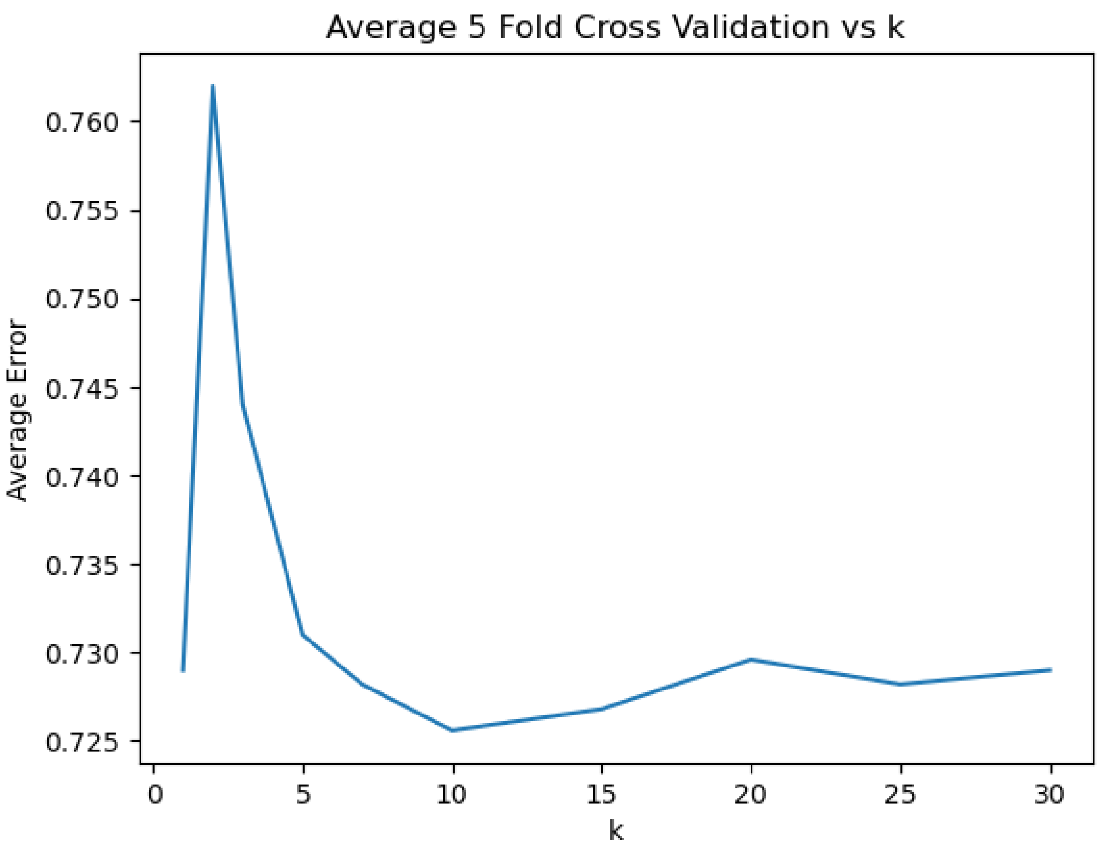
```
#      data points in fold i.
#   y_train_folds is also a list, where y_train_folds[i] contains
#      the corresponding labels for the data in X_train_folds[i]

# ================================================================ #
#Shuffle ?
shuffle = True
if shuffle:
    shuffle_idx = np.random.permutation(X_train.shape[0])

    X_train_folds = np.array_split(X_train[shuffle_idx], num_folds)
    y_train_folds = np.array_split(y_train[shuffle_idx], num_folds)
else:

    X_train_folds = np.array_split(X_train, num_folds)
    y_train_folds = np.array_split(y_train, num_folds)
# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```

2.0.2 Optimizing the number of nearest neighbors hyperparameter.

In this section, we select different numbers of nearest neighbors and assess which one has the lowest k-fold cross validation error.

[11]:
```
time_start =time.time()

ks = [1, 2, 3, 5, 7, 10, 15, 20, 25, 30]

# ================================================================ #
# YOUR CODE HERE:
#   Calculate the cross-validation error for each k in ks, testing
#   the trained model on each of the 5 folds.  Average these errors
#   together and make a plot of k vs. cross-validation error. Since
#   we are assuming L2 distance here, please use the vectorized code!
#   Otherwise, you might be waiting a long time.
# ================================================================ #
error_k_cv = []
for k in ks:
    sum_error = 0
    for i in range(0,num_folds):
        knn = KNN()

        #Assign folds to training and validation datasets.
        X_validation_cv = X_train_folds[i]
        y_validation_cv = y_train_folds[i]
        X_train_cv = np.concatenate(X_train_folds[:i] + X_train_folds[i+1:])
```

```
        y_train_cv = np.concatenate(y_train_folds[:i] + y_train_folds[i+1:])

        #knn
        knn.train(X_train_cv,y_train_cv)
        dists = knn.compute_L2_distances_vectorized(X_validation_cv)
        y_pred = knn.predict_labels(dists, k = k)

        #error
        sum_error += (np.sum(y_pred != y_validation_cv)/y_pred.shape[0])

    error_k_cv.append(sum_error/num_folds) # take average of the errors found␣
 ↪in each cross validation step

plt.plot(ks,error_k_cv)
plt.title('Average 5 Fold Cross Validation vs k')
plt.ylabel('Average Error')
plt.xlabel('k')
plt.show()

best_idx = np.argmin(error_k_cv)
print("Best error is ", error_k_cv[best_idx] ," with k = ", ks[best_idx] )
# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #

print('Computation time: %.2f'%(time.time()-time_start))
```




```
Best error is  0.7256  with k =  10
Computation time: 22.31
```

2.1 Questions:

(1) What value of k is best amongst the tested k's?

(2) What is the cross-validation error for this value of k?

2.2 Answers:

(1) k = 10

(2) cv_error = 0.7256, if the dataset is shuffled.

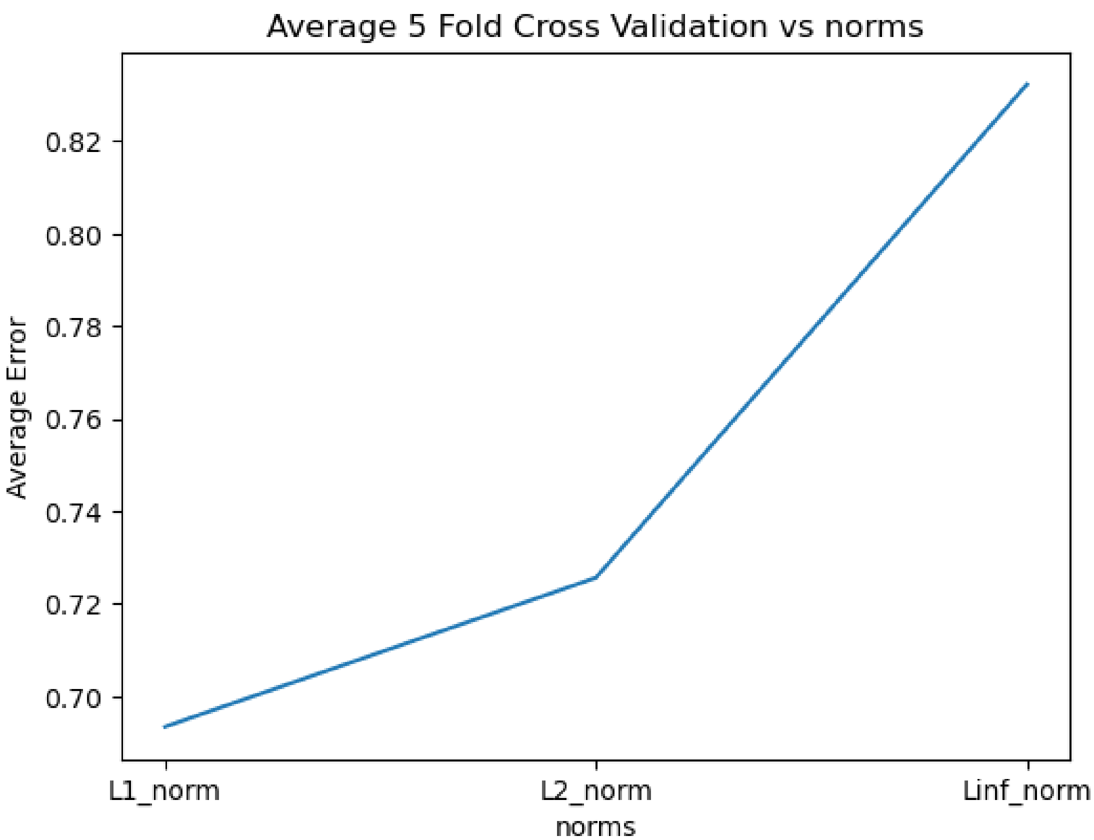
2.2.1 Optimizing the norm

Next, we test three different norms (the 1, 2, and infinity norms) and see which distance metric results in the best cross-validation performance.

[12]:
```
time_start =time.time()
```

```python
L1_norm = lambda x: np.linalg.norm(x, ord=1)
L2_norm = lambda x: np.linalg.norm(x, ord=2)
Linf_norm = lambda x: np.linalg.norm(x, ord= np.inf)
norms = [L1_norm, L2_norm, Linf_norm]

# ================================================================ #
# YOUR CODE HERE:
#   Calculate the cross-validation error for each norm in norms, testing
#   the trained model on each of the 5 folds.  Average these errors
#   together and make a plot of the norm used vs the cross-validation error
#   Use the best cross-validation k from the previous part.
#
#   Feel free to use the compute_distances function.  We're testing just
#   three norms, but be advised that this could still take some time.
#   You're welcome to write a vectorized form of the L1- and Linf- norms
#   to speed this up, but it is not necessary.
# ================================================================ #
norms_list = ["L1_norm", "L2_norm", "Linf_norm"]
error_norm_cv = []

for norm in norms:
    sum_error = 0
    for i in range(0,num_folds):
        knn = KNN()

        #Assign folds to training and validation datasets.
        X_validation_cv = X_train_folds[i]
        y_validation_cv = y_train_folds[i]
        X_train_cv = np.concatenate(X_train_folds[:i] + X_train_folds[i+1:])
        y_train_cv = np.concatenate(y_train_folds[:i] + y_train_folds[i+1:])

        #knn
        knn.train(X_train_cv,y_train_cv)
        dists = knn.compute_distances(X_validation_cv, norm = norm)
        y_pred = knn.predict_labels(dists, k = ks[best_idx])

        #error
        sum_error += (np.sum(y_pred != y_validation_cv)/y_pred.shape[0])
    error_norm_cv.append(sum_error/num_folds) # take average of the errors␣
 ↪found in each cross validation step

plt.plot(norms_list, error_norm_cv)
plt.title('Average 5 Fold Cross Validation vs norms')
plt.ylabel('Average Error')
plt.xlabel('norms')
plt.show()
```

```
print("Errors for L1, L2, Linf: ", error_norm_cv)
best_idx = np.argmin(error_norm_cv)
print("Best error is ", error_norm_cv[best_idx] ," with norm = ",␣
 ↪norms_list[best_idx] )

# =============================================================== #
# END YOUR CODE HERE
# =============================================================== #
print('Computation time: %.2f'%(time.time()-time_start))
```




```
Errors for L1, L2, Linf:  [0.6934, 0.7256, 0.8321999999999999]
Best error is  0.6934  with norm =  L1_norm
Computation time: 418.51
```

2.3 Questions:

(1) What norm has the best cross-validation error?

(2) What is the cross-validation error for your given norm and k?

2.4 Answers:

(1) L1-norm has the best cv error.

(2) Best cross validation error is 0.6934 with L1_norm and k = 10.

3 Evaluating the model on the testing dataset.

Now, given the optimal k and norm you found in earlier parts, evaluate the testing error of the k-nearest neighbors model.

[14]:
```
error = 1

# ================================================================ #
# YOUR CODE HERE:
#   Evaluate the testing error of the k-nearest neighbors classifier
#   for your optimal hyperparameters found by 5-fold cross-validation.
# ================================================================ #

#Optimal k and norm, k = , norm = l1
L1_norm = lambda x: np.linalg.norm(x, ord=1)
k = 10

knn = KNN()
knn.train(X_train,y_train)
dists = knn.compute_distances(X = X_test, norm = L1_norm)
y_pred = knn.predict_labels(dists, k = k)
error = (np.sum(y_pred != y_test)/y_pred.shape[0])

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #

print('Error rate achieved: {}'.format(error))
```

```
Error rate achieved: 0.722
```

3.1 Question:

How much did your error improve by cross-validation over naively choosing $k = 1$ and using the L2-norm?

3.2 Answer:

My error improved from 0.726 to 0.722 by selecting hyperparameters with cross-validation.

[ ]:
```
```

```python
import numpy as np
import pdb


class KNN(object):

  def __init__(self):
    pass

  def train(self, X, y):
    """
    Inputs:
    - X is a numpy array of size (num_examples, D)
    - y is a numpy array of size (num_examples, )
    """
    self.X_train = X
    self.y_train = y

  def compute_distances(self, X, norm=None):
    """
    Compute the distance between each test point in X and each training point
    in self.X_train.

    Inputs:
    - X: A numpy array of shape (num_test, D) containing test data.
    - norm: the function with which the norm is taken.

    Returns:
    - dists: A numpy array of shape (num_test, num_train) where dists[i, j]
      is the Euclidean distance between the ith test point and the jth training
      point.
    """
    if norm is None:
      norm = lambda x: np.sqrt(np.sum(x**2))
      #norm = 2

    num_test = X.shape[0]
    num_train = self.X_train.shape[0]
    dists = np.zeros((num_test, num_train))

    for i in np.arange(num_test):

      for j in np.arange(num_train):
        # ================================================================ #
        # YOUR CODE HERE:
        #   Compute the distance between the ith test point and the jth
        #   training point using norm(), and store the result in dists[i, j].
        # ================================================================ #

        dists[i,j] = norm(self.X_train[j,:] - X[i,:])

        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #

    return dists

  def compute_L2_distances_vectorized(self, X):
    """
    Compute the distance between each test point in X and each training point
    in self.X_train WITHOUT using any for loops.

    Inputs:
    - X: A numpy array of shape (num_test, D) containing test data.

    Returns:
    - dists: A numpy array of shape (num_test, num_train) where dists[i, j]
```

```
      is the Euclidean distance between the ith test point and the jth training
      point.
    """
    num_test = X.shape[0]
    num_train = self.X_train.shape[0]
    dists = np.zeros((num_test, num_train))

    # ================================================================ #
    # YOUR CODE HERE:
    #   Compute the L2 distance between the ith test point and the jth
    #   training point and store the result in dists[i, j].  You may
    #   NOT use a for loop (or list comprehension).  You may only use
    #   numpy operations.
    #
    #   HINT: use broadcasting.  If you have a shape (N,1) array and
    #   a shape (M,) array, adding them together produces a shape (N, M)
    #   array.
    # ================================================================ #

    #Alternative solution which is slower
    #test_norm = np.diag(np.dot(X,X.T)).reshape(num_test,1) # at the diagonals we obtain norm of each sample
    #train_norm = np.diag(np.dot(self.X_train,self.X_train.T)).reshape((1,num_train))

    # Below is faster
    test_norm = np.sum(X**2, axis = 1).reshape((num_test,1)) # add columns together to obtain norm for each sample
    train_norm = np.sum(self.X_train**2, axis = 1).reshape((1,num_train))
    cross_norm = np.dot(X,self.X_train.T)
    dists = np.sqrt(dists + test_norm -  2 * cross_norm + train_norm)

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return dists


  def predict_labels(self, dists, k=1):
    """
    Given a matrix of distances between test points and training points,
    predict a label for each test point.

    Inputs:
    - dists: A numpy array of shape (num_test, num_train) where dists[i, j]
      gives the distance betwen the ith test point and the jth training point.

    Returns:
    - y: A numpy array of shape (num_test,) containing predicted labels for the
      test data, where y[i] is the predicted label for the test point X[i].
    """
    num_test = dists.shape[0]
    y_pred = np.zeros(num_test)
    for i in np.arange(num_test):
      # A list of length k storing the labels of the k nearest neighbors to
      # the ith test point.
      closest_y = []
      # ================================================================ #
      # YOUR CODE HERE:
      #   Use the distances to calculate and then store the labels of
      #   the k-nearest neighbors to the ith test point.  The function
      #   numpy.argsort may be useful.
      #
      #   After doing this, find the most common label of the k-nearest
      #   neighbors.  Store the predicted label of the ith training example
      #   as y_pred[i].  Break ties by choosing the smaller label.
      # ================================================================ #
```

```
            idx = np.argsort(dists[i])
            closest_y = self.y_train[idx[:k]]
            unique, counts = np.unique(closest_y, return_counts=True)
            y_pred[i] = unique[np.argmax(counts)]

            # ================================================================ #
            # END YOUR CODE HERE
            # ================================================================ #

        return y_pred
```

3)

Data: $(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})$, where $x^{(j)} \in \mathbb{R}^n$

$y^{(j)} \in \{1, \dots, c\}$

$j = 1, \dots, m$

m: sample size

n: # of fatures

$\theta = \{w_i, b_i\}_{i=1,\dots,c}$

$$\tilde{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}, \quad \tilde{w}_i = \begin{bmatrix} w_i \\ b_i \end{bmatrix}$$

$$\Downarrow$$

$$a_i(x) = \tilde{w}_i^T \tilde{x}$$

$$\text{softmax}_i(x) = \frac{e^{w_i^T x + b_i}}{\sum_{k=1}^{c} e^{w_k^T x + b_k}}$$

$$p(x^{(1)}, \dots, x^{(m)}, y^{(1)}, \dots, y^{(m)} \mid \theta) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)} \mid \theta)$$

$$= \prod_{i=1}^{m} p(x^{(i)} \mid \theta) \underbrace{p(y^{(i)} \mid x^{(i)}, \theta)}_{\text{softmax}_j(x^{(i)})}$$

$$\arg\max_{\theta} \prod_{i=1}^{m} p(x^{(i)} \mid \theta)\, p(y^{(i)} \mid x^{(i)}, \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}, \theta)$$

Take log:

$$\arg\max_{\theta} \sum_{i=1}^{m} \log \left[\frac{e^{a_{y^{(i)}}(x^{(i)})}}{\sum_{j=1} e^{a_j(x^{(i)})}} \right]$$

$\arg\max_{\theta} f(\theta) = \arg\min_{\theta} -f(\theta)$

$$= \arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \left[\log \left(\sum_{j=1}^{c} e^{a_j(x^{(i)})} \right) - a_{y^{(i)}}(x^{(i)}) \right]$$

$$\nabla_{\tilde{w}_i}\left(\log \sum_{j=1}^{c} e^{a_j(x)}\right)$$

$$= \nabla_{\tilde{w}_i}\left(\log\left[e^{w_1^T\tilde{x}} + e^{w_2^T\tilde{x}} + \dots + e^{w_c^T\tilde{x}}\right]\right)$$

$$= \frac{1}{\sum_{j=1}^{c} e^{a_j(x)}} e^{\tilde{w}_i^T\tilde{x}}\,\tilde{x} \quad \text{, by chain rule}$$

$$= \frac{e^{\tilde{w}_i^T\tilde{x}}}{\sum_{j=1}^{c} e^{a_j(x)}}\tilde{x} = \frac{e^{a_i(x)}}{\sum_{j=1}^{c} e^{a_j(x)}}\tilde{x}$$

$$\nabla_{w_i} a_{y(k)}(x) \Rightarrow \begin{cases} \nabla_{\tilde{w}_i} a_{y(i)}(x) & \text{if } i = y(k) \\ 0 & \text{if } i \neq y(k) \end{cases}$$

If $i = y(k)$

$$\nabla_{\tilde{w}_i} \tilde{w}_i^T\tilde{x} = \tilde{x}$$

$$\nabla_{\tilde{w}_i} \mathcal{L}(\tilde{w}_i) = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{e^{a_i(x^{(j)})}}{\sum_{k=1}^{c} e^{a_k(x^{(j)})}} - \delta_{y(j),i}\right]\tilde{x}^{(j)}$$

$$\boxed{\begin{aligned} \nabla_{w_i} \mathcal{L}(w_i, b_i) &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{e^{a_i(x^{(j)})}}{\sum_{k=1}^{c} e^{a_k(x^{(j)})}} - \delta_{y^{(j)},i}\right] x^{(j)} \\ \nabla_{b_i} \mathcal{L}(w_i, b_i) &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{e^{a_i(x^{(j)})}}{\sum_{k=1}^{c} e^{a_k(x^{(j)})}} - \delta_{y^{(j)},i}\right] \end{aligned}}$$

4)

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(k)}, y^{(k)})\}$$

$$x^{(i)} \in \mathbb{R}^d, \quad y^{(i)} \in \{-1, 1\}$$

$$\boxed{\begin{aligned} \tilde{w} &= \begin{bmatrix} w \\ 1 \end{bmatrix} \\ \tilde{x}^{(i)} &= \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix} \end{aligned}}$$

$$\mathcal{L}(w, b) = \frac{1}{k} \sum_{i=1}^{m} \underbrace{\max(0, 1 - y^{(i)}(w^T x^{(i)} + b))}_{hinge_{y^{(i)}}(x^{(i)})}$$

$$\frac{\partial\, hinge_{y^{(i)}}(x^{(i)})}{\partial \tilde{w}} = \begin{cases} 0 & \text{if } y_i(\tilde{w}^T \tilde{x}^{(i)}) \geq 1 \\ -y_i \tilde{x}^{(i)} & \text{if } y_i(\tilde{w}^T \tilde{x}^{(i)}) < 1 \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{w}} = \frac{1}{k} \sum_{i=1}^{m} \frac{\partial\, hinge_{y^{(i)}}(x^{(i)})}{\partial \tilde{w}} = \frac{1}{k} \sum_{i=1}^{m} \mathbb{1}_{y_i(\tilde{w}^T \tilde{x}^{(i)})} (-y_i \tilde{x})$$

$$\boxed{\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} &= \frac{1}{k} \sum_{i=1}^{m} \mathbb{1}_{y_i(w^T x^{(i)} + b)} (-y_i x^{(i)}) \\ \frac{\partial \mathcal{L}}{\partial b} &= \frac{1}{k} \sum_{i=1}^{m} \mathbb{1}_{y_i(w^T x^{(i)} + b)} (-y_i) \end{aligned}}$$

softmax_nosol

January 28, 2023

0.1 This is the softmax workbook for ECE C147/C247 Assignment #2

Please follow the notebook linearly to implement a softmax classifier.

Please print out the workbook entirely when completed.

The goal of this workbook is to give you experience with training a softmax classifier.

[1]:
```
import random
import numpy as np
from utils.data_utils import load_CIFAR10
import matplotlib.pyplot as plt

%matplotlib inline
%load_ext autoreload
%autoreload 2
```

[2]:
```
def get_CIFAR10_data(num_training=49000, num_validation=1000, num_test=1000,␣
 ↪num_dev=500):
    """
    Load the CIFAR-10 dataset from disk and perform preprocessing to prepare
    it for the linear classifier. These are the same steps as we used for the
    SVM, but condensed to a single function.
    """
    # Load the raw CIFAR-10 data
    cifar10_dir = 'cifar-10-batches-py' # You need to update this line
    X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

    # subsample the data
    mask = list(range(num_training, num_training + num_validation))
    X_val = X_train[mask]
    y_val = y_train[mask]
    mask = list(range(num_training))
    X_train = X_train[mask]
    y_train = y_train[mask]
    mask = list(range(num_test))
    X_test = X_test[mask]
    y_test = y_test[mask]
    mask = np.random.choice(num_training, num_dev, replace=False)
```

```
    X_dev = X_train[mask]
    y_dev = y_train[mask]

    # Preprocessing: reshape the image data into rows
    X_train = np.reshape(X_train, (X_train.shape[0], -1))
    X_val = np.reshape(X_val, (X_val.shape[0], -1))
    X_test = np.reshape(X_test, (X_test.shape[0], -1))
    X_dev = np.reshape(X_dev, (X_dev.shape[0], -1))

    # Normalize the data: subtract the mean image
    mean_image = np.mean(X_train, axis = 0)
    X_train -= mean_image
    X_val -= mean_image
    X_test -= mean_image
    X_dev -= mean_image

    # add bias dimension and transform into columns
    X_train = np.hstack([X_train, np.ones((X_train.shape[0], 1))])
    X_val = np.hstack([X_val, np.ones((X_val.shape[0], 1))])
    X_test = np.hstack([X_test, np.ones((X_test.shape[0], 1))])
    X_dev = np.hstack([X_dev, np.ones((X_dev.shape[0], 1))])

    return X_train, y_train, X_val, y_val, X_test, y_test, X_dev, y_dev


# Invoke the above function to get our data.
X_train, y_train, X_val, y_val, X_test, y_test, X_dev, y_dev = 
 ↪get_CIFAR10_data()
print('Train data shape: ', X_train.shape)
print('Train labels shape: ', y_train.shape)
print('Validation data shape: ', X_val.shape)
print('Validation labels shape: ', y_val.shape)
print('Test data shape: ', X_test.shape)
print('Test labels shape: ', y_test.shape)
print('dev data shape: ', X_dev.shape)
print('dev labels shape: ', y_dev.shape)
```

```
Train data shape:  (49000, 3073)
Train labels shape:  (49000,)
Validation data shape:  (1000, 3073)
Validation labels shape:  (1000,)
Test data shape:  (1000, 3073)
Test labels shape:  (1000,)
dev data shape:  (500, 3073)
dev labels shape:  (500,)
```

0.2 Training a softmax classifier.

The following cells will take you through building a softmax classifier. You will implement its loss function, then subsequently train it with gradient descent. Finally, you will choose the learning rate of gradient descent to optimize its classification performance.

[3]:
```
from nndl import Softmax
```

[4]:
```
# Declare an instance of the Softmax class.
# Weights are initialized to a random value.
# Note, to keep people's first solutions consistent, we are going to use a
 ↪random seed.

np.random.seed(1)

num_classes = len(np.unique(y_train))
num_features = X_train.shape[1]

softmax = Softmax(dims=[num_classes, num_features])
```

Softmax loss

[5]:
```
## Implement the loss function of the softmax using a for loop over
#  the number of examples

loss = softmax.loss(X_train, y_train)
```

[6]:
```
print(loss)
```

```
2.3277607028048863
```

0.3 Question:

You'll notice the loss returned by the softmax is about 2.3 (if implemented correctly). Why does this make sense?

0.4 Answer:

Weights are randomly initilized, therefore classifier should randomly guess the class. Randomly guessing a class means the probability of choosing that class is 1/10 where 10 is the number of classes. Since our loss function is negative log-probability, -log(1/10) = log(10) = 2.3.

Softmax gradient

[7]:
```
## Calculate the gradient of the softmax loss in the Softmax class.
# For convenience, we'll write one function that computes the loss
#   and gradient together, softmax.loss_and_grad(X, y)
# You may copy and paste your loss code from softmax.loss() here, and then
#   use the appropriate intermediate values to calculate the gradient.

loss, grad = softmax.loss_and_grad(X_dev,y_dev)
```

```
# Compare your gradient to a gradient check we wrote.
# You should see relative gradient errors on the order of 1e-07 or less if you
↪implemented the gradient correctly.
softmax.grad_check_sparse(X_dev, y_dev, grad)
```

```
numerical: -0.960939 analytic: -0.960939, relative error: 1.426816e-08
numerical: 2.595124 analytic: 2.595124, relative error: 8.445868e-09
numerical: -0.504541 analytic: -0.504541, relative error: 2.965655e-08
numerical: 2.796462 analytic: 2.796462, relative error: 2.667713e-09
numerical: -1.801980 analytic: -1.801981, relative error: 3.141575e-08
numerical: 1.388577 analytic: 1.388577, relative error: 2.193987e-08
numerical: -1.940394 analytic: -1.940394, relative error: 4.875162e-09
numerical: -1.064656 analytic: -1.064656, relative error: 2.460331e-08
numerical: 0.278238 analytic: 0.278238, relative error: 1.142827e-08
numerical: -1.554694 analytic: -1.554694, relative error: 4.056617e-08
```

0.5 A vectorized version of Softmax

To speed things up, we will vectorize the loss and gradient calculations. This will be helpful for stochastic gradient descent.

[8]:
```
import time
```

[9]:
```
## Implement softmax.fast_loss_and_grad which calculates the loss and gradient
#    WITHOUT using any for loops.

# Standard loss and gradient
tic = time.time()
loss, grad = softmax.loss_and_grad(X_dev, y_dev)
toc = time.time()
print('Normal loss / grad_norm: {} / {} computed in {}s'.format(loss, np.linalg.
↪norm(grad, 'fro'), toc - tic))

tic = time.time()
loss_vectorized, grad_vectorized = softmax.fast_loss_and_grad(X_dev, y_dev)
toc = time.time()
print('Vectorized loss / grad: {} / {} computed in {}s'.format(loss_vectorized,
↪np.linalg.norm(grad_vectorized, 'fro'), toc - tic))

# The losses should match but your vectorized implementation should be much
↪faster.
print('difference in loss / grad: {} /{} '.format(loss - loss_vectorized, np.
↪linalg.norm(grad - grad_vectorized)))

# You should notice a speedup with the same output.
```

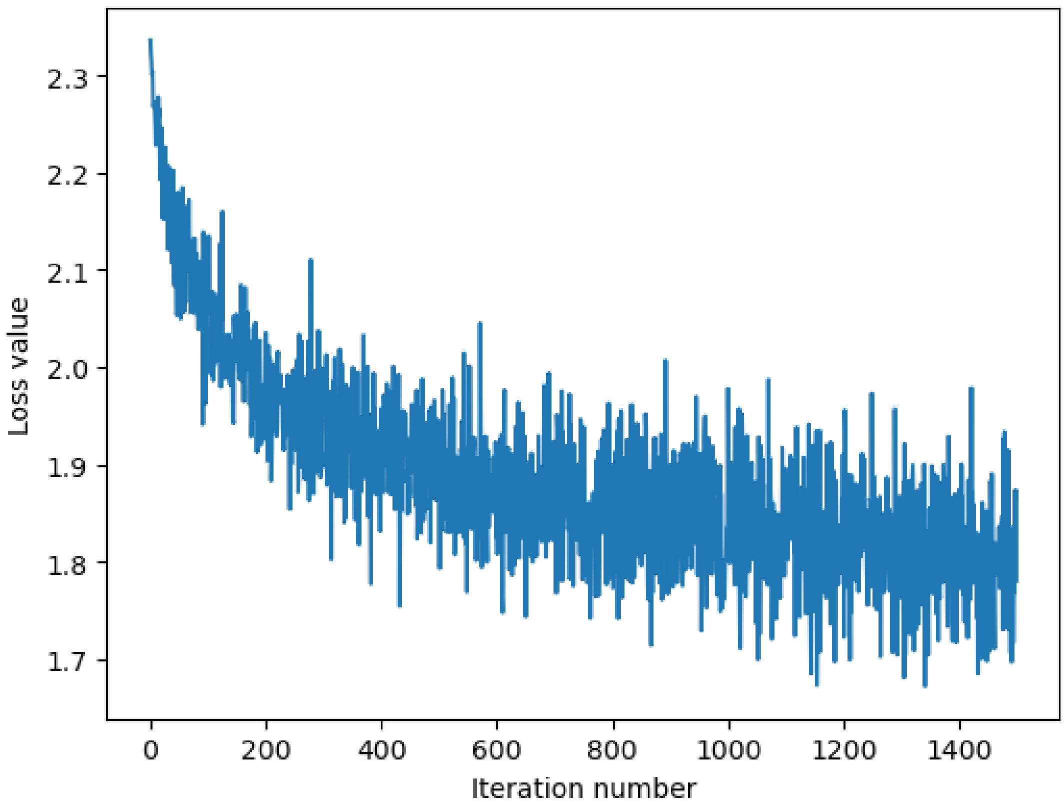
```
Normal loss / grad_norm: 2.3319379211333056 / 310.07680893055635 computed in
0.015015840530395508s
Vectorized loss / grad: 2.331937921133306 / 310.07680893055635 computed in
0.002000093460083008s
difference in loss / grad: -4.440892098500626e-16 /2.7866450218004654e-13
```

0.6 Stochastic gradient descent

We now implement stochastic gradient descent. This uses the same principles of gradient descent we discussed in class, however, it calculates the gradient by only using examples from a subset of the training set (so each gradient calculation is faster).

[10]:
```python
# Implement softmax.train() by filling in the code to extract a batch of data
# and perform the gradient step.
import time


tic = time.time()
loss_hist = softmax.train(X_train, y_train, learning_rate=1e-7,
                      num_iters=1500, verbose=True)
toc = time.time()
print('That took {}s'.format(toc - tic))

plt.plot(loss_hist)
plt.xlabel('Iteration number')
plt.ylabel('Loss value')
plt.show()
```

```
iteration 0 / 1500: loss 2.3365926606637544
iteration 100 / 1500: loss 2.0557222613850827
iteration 200 / 1500: loss 2.0357745120662813
iteration 300 / 1500: loss 1.9813348165609888
iteration 400 / 1500: loss 1.9583142443981614
iteration 500 / 1500: loss 1.8622653073541355
iteration 600 / 1500: loss 1.8532611454359382
iteration 700 / 1500: loss 1.8353062223725827
iteration 800 / 1500: loss 1.829389246882764
iteration 900 / 1500: loss 1.8992158530357484
iteration 1000 / 1500: loss 1.97835035402523
iteration 1100 / 1500: loss 1.8470797913532633
iteration 1200 / 1500: loss 1.8411450268664082
iteration 1300 / 1500: loss 1.7910402495792102
iteration 1400 / 1500: loss 1.8705803029382257
That took 4.577153205871582s
```



0.6.1 Evaluate the performance of the trained softmax classifier on the validation data.

[11]:
```
## Implement softmax.predict() and use it to compute the training and testing
↪error.

y_train_pred = softmax.predict(X_train)
print('training accuracy: {}'.format(np.mean(np.equal(y_train,y_train_pred), )))
y_val_pred = softmax.predict(X_val)
print('validation accuracy: {}'.format(np.mean(np.equal(y_val, y_val_pred)), ))
```

```
training accuracy: 0.3811428571428571
validation accuracy: 0.398
```

0.7 Optimize the softmax classifier

[12]:
```
np.finfo(float).eps
```

[12]:
```
2.220446049250313e-16
```

[13]:
```python
# ================================================================ #
# YOUR CODE HERE:
#   Train the Softmax classifier with different learning rates and
#     evaluate on the validation data.
#   Report:
#     - The best learning rate of the ones you tested.
#     - The best validation accuracy corresponding to the best validation error.
#
#   Select the SVM that achieved the best validation error and report
#     its error rate on the test set.
# ================================================================ #
learning_rates = np.geomspace(1e-9, 1e-3, num= 7)
acc = []

for learning_rate in learning_rates:
    clf = Softmax()
    clf.train(X_train,y_train,learning_rate = learning_rate, num_iters = 1500,␣
 ↪verbose = False)
    prediction = clf.predict(X_val)
    accuracy = np.sum(prediction == y_val) / y_val.shape[0]
    acc.append(accuracy)

best_idx = np.argmax(acc)
print("All accuracies : ", acc)
print("Best validation accuracy is ", acc[best_idx], " with learning rate = ",␣
 ↪learning_rates[best_idx])
print("Best validation error is ", 1 - acc[best_idx], " with learning rate = ",␣
 ↪learning_rates[best_idx])

clf = Softmax()
clf.train(X_train,y_train,learning_rate = learning_rates[best_idx], num_iters =␣
 ↪1500, verbose = False)
prediction = clf.predict(X_test)
error_test = np.sum(prediction != y_test) / y_test.shape[0]

print("Error rate on the test set is ", error_test, " with learning rate = ",␣
 ↪learning_rates[best_idx])
# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```

```
All accuracies :  [0.16, 0.304, 0.395, 0.407, 0.33, 0.262, 0.288]
Best validation accuracy is  0.407  with learning rate =  1e-06
Best validation error is  0.593  with learning rate =  1e-06
Error rate on the test set is  0.608  with learning rate =  1e-06
```

[ ]:

```python
import numpy as np


class Softmax(object):

  def __init__(self, dims=[10, 3073]):
    self.init_weights(dims=dims)

  def init_weights(self, dims):
    """
    Initializes the weight matrix of the Softmax classifier.
    Note that it has shape (C, D) where C is the number of
    classes and D is the feature size.
    """
    self.W = np.random.normal(size=dims) * 0.0001

  def loss(self, X, y):
    """
    Calculates the softmax loss.

    Inputs have dimension D, there are C classes, and we operate on minibatches
    of N examples.

    Inputs:
    - X: A numpy array of shape (N, D) containing a minibatch of data.
    - y: A numpy array of shape (N,) containing training labels; y[i] = c means
      that X[i] has label c, where 0 <= c < C.

    Returns a tuple of:
    - loss as single float
    """

    # Initialize the loss to zero.
    loss = 0.0

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the normalized softmax loss.  Store it as the variable loss.
    #   (That is, calculate the sum of the losses of all the training
    #   set margins, and then normalize the loss by the number of
    #   training examples.)
    # ================================================================ #
    num_samples = X.shape[0]

    all_scores = np.dot(X,self.W.T)
    for i in range(num_samples):
        sample_scores = all_scores[i] - np.max(all_scores[i])
        sample_class_score = sample_scores[y[i]]
        exp_sum = np.sum(np.exp(sample_scores))
        loss = loss + np.log(exp_sum) - sample_class_score

    loss = loss/num_samples

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss

  def loss_and_grad(self, X, y):
    """
    Same as self.loss(X, y), except that it also returns the gradient.

    Output: grad -- a matrix of the same dimensions as W containing
      the gradient of the loss with respect to W.
    """

```

```python
    # Initialize the loss and gradient to zero.
    loss = 0.0
    grad = np.zeros_like(self.W)

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the softmax loss and the gradient. Store the gradient
    #   as the variable grad.
    # ================================================================ #
    num_sample = X.shape[0]

    all_scores = np.dot(X,self.W.T)
    for i in range(num_sample):
        sample_scores = all_scores[i] - np.max(all_scores[i])
        sample_class_score = sample_scores[y[i]]

        exp_sum = np.sum(np.exp(sample_scores))
        loss = loss + np.log(exp_sum) - sample_class_score

        sftmax = (np.exp(sample_scores) / exp_sum).reshape(-1,1)
        grad = grad + np.dot(sftmax, X[i].reshape(1,-1))
        grad[y[i]] = grad[y[i]] - X[i]

    loss = loss/num_sample
    grad = grad/num_sample

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss, grad

  def grad_check_sparse(self, X, y, your_grad, num_checks=10, h=1e-5):
    """
    sample a few random elements and only return numerical
    in these dimensions.
    """

    for i in np.arange(num_checks):
      ix = tuple([np.random.randint(m) for m in self.W.shape])

      oldval = self.W[ix]
      self.W[ix] = oldval + h # increment by h
      fxph = self.loss(X, y)
      self.W[ix] = oldval - h # decrement by h
      fxmh = self.loss(X,y) # evaluate f(x - h)
      self.W[ix] = oldval # reset

      grad_numerical = (fxph - fxmh) / (2 * h)
      grad_analytic = your_grad[ix]
      rel_error = abs(grad_numerical - grad_analytic) / (abs(grad_numerical) + abs(grad_analytic))
      print('numerical: %f analytic: %f, relative error: %e' % (grad_numerical, grad_analytic, rel_error))

  def fast_loss_and_grad(self, X, y):
    """
    A vectorized implementation of loss_and_grad. It shares the same
    inputs and ouptuts as loss_and_grad.
    """
    loss = 0.0
    grad = np.zeros(self.W.shape) # initialize the gradient as zero

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the softmax loss and gradient WITHOUT any for loops.
    # ================================================================ #
```

```
    num_sample = X.shape[0]

    all_scores = np.dot(X,self.W.T)
    all_scores_stable = all_scores - np.max(all_scores, axis = 1, keepdims = True) # sample x class

    exp_sum = np.sum(np.exp(all_scores_stable),axis = 1, keepdims = True)
    sftmax = np.exp(all_scores_stable)/exp_sum
    sftmax = sftmax.clip(min = np.finfo(float).eps) # Added to avoid log0
    loss = np.sum(-np.log(sftmax[np.arange(num_sample),y]))

    sftmax[np.arange(num_sample),y] -= 1
    grad = np.dot(sftmax.T,X)

    loss = loss/num_sample
    grad = grad/num_sample

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss, grad

  def train(self, X, y, learning_rate=1e-3, num_iters=100,
            batch_size=200, verbose=False):
    """
    Train this linear classifier using stochastic gradient descent.

    Inputs:
    - X: A numpy array of shape (N, D) containing training data; there are N
      training samples each of dimension D.
    - y: A numpy array of shape (N,) containing training labels; y[i] = c
      means that X[i] has label 0 <= c < C for C classes.
    - learning_rate: (float) learning rate for optimization.
    - num_iters: (integer) number of steps to take when optimizing
    - batch_size: (integer) number of training examples to use at each step.
    - verbose: (boolean) If true, print progress during optimization.

    Outputs:
    A list containing the value of the loss function at each training iteration.
    """
    num_train, dim = X.shape
    num_classes = np.max(y) + 1 # assume y takes values 0...K-1 where K is number of classes

    self.init_weights(dims=[np.max(y) + 1, X.shape[1]]) # initializes the weights of self.W

    # Run stochastic gradient descent to optimize W
    loss_history = []

    for it in np.arange(num_iters):
      X_batch = None
      y_batch = None

      # ================================================================ #
      # YOUR CODE HERE:
      #   Sample batch_size elements from the training data for use in
      #     gradient descent.  After sampling,
      #     - X_batch should have shape: (batch_size, dim)
      #     - y_batch should have shape: (batch_size,)
      #   The indices should be randomly generated to reduce correlations
      #   in the dataset.  Use np.random.choice.  It's okay to sample with
      #   replacement.
      # ================================================================ #
      idx = np.random.choice(num_train, batch_size, replace = True)
      X_batch = X[idx]
```

```
      y_batch = y[idx]
      # ================================================================ #
      # END YOUR CODE HERE
      # ================================================================ #

      # evaluate loss and gradient
      loss, grad = self.fast_loss_and_grad(X_batch, y_batch)
      loss_history.append(loss)

      # ================================================================ #
      # YOUR CODE HERE:
      #   Update the parameters, self.W, with a gradient step
      # ================================================================ #
      self.W = self.W - (learning_rate * grad)

      # ================================================================ #
      # END YOUR CODE HERE
      # ================================================================ #

      if verbose and it % 100 == 0:
        print('iteration {} / {}: loss {}'.format(it, num_iters, loss))

    return loss_history

  def predict(self, X):
    """
    Inputs:
    - X: N x D array of training data. Each row is a D-dimensional point.

    Returns:
    - y_pred: Predicted labels for the data in X. y_pred is a 1-dimensional
      array of length N, and each element is an integer giving the predicted
      class.
    """
    y_pred = np.zeros(X.shape[1])
    # ================================================================ #
    # YOUR CODE HERE:
    #   Predict the labels given the training data.
    # ================================================================ #
    all_scores = np.dot(X,self.W.T)
    y_pred = np.argmax(all_scores, axis = 1)
    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return y_pred
```