kNN

November 11, 2020

1 CS145 Howework 3, Part 1: kNN

Important Note: HW3 is due on **11:59 PM PT, Nov 9 (Monday, Week 6)**. Please submit through GradeScope.

Note that, Howework #3 has two jupyter notebooks to complete (Part 1: kNN and Part 2: Neural Network).

1.1 Print Out Your Name and UID

Name: Devyan Biswas, UID: 804988161

1.2 Before You Start

You need to first create HW2 conda environment by the given `cs145hw3.yml` file, which provides the name and necessary packages for this tasks. If you have `conda` properly installed, you may create, activate or deactivate by the following commands:

```
conda env create -f cs145hw3.yml
conda activate hw3
conda deactivate
```

OR

```
conda env create --name NAMEOFYOURCHOICE -f cs145hw3.yml
conda activate NAMEOFYOURCHOICE
conda deactivate
```

To view the list of your environments, use the following command:

```
conda env list
```

More useful information about managing environments can be found here.

You may also quickly review the usage of basic Python and Numpy package, if needed in coding for matrix operations.

In this notebook, you must not delete any code cells in this notebook. If you change any code outside the blocks (such as hyperparameters) that you are allowed to edit (between `STRART/END YOUR CODE HERE`), you need to highlight these changes. You may add some additional cells to help explain your results and observations.

1.3 Download and prepare the dataset

Download the CIFAR-10 dataset (file size: ~163M). Run the following from the HW3 directory:

```
cd hw3/data/datasets
./get_datasets.sh
```

Make sure you put the dataset downloaded under hw3/data/datasets folder. After downloading the dataset, you can start your notebook from the HW3 directory. Note that the dataset is used in both jupyter notebooks (kNN and Neural Networks). You only need to download the dataset once for HW3.

1.4 Import the appropriate libraries

```
In [56]: import numpy as np # for doing most of our calculations
         import matplotlib.pyplot as plt# for plotting
         from data.data_utils import load_CIFAR10 # function to load the CIFAR-10 dataset.

         # 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
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

Now, to verify that the dataset has been successfully set up, the following code will print out the shape of train/test data and labels. The output shapes for train/test data are (50000, 32, 32, 3) and (10000, 32, 32, 3), while the labels are (50000,) and (10000,) respectively.

```
In [57]: # Set the path to the CIFAR-10 data
         cifar10_dir = './data/datasets/cifar-10-batches-py'
         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,)
```

Now we visualize some examples from the dataset by showing a few examples of training images from each class.

In [58]:
```
classes = ['plane', 'car', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', 'truc
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()
```

In [59]:
```
# 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)
```

1.5 Implement K-nearest neighbors algorithms

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

```
In [60]: # Import the KNN class
         from hw3code import KNN
```

```
In [61]: # 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)
```

Questions

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

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

Answers

1. The `knn.train()` function is just assigning the passed in X and y values into the model's X_train and y_train. We are using a lazy learning approach to KNN here, meaning that we don't do any real training; instead, we store the training values and then do our prediction/classification only on receiving a new test tuple.
2. The biggest cons to this method are that the time in prediction goes up significantly, as well as the amount of memory needed. However, the "training" time is much less. Additionally, unlike the eager approach, the lazy learning approach doesn't stick to one hypothesis. Instead, it is able to use many local linear functions that can actually emulate a global function for the dataset (what the notes term as a global approximation to the target function).

1.6 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.

```
In [62]: # 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: 50.123364210128784
Frobenius norm of L2 distances: 7906696.077040902
```

1.6.1 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. Normally it may takes 20-40 seconds.

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

1.6.2 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.

```
In [63]: # 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): {}'.f
```

```
Time to run code: 0.388016939163208
Difference in L2 distances between your KNN implementations (should be 0): 0.0
```

1.6.3 Speedup

Depending on your computer speed, you should see a 20-100x speed up from vectorization and no difference in L2 distances between two implementations.

On our computer, the vectorized form took 0.20 seconds while the naive implementation took 26.88 seconds.

1.7 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.

In [64]:
```
# 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

# ================================================================ #
# START 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.
# ================================================================ #
pred_labels = knn.predict_labels(dists_L2_vectorized, k = 1)
# print(pred_labels)
num_samples = pred_labels.shape[0]
count = 0
for i in range(num_samples):
    count += (pred_labels[i] != y_test[i])
error = count / num_samples
# ================================================================ #
# 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.

1.7.1 Questions:

What could you do to improve the accuracy of the k-nearest neighbor classifier you just implemented? Write down your answer in less than 30 words.

1.7.2 Answers:

One way to resolve this is to do some processing on the dataset to make it easier to classify, like with feature/data scaling.

1.8 Optimizing KNN hyperparameters k

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

If you are not familiar with cross validation, cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern. More specifically, in k-fold cross-validation, you evenly split the input data into k subsets of data (also known as folds). You train an ML model on all but one (k-1) of the subsets, and then evaluate the model on the subset that was not used for training. This process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time.

More details of cross validation can be found here. However, you are not allowed to use sklean in your implementation.

1.8.1 Create training and validation folds

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

In [65]:
```
# Create the dataset folds for cross-valdiation.
num_folds = 5

X_train_folds = []
y_train_folds =  []

# ================================================================ #
# START 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]
# ================================================================ #
num_training_exs = X_train.shape[0]
entries_per_fold = int(num_training_exs / num_folds)
rand_indices = np.random.permutation(num_training_exs)
X_train_folds = np.split(X_train[rand_indices], num_folds)
y_train_folds = np.split(y_train[rand_indices], num_folds)

X_train_folds = np.asarray(X_train_folds)
y_train_folds = np.asarray(y_train_folds)

print(X_train_folds)
```

```
        print(y_train_folds)
        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #
```

```
[[[122.  98.  62. ... 184. 117.  76.]
  [206. 235. 226. ... 135. 120. 101.]
  [121. 106. 102. ... 127. 107.  98.]
  ...
  [113. 123.  64. ... 179. 171. 143.]
  [ 53.  65.  53. ...  49.  50.  41.]
  [135. 163. 168. ...  86.  67.  57.]]

 [[209. 205. 197. ... 193. 161.  92.]
  [ 73. 102.  95. ...  40.  65.  38.]
  [186. 180. 184. ... 112. 117. 103.]
  ...
  [219. 205. 205. ...  18.  16.  30.]
  [ 41.  53.  40. ... 186. 190. 179.]
  [ 48. 112. 172. ...  50.  51.  48.]]

 [[ 67.  60.  44. ... 181. 153. 120.]
  [ 49.  41.  30. ... 204. 184. 178.]
  [150. 161. 174. ...  94. 108. 109.]
  ...
  [ 32.  48.  96. ...  27.  42.  83.]
  [149. 158. 189. ... 127. 132. 140.]
  [190. 195. 181. ... 155. 140. 122.]]

 [[ 16.  37.  79. ...  47.  96. 153.]
  [182. 200. 212. ...  74. 163. 135.]
  [125. 127. 126. ... 106. 132.  91.]
  ...
  [152. 161. 175. ...  96.  79.  79.]
  [ 35.  41.  55. ... 104. 120. 107.]
  [ 32.  43.  60. ...  18.  25.  41.]]

 [[180.  63.  80. ...  20.  19.  17.]
  [114. 110. 104. ...  51.  43.  32.]
  [250. 253. 249. ... 254. 253. 254.]
  ...
  [253. 253. 253. ... 172. 163. 102.]
  [ 55.  60.  68. ... 116. 122. 115.]
  [232. 215. 187. ... 215. 209. 197.]]]
[[2 9 2 ... 7 4 9]
 [2 4 7 ... 9 8 0]
 [4 6 8 ... 6 8 5]
 [8 8 2 ... 2 4 8]
```

```
 [9 3 8 ... 4 5 9]]
```

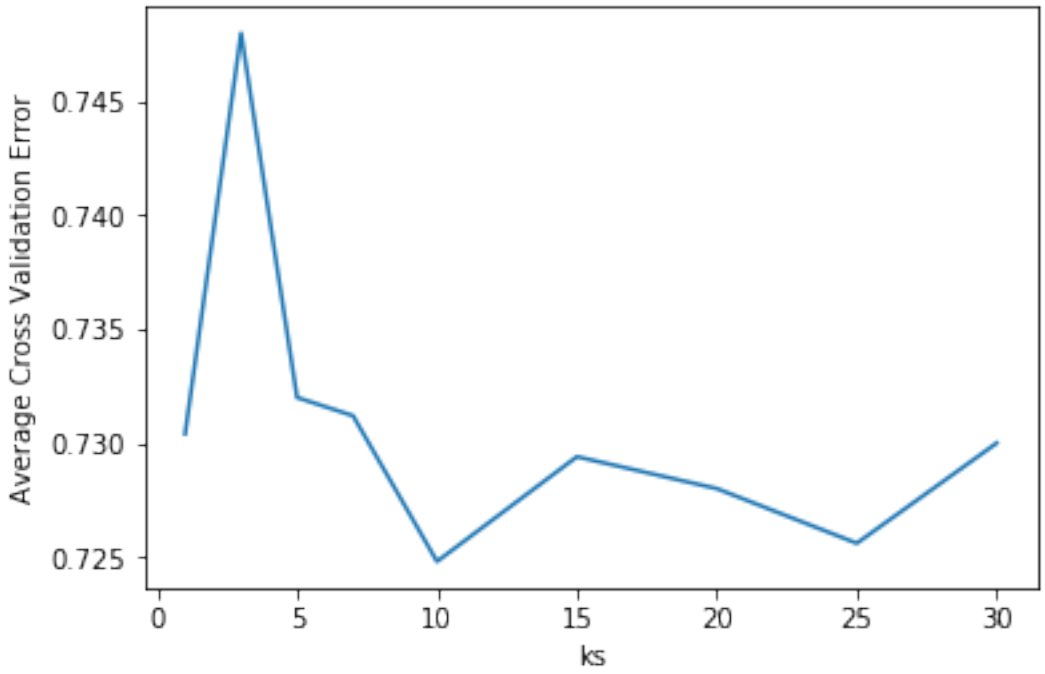
1.8.2 Optimizing the number of nearest neighbors hyperparameter.

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

In [66]:
```
time_start =time.time()

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

# ================================================================ #
# START 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. average cross-validation error.
#   Since we assume L2 distance here, please use the vectorized code!
#   Otherwise, you might be waiting a long time.
# ================================================================ #
knn = KNN()
res = np.zeros(len(ks))
for index,k in enumerate(ks):
    error = 0
    for j in range(num_folds):
        x_t_folds = np.concatenate([X_train_folds[fold] for fold in range(num_folds) if
        y_t_folds = np.concatenate([y_train_folds[fold] for fold in range(num_folds) if
        x_test_fold = X_train_folds[j]
        y_test_fold = y_train_folds[j]
        knn.train(X=x_t_folds, y=y_t_folds) #train on the n-1 folds
        distances = knn.compute_L2_distances_vectorized(X=x_test_fold) # check distance
        pred = knn.predict_labels(distances, k=k) # run prediction
        num_incorrect = np.sum(pred != y_test_fold) #check number of wrong cases
        error += num_incorrect / y_test_fold.shape[0] #create average error rate for mo
    res[index] = error / num_folds # add average error rate to red[index]

ks_min = ks[np.argsort(res)[0]]
results_min = min(res)
# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
print('Set k = {0} and get minimum error as {1}'.format(ks_min,results_min))
plt.plot(ks,res)
plt.xlabel('ks')
plt.ylabel('Average Cross Validation Error')
plt.show()
```

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

```
Set k = 10 and get minimum error as 0.7247999999999999
```




```
Computation time: 47.53
```

Questions:

(1) Why do we typically choose k as an odd number (for exmple in `ks`)

(2) What value of k is best amongst the tested k's? What is the cross-validation error for this value of k?

Answers 1. An odd number for `k` is generally done to avoid situations in which a datapoint has multiple options for the label it could be assigned. Essentially, it's to overcome ties. 2. This value depends on the subset we took of the data at the start, but for the version on this notebook, the best `k=10`, and the cross-validation error is ≈ 0.7248

1.9 Evaluating the model on the testing dataset.

Now, given the optimal k which you have learned, evaluate the testing error of the k-nearest neighbors model.

```
In [67]: error = 1

         # ================================================================ #
         # START YOUR CODE HERE
         # ================================================================ #
         #   Evaluate the testing error of the k-nearest neighbors classifier
         #   for your optimal hyperparameters found by 5-fold cross-validation.
         # ================================================================ #
         knn.train(X=X_train, y=y_train)
         dists_L2_vectorized = knn.compute_L2_distances_vectorized(X=X_test)
         pred_labels = knn.predict_labels(dists_L2_vectorized,k=ks_min)
         num_samples = pred_labels.shape[0]
         num_errors = 0
         for i in range(num_samples):
             if pred_labels[i] != y_test[i]:
                 num_errors = num_errors + 1
         error = num_errors/num_samples

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

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

Error rate achieved: 0.718
```

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

Answers Well, the error went from 0.726 to 0.718, so it's an improvement of 0.08. The dataset may be difficult to separate or classify without further processing of the data, but it's still a pretty good improvement.

1.10 End of Homework 3, Part 1 :)

After you've finished both parts the homework, please print out the both of the entire `ipynb` notebooks and `py` files into one PDF file. Make sure you include the output of code cells and answers for questions. Prepare submit it to GradeScope. Do not include any dataset in your submission.

```python
import numpy as np
import pdb

"""
This code was based off of code from cs231n at Stanford University, and modified for CS145 at UCLA.
"""

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, )
        """
        # ========================================================== #
        # START YOUR CODE HERE
        # ========================================================== #
        #   Hint: KNN does not do any further processsing, just store the training
        #   samples with labels into as self.X_train and self.y_train
        # ========================================================== #
        self.X_train = X
        self.y_train = y
        # ========================================================== #
        # END YOUR CODE HERE
        # ========================================================== #

    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):
            #
========================================================== #
```

```
            # START 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].
            #
=========================================================== #
               diff = X[i] - self.X_train[j]
               dists[i][j] = norm(diff)
            #
=========================================================== #
            # 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))

        # ========================================================== #
        # START 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.
        # ========================================================== #
        M = np.dot(X, (self.X_train).T) # dot procuct between testing and
training data
        train_data_squares = np.square(self.X_train).sum(axis = 1) # sums
across cols the squares of each entry in train data
        # print(train_data_squares.shape)
        test_data_squares = np.square(X).sum(axis = 1) # sums across cols the
squares of each entry in test data
        # print(X)
```

```
        # print("Test Data Squares size and entries")
        # print(test_data_squares.shape)
        # print(test_data_squares)
        test_columnar = test_data_squares.reshape((num_test, 1)) #transforms the sum into a vector form of N,1
        # print("Test Broadcast size and entries")
        # print(test_broadcast.shape)
        # print(test_broadcast)
        dists = np.sqrt(test_columnar + train_data_squares - 2 * M) # Formula from notes
        # ========================================================== #
        # END YOUR CODE HERE
        # ========================================================== #
        # print(dists)
        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 range(num_test):
            # A list of length k storing the labels of the k nearest neighbors to
            # the ith test point.

            closest_y = []

            #
            # ========================================================== #
            # START 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.
            #
            # ========================================================== #
            sorted_dists = np.argsort(dists[i,:])
            closest_ys= sorted_dists[:k]
```

```
151             labels = self.y_train[closest_ys]
152             y_pred[i] = np.bincount(labels).argmax()
153             #
    ========================================================================= #
154             # END YOUR CODE HERE
155             #
    ========================================================================= #
156         return y_pred
157
```

toy_nn

November 11, 2020

1 CS145 Howework 3, Part 2: Neural Networks

Important Note: HW3 is due on **11:59 PM PT, Nov 9 (Monday, Week 6)**. Please submit through GradeScope.

Note that, Howework #3 has two jupyter notebooks to complete (Part 1: kNN and Part 2: Neural Network).

1.1 Print Out Your Name and UID

Name: Devyan Biswas, UID: 804988161

1.2 Before You Start

You need to first create HW3 conda environment by the given `cs145hw3.yml` file, which provides the name and necessary packages for this tasks. If you have `conda` properly installed, you may create, activate or deactivate by the following commands:

```
conda env create -f cs145hw3.yml
conda activate hw3
conda deactivate
```

OR

```
conda env create --name NAMEOFYOURCHOICE -f cs145hw3.yml
conda activate NAMEOFYOURCHOICE
conda deactivate
```

To view the list of your environments, use the following command:

```
conda env list
```

More useful information about managing environments can be found here.

You may also quickly review the usage of basic Python and Numpy package, if needed in coding for matrix operations.

In this notebook, you must not delete any code cells in this notebook. If you change any code outside the blocks (such as hyperparameters) that you are allowed to edit (between `STRART/END YOUR CODE HERE`), you need to highlight these changes. You may add some additional cells to help explain your results and observations.

1.3 Section 1: Backprop in a neural network

Note: Section 1 is "question-answer" style problem. You do not need to code anything and you are required to calculate by hand (with a scientific calculator), which helps you understand the back propagation in neural networks.

In this question, let's consider a simple two-layer neural network and manually do the forward and backward pass. For simplicity, we assume our input data is two dimension. Then the model architecture looks like the following. Notice that in the example we saw in class, the bias term `b` was not explicit listed in the architecture diagram. Here we include the term `b` explicitly for each layer in the diagram. Recall the formula for computing $\mathbf{x}^{(l)}$ in the l-th layer from $\mathbf{x}^{(l-1)}$ in the $(l-1)$-th layer is $\mathbf{x}^{(l)} = \mathbf{f}^{(l)}(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)})$. The activation function $\mathbf{f}^{(l)}$ we choose is the `sigmoid` function for all layers, i.e. $\mathbf{f}^{(l)}(z) = \frac{1}{1+\exp(-z)}$. The final loss function is $\frac{1}{2}$ of the `mean squared error` loss, i.e. $l(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2}||\mathbf{y} - \hat{\mathbf{y}}||^2$.

We initialize our weights as

$$\mathbf{W}^{(1)} = \begin{bmatrix} 0.15 & 0.2 \\ 0.25 & 0.3 \end{bmatrix}, \quad \mathbf{W}^{(2)} = [0.4, 0.45], \quad \mathbf{b}^{(1)} = [0.35, 0.35], \quad \mathbf{b}^{(2)} = 0.6$$

1.3.1 Forward pass

Questions

1. When the input $\mathbf{x}^{(0)} = [0.05, 0.1]$, what will be the value of $\mathbf{x}^{(1)}$ in the hidden layer? (Show your work).
2. Based on the value $\mathbf{x}^{(1)}$ you computed, what will be the value of $\mathbf{x}^{(2)}$ in the output layer? (Show your work).
3. When the target value of this input is $y = 0.01$, based on the value $\mathbf{x}^{(2)}$ you computed, what will be the loss? (Show your work).

Answers:

1. $\mathbf{x}^{(1)} = \mathbf{f}^{(1)}(\begin{bmatrix} 0.15 & 0.2 \\ 0.25 & 0.3 \end{bmatrix} * [0.05, 0.1] + [0.35, 0.35]) \approx [0.593, 0.597]$
2. $\mathbf{x}^{(2)} = \mathbf{f}^{(2)}(\begin{bmatrix} 0.4 & 0.45 \end{bmatrix} * [0.593, 0.597] + 0.6) \approx 0.751$
3. $\frac{(0.01-0.751)^2}{2} \approx 0.275$

1.3.2 Backward pass

With the loss computed below, we are ready for a backward pass to update the weights in the neural network. Kindly remind that the gradients of a variable should have the same shape with the variable.

Questions

1. Consider the loss l of the same input $\mathbf{x}^{(0)} = [0.05, 0.1]$, what will be the update of $\mathbf{W}^{(2)}$ and $\mathbf{b}^{(2)}$ when we backprop, i.e. $\frac{\partial l}{\partial \mathbf{W}^{(2)}}$, $\frac{\partial l}{\partial \mathbf{b}^{(2)}}$ (Show your work in detailed calculation steps. Answers without justification will not be credited.).
2. Based on the result you computed in part 1, when we keep backproping, what will be the update of $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$, i.e. $\frac{\partial l}{\partial \mathbf{W}^{(1)}}$, $\frac{\partial l}{\partial \mathbf{b}^{(1)}}$ (Show your work in details calculation steps. Answers without justification will not be credited.).

Answers:

1. $\frac{\partial l}{\partial \mathbf{W}^{(2)}} => -(y - x^{(2)}) * f^{'}(z^{(2)}) * x^{(1)}$

$z^{(2)} = 0.665$
$f^{'}(z) = \sigma(z)(1 - \sigma(z))$; plugging and chugging:
$-(0.01 - 0.75) * (\sigma(0.665) * (1 - \sigma(0.665)) = \delta^{(2)} = \frac{\partial l}{\partial \mathbf{b}^{(2)}} \approx \boxed{0.1662}$
$\frac{\partial l}{\partial \mathbf{W}^{(2)}} = \delta^{(2)} * x^{(1)} = \boxed{[0.0986\ ,\ 0.0992]}$
2 (Doing one calc for all, since same nodes propogate backward and only one output going back through both)
$\frac{\partial l}{\partial \mathbf{W}_{\mathbf{k}}^{(1)}} => \delta^{(2)} \mathbf{W}_{\mathbf{j}}^{(\mathbf{1})} * f^{'}(z^{(1)}) * x_k^{(0)}$
$z^{(1)} = [0.3775, 0.3925]$
$0.1662 * [0.4, 0.45] * (\sigma(z^{(1)}) * (1 - \sigma(z^{(1)})) = \delta^{(1)} = \frac{\partial l}{\partial \mathbf{b}^{(1)}} \approx \boxed{[0.0321\ ,\ 0.0361]}$
$\frac{\partial l}{\partial \mathbf{W}_{\mathbf{k}}^{(1)}} = \delta^{(1)} * x^{(0)} = \begin{bmatrix} 0.00522 & 0.00522 \\ 0.00522 & 0.00522 \end{bmatrix}$

1.4 Section 2: Coding a two-layer neural network

Import libraries and define relative error function, which is used to check results later.

```
In [23]: import random
         import numpy as np
         from data.data_utils import load_CIFAR10
         import matplotlib.pyplot as plt

         %matplotlib inline
         plt.rcParams['figure.figsize'] = (10.0, 8.0)
         %load_ext autoreload
         %autoreload 2

         def rel_error(x, y):
             """ returns relative error """
             return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

1.5 Toy example

Before loading CIFAR-10, there will be a toy example to test your implementation of the forward and backward pass.

```
In [24]: from hw3code.neural_net import TwoLayerNet
```

```
In [25]: # Create a small net and some toy data to check your implementations.
         # Note that we set the random seed for repeatable experiments.
```

```
input_size = 4
hidden_size = 10
num_classes = 3
num_inputs = 5

def init_toy_model():
    np.random.seed(0)
    return TwoLayerNet(input_size, hidden_size, num_classes, std=1e-1)

def init_toy_data():
    np.random.seed(1)
    X = 10 * np.random.randn(num_inputs, input_size)
    y = np.array([0, 1, 2, 2, 1])
    return X, y

net = init_toy_model()
X, y = init_toy_data()
```

1.5.1 Compute forward pass scores

In [26]:
```
## Implement the forward pass of the neural network.

# Note, there is a statement if y is None: return scores, which is why
# the following call will calculate the scores.
scores = net.loss(X)
print('Your scores:')
print(scores)
print()
print('correct scores:')
correct_scores = np.asarray([
    [-1.07260209,  0.05083871, -0.87253915],
    [-2.02778743, -0.10832494, -1.52641362],
    [-0.74225908,  0.15259725, -0.39578548],
    [-0.38172726,  0.10835902, -0.17328274],
    [-0.64417314, -0.18886813, -0.41106892]])
print(correct_scores)
print()

# The difference should be very small. We get < 1e-7
print('Difference between your scores and correct scores:')
print(np.sum(np.abs(scores - correct_scores)))
```

```
Your scores:
[[-1.07260209  0.05083871 -0.87253915]
 [-2.02778743 -0.10832494 -1.52641362]
 [-0.74225908  0.15259725 -0.39578548]
 [-0.38172726  0.10835902 -0.17328274]
```

```
 [-0.64417314 -0.18886813 -0.41106892]]

correct scores:
[[-1.07260209  0.05083871 -0.87253915]
 [-2.02778743 -0.10832494 -1.52641362]
 [-0.74225908  0.15259725 -0.39578548]
 [-0.38172726  0.10835902 -0.17328274]
 [-0.64417314 -0.18886813 -0.41106892]]

Difference between your scores and correct scores:
3.381231204052648e-08
```

1.5.2 Forward pass loss

The total loss includes data loss (MSE) and regularization loss, which is,

$$L = L_{data} + L_{reg} = \frac{1}{2N}\sum_{i=1}^{N}\left(\boldsymbol{y}_{\text{pred}} - \boldsymbol{y}_{\text{target}}\right)^2 + \frac{\lambda}{2}\left(||W_1||^2 + ||W_2||^2\right)$$

More specifically in multi-class situation, if the output of neural nets from one sample is $y_{\text{pred}} = (0.1, 0.1, 0.8)$ and $y_{\text{target}} = (0, 0, 1)$ from the given label, then the MSE error will be $Error = (0.1 - 0)^2 + (0.1 - 0)^2 + (0.8 - 1)^2 = 0.06$

Implement data loss and regularization loss. In the MSE function, you also need to return the gradients which need to be passed backward. This is similar to batch gradient in linear regression. Test your implementation of loss functions. The Difference should be less than 1e-12.

```
In [27]: loss, _ = net.loss(X, y, reg=0.05)
         correct_loss_MSE = 1.8973332763705641

         # should be very small, we get < 1e-12
         print('Difference between your loss and correct loss:')
         print(np.sum(np.abs(loss - correct_loss_MSE)))
```

```
Difference between your loss and correct loss:
0.0
```

1.5.3 Backward pass (You do not need to implemented this part)

We have already implemented the backwards pass of the neural network for you. Run the block of code to check your gradients with the gradient check utilities provided. The results should be automatically correct (tiny relative error).

If there is a gradient error larger than 1e-8, the training for neural networks later will be negatively affected.

```
In [28]: from data.gradient_check import eval_numerical_gradient

         # Use numeric gradient checking to check your implementation of the backward pass.
```

```
        # If your implementation is correct, the difference between the numeric and
        # analytic gradients should be less than 1e-8 for each of W1, W2, b1, and b2.

        loss, grads = net.loss(X, y, reg=0.05)

        # these should all be less than 1e-8 or so
        for param_name in grads:
            f = lambda W: net.loss(X, y, reg=0.05)[0]
            param_grad_num = eval_numerical_gradient(f, net.params[param_name], verbose=False)
            print('{} max relative error: {}'.format(param_name, rel_error(param_grad_num, grad
```

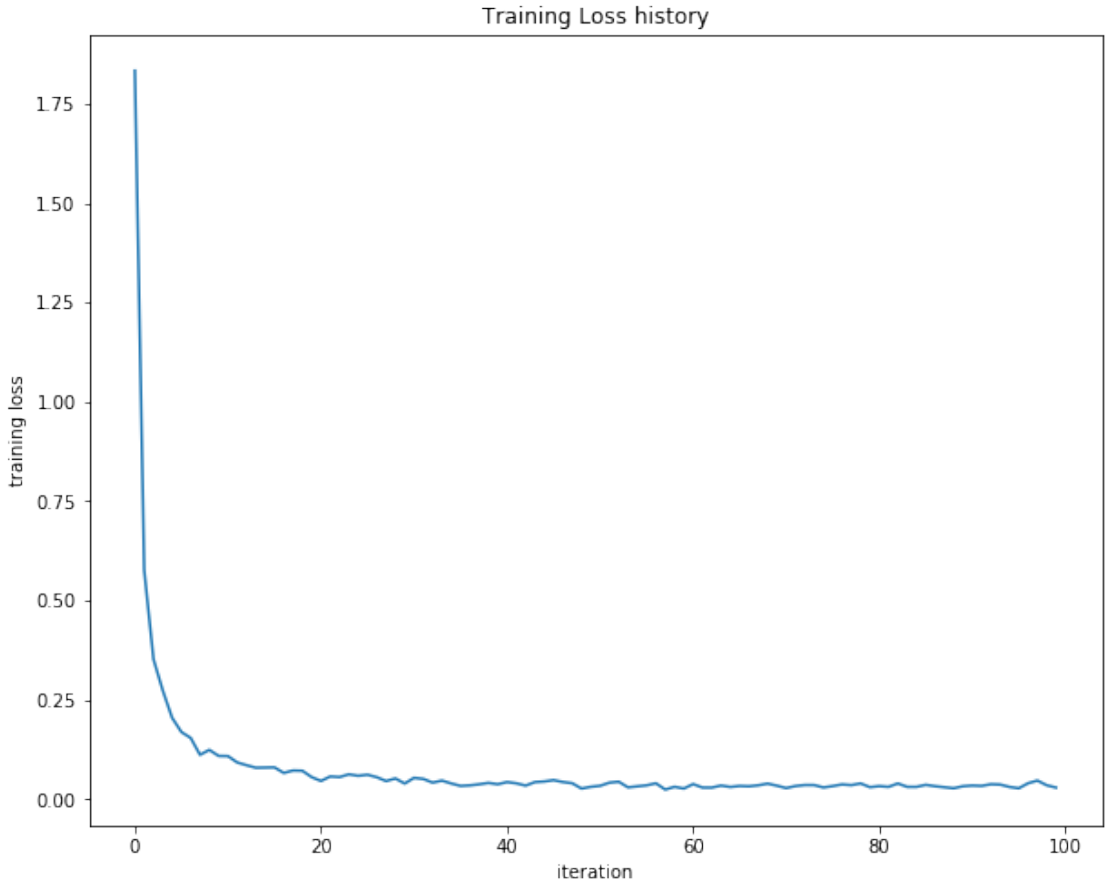
```
W2 max relative error: 8.80091875172355e-11
b2 max relative error: 2.4554844805570154e-11
W1 max relative error: 1.7476665046687833e-09
b1 max relative error: 7.382451041178829e-10
```

1.5.4 Training the network

Implement neural_net.train() to train the network via stochastic gradient descent, much like the linear regression.

```
In [29]: net = init_toy_model()
         stats = net.train(X, y, X, y,
                     learning_rate=1e-1, reg=5e-6,
                     num_iters=100, verbose=False)

         print('Final training loss: ', stats['loss_history'][-1])

         # plot the loss history
         plt.plot(stats['loss_history'])
         plt.xlabel('iteration')
         plt.ylabel('training loss')
         plt.title('Training Loss history')
         plt.show()
```

```
Final training loss:  0.02950555626206818
```



1.6 Classify CIFAR-10

Do classification on the CIFAR-10 dataset.

```
In [30]: from data.data_utils import load_CIFAR10

         def get_CIFAR10_data(num_training=49000, num_validation=1000, num_test=1000):
             """
             Load the CIFAR-10 dataset from disk and perform preprocessing to prepare
             it for the two-layer neural net 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 = './data/datasets/cifar-10-batches-py'
             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]

    # 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

    # Reshape data to rows
    X_train = X_train.reshape(num_training, -1)
    X_val = X_val.reshape(num_validation, -1)
    X_test = X_test.reshape(num_test, -1)

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


# Invoke the above function to get our data.
X_train, y_train, X_val, y_val, X_test, y_test = 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)
```

```
Train data shape:  (49000, 3072)
Train labels shape:  (49000,)
Validation data shape:  (1000, 3072)
Validation labels shape:  (1000,)
Test data shape:  (1000, 3072)
Test labels shape:  (1000,)
```

1.6.1 Running SGD

If your implementation is correct, you should see a validation accuracy of around 15-18%.

```
In [40]: input_size = 32 * 32 * 3
         hidden_size = 50
         num_classes = 10
         net = TwoLayerNet(input_size, hidden_size, num_classes)
```

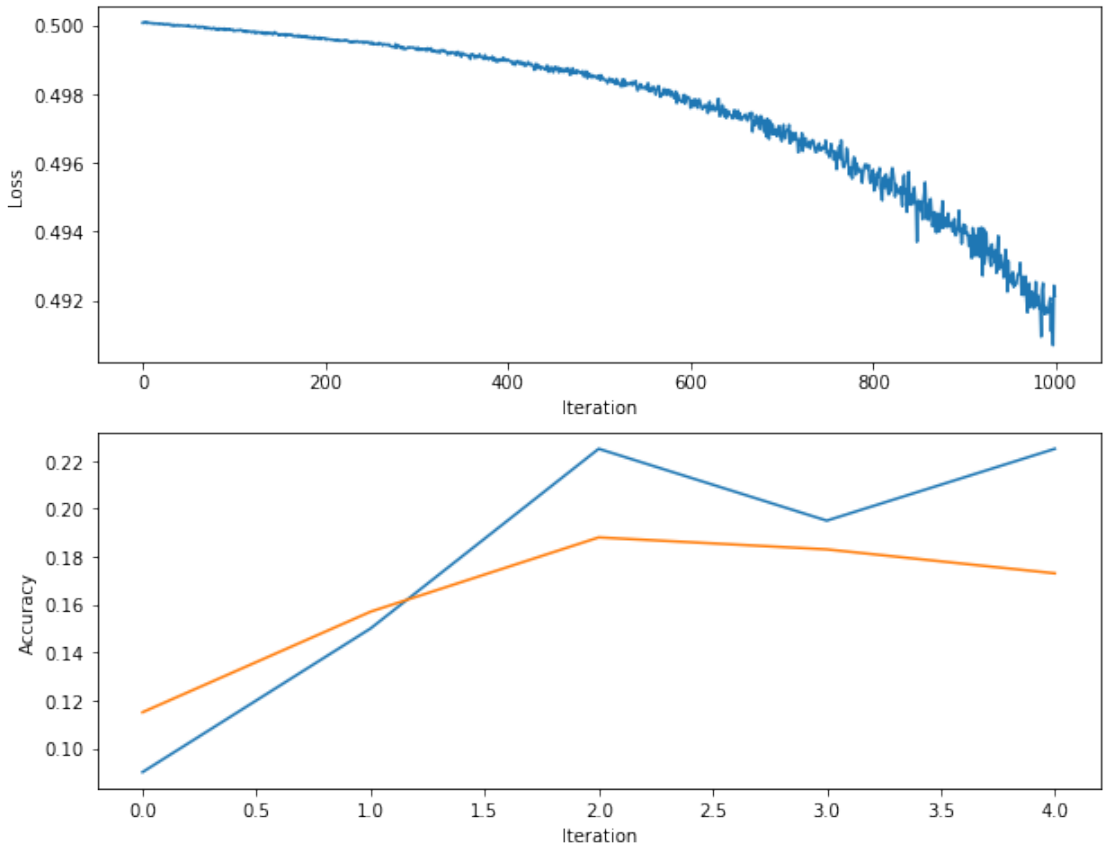
```
        # Train the network
        stats = net.train(X_train, y_train, X_val, y_val,
                    num_iters=1000, batch_size=200,
                    learning_rate=1e-5, learning_rate_decay=0.95,
                    reg=0.1, verbose=True)

        # Predict on the validation set
        val_acc = (net.predict(X_val) == y_val).mean()
        print('Validation accuracy: ', val_acc)

        # Save this net as the variable subopt_net for later comparison.
        subopt_net = net
        test_acc = (subopt_net.predict(X_test) == y_test).mean()
        print('Test accuracy (subopt_net): ', test_acc)
```

```
iteration 0 / 1000: loss 0.5000653841132241
iteration 100 / 1000: loss 0.4998690749422438
iteration 200 / 1000: loss 0.4996798178691218
iteration 300 / 1000: loss 0.4994397950516031
iteration 400 / 1000: loss 0.4991026153222729
iteration 500 / 1000: loss 0.49884677571338243
iteration 600 / 1000: loss 0.4981137879077918
iteration 700 / 1000: loss 0.49755140486802873
iteration 800 / 1000: loss 0.4965314440354824
iteration 900 / 1000: loss 0.49479052775342364
Validation accuracy:  0.173
Test accuracy (subopt_net):  0.186
```

```
In [32]: stats['train_acc_history']
```

```
Out[32]: [0.09, 0.15, 0.225, 0.195, 0.225]
```

```
In [33]: # Plot the loss function and train / validation accuracies
         plt.subplot(2, 1, 1)
         plt.plot(stats['loss_history'])
         plt.xlabel('Iteration')
         plt.ylabel('Loss')

         plt.subplot(2, 1, 2)
         plt.plot(stats['train_acc_history'], label='train')
         plt.plot(stats['val_acc_history'], label='val')
         plt.xlabel('Iteration')
         plt.ylabel('Accuracy')

         plt.show()
```



Questions:

The training accuracy isn't great. It seems even worse than simple KNN model, which is not as good as expected.

(1) What are some of the reasons why this is the case? Based on previous observations, please provide at least two possible reasons with justification.

(2) How should you fix the problems you identified in (1)?

Answers:

1. The first thing that comes to mind is that this is only a 2-Layer network, so low accuracy is to be expected since our dataset likely cannot be accurately handled with a low-powered neural network. Another reason could be that our hyperparams are not accurate; given that we only tested with a small set of possible hyperparameters, our neural network is limited.
2. Of course, with tuning, our issue of hyperparameter inaccuracy will go down. Additionally, we can increase the number of hidden layers, allowing for more complex data processing.

1.7 Optimize the neural network

Use the following part of the Jupyter notebook to optimize your hyperparameters on the validation set. Store your nets as best_net. To get the full credit of the neural nets, you should get at least **45%** accuracy on validation set.

Reminder: Think about whether you should retrain a new model from scratch every time your try a new set of hyperparameters.

In [34]:
```
best_net = None # store the best model into this

# ================================================================ #
# START YOUR CODE HERE:
# ================================================================ #
#   Optimize over your hyperparameters to arrive at the best neural
#   network.  You should be able to get over 45% validation accuracy.
#   For this part of the notebook, we will give credit based on the
#   accuracy you get.  Your score on this question will be multiplied by:
#      min(floor((X - 23%)) / %22, 1)
#   where if you get 50% or higher validation accuracy, you get full
#   points.
#
#   Note, you need to use the same network structure (keep hidden_size = 50)!
# ================================================================ #

# todo: optimal parameter search (you may use grid search by for-loops )
input_size = 32 * 32 * 3
hidden_size = 50
num_classes = 10
best_valid_acc = 0

# net = TwoLayerNet(input_size, hidden_size, num_classes)
# net = init_toy_model()
# Train the network and find best parameter:

opt_settings_str = ''
lrs = [1e-4, 1e-3]
decays = [0.8, 0.75, 0.6]
regularizations = [1, 2, 4]

for lr in lrs:
    for d in decays:
        for regz in regularizations:
            stats = net.train(X_train, y_train, X_val, y_val, num_iters=1000, batch_siz
            valid_acc = (net.predict(X_val) == y_val).mean()
            print('Validation accuracy: ', valid_acc)
            print('learning_rate: {}, iterations: {}, batch_sizes: {}, lrdecay: {}, reg
            if valid_acc > best_valid_acc:
                best_valid_acc = valid_acc
                best_net = net
                opt_settings_str = 'learning_rate: {}, iterations: {}, batch_sizes: {},


# ================================================================ #
```

```
            # END YOUR CODE HERE
            # ================================================================ #
            # Output your results
            print("== Best parameter settings ==")
            print(opt_settings_str)
            print("Best accuracy on validation set: {}".format(best_valid_acc))
```

```
iteration 0 / 1000: loss 0.491828519168669
iteration 100 / 1000: loss 0.4593743449019668
iteration 200 / 1000: loss 0.4480651160670165
iteration 300 / 1000: loss 0.44335528959711595
iteration 400 / 1000: loss 0.4452274893069379
iteration 500 / 1000: loss 0.43910652028271485
iteration 600 / 1000: loss 0.4344220339248548
iteration 700 / 1000: loss 0.4354469254224735
iteration 800 / 1000: loss 0.4376013062916228
iteration 900 / 1000: loss 0.428242941065242
Validation accuracy:  0.264
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.8, regularization: 1
iteration 0 / 1000: loss 0.4340652324416934
iteration 100 / 1000: loss 0.4306596476220053
iteration 200 / 1000: loss 0.42239392452453467
iteration 300 / 1000: loss 0.4154876727613203
iteration 400 / 1000: loss 0.41857104571277093
iteration 500 / 1000: loss 0.4154808619073914
iteration 600 / 1000: loss 0.41937266564593945
iteration 700 / 1000: loss 0.4196167011193107
iteration 800 / 1000: loss 0.4126504399414269
iteration 900 / 1000: loss 0.4172734956739185
Validation accuracy:  0.307
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.8, regularization: 2
iteration 0 / 1000: loss 0.42541767522510265
iteration 100 / 1000: loss 0.41454760918284805
iteration 200 / 1000: loss 0.40817473528668957
iteration 300 / 1000: loss 0.4121884993652861
iteration 400 / 1000: loss 0.41326292821382576
iteration 500 / 1000: loss 0.4085735662435541
iteration 600 / 1000: loss 0.41198020737559043
iteration 700 / 1000: loss 0.40466962529437694
iteration 800 / 1000: loss 0.40684672049226916
iteration 900 / 1000: loss 0.40830316077315343
Validation accuracy:  0.337
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.8, regularization: 4
iteration 0 / 1000: loss 0.40385631596372557
iteration 100 / 1000: loss 0.39553689152016397
iteration 200 / 1000: loss 0.40809878914139663
iteration 300 / 1000: loss 0.39564180747654276
iteration 400 / 1000: loss 0.4040007266188983
```

```
iteration 500 / 1000: loss 4.0020189655775834
iteration 600 / 1000: loss 3.9857880112706795
iteration 700 / 1000: loss 4.0049433828780739
iteration 800 / 1000: loss 3.9986725078900686
iteration 900 / 1000: loss 4.0429988779609255
Validation accuracy:  0.359
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.75, regularization: 1
iteration 0 / 1000: loss 4.0026028110560171
iteration 100 / 1000: loss 4.0058554045957627
iteration 200 / 1000: loss 3.9806453983719103
iteration 300 / 1000: loss 3.9582881424568783
iteration 400 / 1000: loss 3.9993962018531759
iteration 500 / 1000: loss 4.0010262521148996
iteration 600 / 1000: loss 3.9958181890379916
iteration 700 / 1000: loss 3.9947632080731116
iteration 800 / 1000: loss 3.9129350845641403
iteration 900 / 1000: loss 3.9922696777303684
Validation accuracy:  0.375
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.75, regularization: 2
iteration 0 / 1000: loss 4.0090515871947538
iteration 100 / 1000: loss 4.0055587388220986
iteration 200 / 1000: loss 3.9312697122879464
iteration 300 / 1000: loss 3.9977079524053877
iteration 400 / 1000: loss 4.0079274296376304
iteration 500 / 1000: loss 3.9943891282957726
iteration 600 / 1000: loss 4.0528261145629546
iteration 700 / 1000: loss 4.0779912107005195
iteration 800 / 1000: loss 4.0015813000859951
iteration 900 / 1000: loss 4.0269052270689804
Validation accuracy:  0.389
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.75, regularization: 4
iteration 0 / 1000: loss 3.9152767988323434
iteration 100 / 1000: loss 3.8850913545049704
iteration 200 / 1000: loss 3.93466663692234
iteration 300 / 1000: loss 3.8843112784972863
iteration 400 / 1000: loss 3.9002898305056159
iteration 500 / 1000: loss 3.8994587314742592
iteration 600 / 1000: loss 3.8825700638736951
iteration 700 / 1000: loss 3.8860121914990725
iteration 800 / 1000: loss 3.9074373916264726
iteration 900 / 1000: loss 3.8850524936569003
Validation accuracy:  0.405
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 1
iteration 0 / 1000: loss 3.9028015185169188
iteration 100 / 1000: loss 3.8992813120306964 4
iteration 200 / 1000: loss 3.9038844873086486
iteration 300 / 1000: loss 3.8476176183982563
iteration 400 / 1000: loss 3.8488054412851075
```

```
iteration 500 / 1000: loss 0.3843954747814342
iteration 600 / 1000: loss 0.3916274156343006
iteration 700 / 1000: loss 0.3832640840674542
iteration 800 / 1000: loss 0.395536014188009
iteration 900 / 1000: loss 0.3899887218715062
Validation accuracy:  0.416
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 2
iteration 0 / 1000: loss 0.39655607477874205
iteration 100 / 1000: loss 0.40032458804960136
iteration 200 / 1000: loss 0.4005033239302925
iteration 300 / 1000: loss 0.3957050390554905
iteration 400 / 1000: loss 0.401247187267484
iteration 500 / 1000: loss 0.4040651440203802
iteration 600 / 1000: loss 0.3961496951348371
iteration 700 / 1000: loss 0.3961830514682346
iteration 800 / 1000: loss 0.39390753937566975
iteration 900 / 1000: loss 0.400503938138193
Validation accuracy:  0.424
learning_rate: 0.0001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 4
iteration 0 / 1000: loss 0.39509424094899426
iteration 100 / 1000: loss 0.38422287282679674
iteration 200 / 1000: loss 0.39042892398391316
iteration 300 / 1000: loss 0.36675894740367404
iteration 400 / 1000: loss 0.37080681040512925
iteration 500 / 1000: loss 0.37788172958659516
iteration 600 / 1000: loss 0.3724291575982607
iteration 700 / 1000: loss 0.3633699006934623
iteration 800 / 1000: loss 0.36048350399392637
iteration 900 / 1000: loss 0.3707601036243043
Validation accuracy:  0.476
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.8, regularization: 1
iteration 0 / 1000: loss 0.3688431877691851
iteration 100 / 1000: loss 0.3899867335338655
iteration 200 / 1000: loss 0.38068339277648333
iteration 300 / 1000: loss 0.37525616397667666
iteration 400 / 1000: loss 0.3706662612130924
iteration 500 / 1000: loss 0.3742766362521676
iteration 600 / 1000: loss 0.38262336044069034
iteration 700 / 1000: loss 0.3800716725313875
iteration 800 / 1000: loss 0.37358215692614993
iteration 900 / 1000: loss 0.37120669073102025
Validation accuracy:  0.474
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.8, regularization: 2
iteration 0 / 1000: loss 0.39575525869352207
iteration 100 / 1000: loss 0.3928015609552376
iteration 200 / 1000: loss 0.38916966122415503
iteration 300 / 1000: loss 0.3797780923172795
iteration 400 / 1000: loss 0.3903785101935017
```

```
iteration 500 / 1000: loss 0.38711647185236386
iteration 600 / 1000: loss 0.3958934468662587
iteration 700 / 1000: loss 0.38876964215132825
iteration 800 / 1000: loss 0.3836799607901521
iteration 900 / 1000: loss 0.39241119383675616
Validation accuracy:  0.458
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.8, regularization: 4
iteration 0 / 1000: loss 0.36340771337618044
iteration 100 / 1000: loss 0.37142853822435845
iteration 200 / 1000: loss 0.37187505056103887
iteration 300 / 1000: loss 0.3646539629783366
iteration 400 / 1000: loss 0.3673818393854741
iteration 500 / 1000: loss 0.3601854811160642
iteration 600 / 1000: loss 0.35799604936668106
iteration 700 / 1000: loss 0.36772029079358937
iteration 800 / 1000: loss 0.36599811158486495
iteration 900 / 1000: loss 0.3643512630912791
Validation accuracy:  0.48
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.75, regularization: 1
iteration 0 / 1000: loss 0.3771556355523204
iteration 100 / 1000: loss 0.36430739663007317
iteration 200 / 1000: loss 0.37719774454599436
iteration 300 / 1000: loss 0.36174202905625413
iteration 400 / 1000: loss 0.3771642404024748
iteration 500 / 1000: loss 0.37721489889045107
iteration 600 / 1000: loss 0.3774499236730192
iteration 700 / 1000: loss 0.3703718348821798
iteration 800 / 1000: loss 0.36762177220887654
iteration 900 / 1000: loss 0.3830401480792569
Validation accuracy:  0.483
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.75, regularization: 2
iteration 0 / 1000: loss 0.3867937184721809
iteration 100 / 1000: loss 0.38203997973860415
iteration 200 / 1000: loss 0.390956099625095
iteration 300 / 1000: loss 0.3956644182966086
iteration 400 / 1000: loss 0.3874746849416148
iteration 500 / 1000: loss 0.38488345004615176
iteration 600 / 1000: loss 0.39746644989955304
iteration 700 / 1000: loss 0.390113385197794
iteration 800 / 1000: loss 0.3909561111766971
iteration 900 / 1000: loss 0.39243666230239665
Validation accuracy:  0.475
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.75, regularization: 4
iteration 0 / 1000: loss 0.3658066954949327
iteration 100 / 1000: loss 0.3626337811809807
iteration 200 / 1000: loss 0.35084285352458483
iteration 300 / 1000: loss 0.355869653546313
iteration 400 / 1000: loss 0.36383910950187764
```

```
iteration 500 / 1000: loss 0.35283204614912134
iteration 600 / 1000: loss 0.35836518538838724
iteration 700 / 1000: loss 0.3638619927574204
iteration 800 / 1000: loss 0.3542773530403485
iteration 900 / 1000: loss 0.35525510994313925
Validation accuracy:  0.489
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 1
iteration 0 / 1000: loss 0.3599081502883241
iteration 100 / 1000: loss 0.3693910462572361
iteration 200 / 1000: loss 0.36494867355108446
iteration 300 / 1000: loss 0.3733823385804047
iteration 400 / 1000: loss 0.36671532119959327
iteration 500 / 1000: loss 0.3718087493182603
iteration 600 / 1000: loss 0.37440713756780397
iteration 700 / 1000: loss 0.37778253025779024
iteration 800 / 1000: loss 0.3756751103797165
iteration 900 / 1000: loss 0.37839581396212163
Validation accuracy:  0.483
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 2
iteration 0 / 1000: loss 0.38745729727112105
iteration 100 / 1000: loss 0.38286401747762755
iteration 200 / 1000: loss 0.39013155795851434
iteration 300 / 1000: loss 0.38659207567764176
iteration 400 / 1000: loss 0.3945722207389524
iteration 500 / 1000: loss 0.3903150274614551
iteration 600 / 1000: loss 0.3892858420266587
iteration 700 / 1000: loss 0.39455983404110007
iteration 800 / 1000: loss 0.39312300155899255
iteration 900 / 1000: loss 0.3868618911507201
Validation accuracy:  0.48
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 4
== Best parameter settings ==
learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 1
Best accuracy on validation set: 0.489
```

Quesions

(1) What is your best parameter settings? (Output from the previous cell)

(2) What parameters did you tune? How are they changing the performance of nerural network? You can discuss any observations from the optimization.

Answers

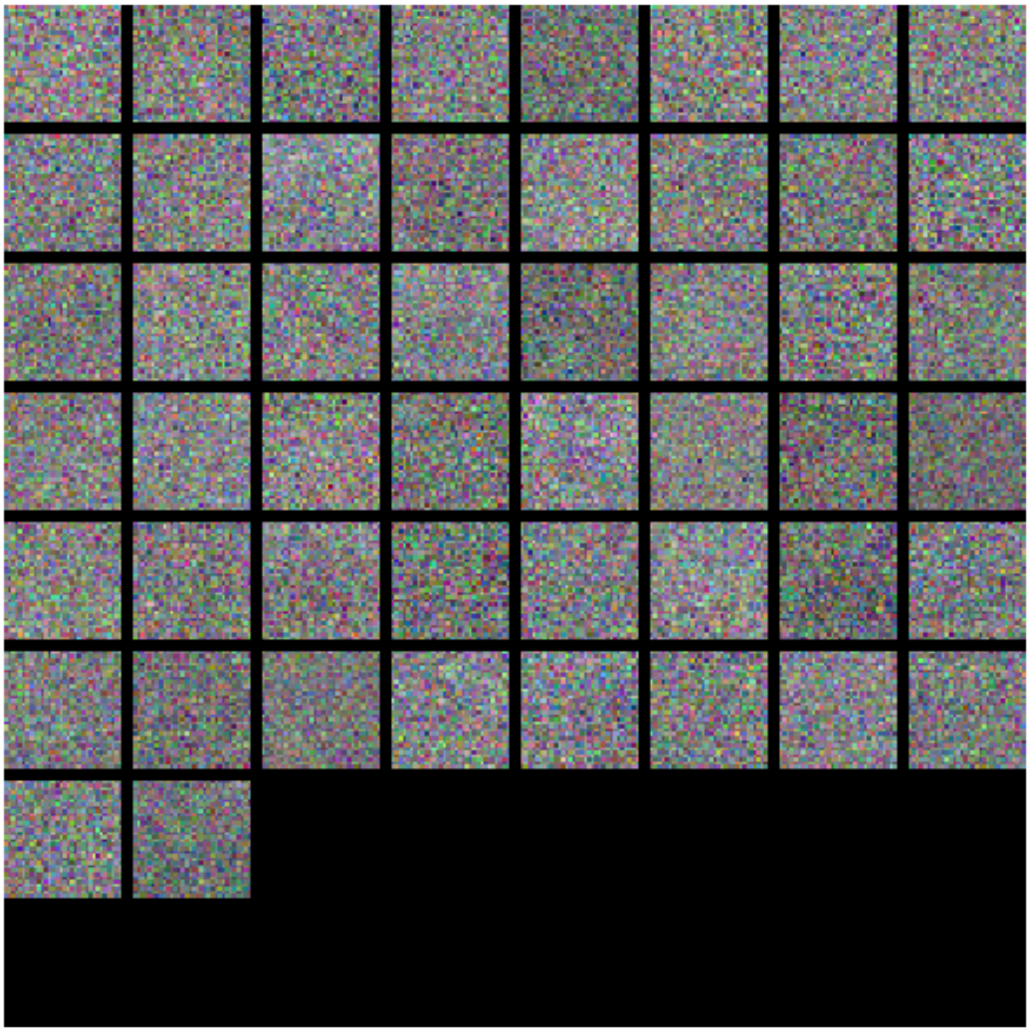
1. Based on the above, the ideal parameters for this subset are: `learning_rate: 0.001, iterations: 1000, batch_sizes: 500, lrdecay: 0.6, regularization: 1`
2. Given that the params that result in the greatest change in response to tuning were the learning rate, learning rate decay, and regularization, I chose to optimize on those parameters, using a few subsets of values for each.

1.8 Visualize the weights of your neural networks

In [42]:
```
from data.vis_utils import visualize_grid

# Visualize the weights of the network

def show_net_weights(net):
    W1 = net.params['W1']
    W1 = W1.T.reshape(32, 32, 3, -1).transpose(3, 0, 1, 2)
    plt.imshow(visualize_grid(W1, padding=3).astype('uint8'))
    plt.gca().axis('off')
    plt.show()

#Had to re-run the subopt one to get the right params, since I used the net obj kinda r
print("Subpopt")
print(subopt_net.params['W1'])
print("Opt")
print(best_net.params['W1'])
show_net_weights(subopt_net)
show_net_weights(best_net)
```

```
Subpopt
[[-7.40695078e-05 -3.49500350e-05 -4.19172611e-05 ...  7.22647179e-06
  -1.15812816e-06 -3.25142389e-05]
 [ 5.35857596e-07 -2.15522067e-06  6.72090349e-05 ...  2.62624771e-06
   6.89106405e-05  2.66278742e-05]
 [ 3.56652442e-05  1.78354344e-04  1.20388108e-05 ...  7.67448887e-05
  -1.49414583e-04  7.05453706e-05]
 ...
 [ 9.33278981e-06  9.96739146e-05 -6.16069902e-05 ... -2.63987006e-05
   7.74001900e-05  1.59241077e-04]
 [-1.03445726e-04  5.77482886e-05  1.20718286e-04 ...  2.07328612e-05
   2.96483766e-05 -6.06061271e-05]
 [-1.41636813e-04  1.29133814e-04  7.17963843e-05 ...  2.49462398e-06
  -7.40524730e-05  1.36491449e-05]]
Opt
[[-3.80722349e-04 -4.88864580e-04 -2.54289992e-04 ...  1.32588075e-04
   1.80212506e-04  4.14105981e-05]
 [-6.76622138e-04 -5.30150255e-04 -8.52004578e-05 ...  1.95883085e-04
   7.29646147e-05  4.45374032e-04]
 [-1.94392771e-04 -1.38243373e-04 -2.72851410e-04 ...  2.93340172e-04
   1.20403374e-04  1.43483987e-04]
 ...
 [-9.88778318e-05 -5.11932411e-05  1.07321192e-04 ...  3.91029391e-05
  -3.77051506e-04 -6.95368976e-04]
 [-5.50675183e-05  7.72192394e-05  1.74916413e-04 ... -6.17585636e-04
```

```
 -5.44980610e-04 -3.38554647e-04]
[ 2.69372922e-04  1.74520235e-04 -5.83476635e-04 ...  1.82538235e-04
  3.01571298e-05 -7.28183748e-05]]
```



Questions:

What differences do you see in the weights between the suboptimal net and the best net you arrived at? What do the weights in neural networks probably learn after training?

Answer:

The weights are far less noisy, and the features are more distinct in our optimized neural network. The weights likely "learned" to distinguish the different attributes in the CIFAR dataset.

1.9 Evaluate on test set

```
In [43]: test_acc = (best_net.predict(X_test) == y_test).mean()
         print('Test accuracy (best_net): ', test_acc)

Test accuracy (best_net):  0.465
```

Questions:

(1) What is your test accuracy by using the best NN you have got? How much does the performance increase compared with kNN? Why can neural networks perform better than kNN?

(2) Do you have any other ideas or suggestions to further improve the performance of neural networks other than the parameters you have tried in the homework?

Answers:

1. The test accuracy is 46.5%. It's a lot better than out kNN approach. That had an error rate of 71.8%, so an accuracy of 28.2%. kNN, while a quick training algo, loses out overall. kNN is limited to the weights and data in a direct way: where it is measuring the distance between features of datapoints. Neural nets, on the other hand, have multiple hidden layers whose outputs cascade into others. This allows for a finer granularity in making distinctions between features and, therefore, increase the accuracy of classifying.
2. Only thing I can think of is what I said before, which is more layers to leverage what I said in 1.

1.10 Bonus Question: Change MSE Loss to Cross Entropy Loss

This is a bonus question. If you finish this (cross entropy loss) correctly, you will get **up to 10 points** (add up to your HW3 score).

Note: From grading policy of this course, your maximum points from homework are still 25 out of 100, but you can use the bonus question to make up other deduction of other assignments.

Pass output scores in networks from forward pass into softmax function. The softmax function is defined as,

$$p_j = \sigma(z_j) = \frac{e^{z_j}}{\sum_{c=1}^{C} e^{z_c}}$$

After softmax, the scores can be considered as probability of j-th class.

The cross entropy loss is defined as,

$$L = L_{\text{CE}} + L_{reg} = \frac{1}{N}\sum_{i=1}^{N} \log\left(p_{i,j}\right) + \frac{\lambda}{2}\left(||W_1||^2 + ||W_2||^2\right)$$

To take derivative of this loss, you will get the gradient as,

$$\frac{\partial L_{\text{CE}}}{\partial o_i} = p_i - y_i$$

More details about multi-class cross entropy loss, please check http://cs231n.github.io/linear-classify/ and more explanation about the derivative of cross entropy.

Change the loss from MSE to cross entropy, you only need to change you `MSE_loss(x,y)` in `TwoLayerNet.loss()` function to `softmax_loss(x,y)`.

Now you are free to use any code to show your results of the two-layer networks with newly-implemented cross entropy loss. You can use code from previous cells.

```
In [ ]: # Start training your networks and show your results
        # ================================================================ #
        # START YOUR CODE HERE:
        # ================================================================ #
        pass
        # ================================================================ #
        # END YOUR CODE HERE
        # ================================================================ #
```

1.11 End of Homework 3, Part 2 :)

After you've finished both parts the homework, please print out the both of the entire `ipynb` notebooks and `py` files into one PDF file. Make sure you include the output of code cells and answers for questions. Prepare submit it to GradeScope. Do not include any dataset in your submission.

```python
import numpy as np
import matplotlib.pyplot as plt

class TwoLayerNet(object):
    """
    A two-layer fully-connected neural network. The net has an input dimension of
    N, a hidden layer dimension of H, and performs classification over C classes.
    We train the network with a softmax loss function and L2 regularization on the
    weight matrices. The network uses a ReLU nonlinearity after the first fully
    connected layer.

    In other words, the network has the following architecture:

    input - fully connected layer - ReLU - fully connected layer - MSE Loss

    ReLU function:
    (i) x = x if x >= 0  (ii) x = 0 if x < 0

    The outputs of the second fully-connected layer are the scores for each class.
    """

    def __init__(self, input_size, hidden_size, output_size, std=1e-4):
        """
        Initialize the model. Weights are initialized to small random values and
        biases are initialized to zero. Weights and biases are stored in the
        variable self.params, which is a dictionary with the following keys:

        W1: First layer weights; has shape (H, D)
        b1: First layer biases; has shape (H,)
        W2: Second layer weights; has shape (C, H)
        b2: Second layer biases; has shape (C,)

        Inputs:
        - input_size: The dimension D of the input data.
        - hidden_size: The number of neurons H in the hidden layer.
        - output_size: The number of classes C.
        """
        self.params = {}
        self.params['W1'] = std * np.random.randn(hidden_size, input_size)
        self.params['b1'] = np.zeros(hidden_size)
        self.params['W2'] = std * np.random.randn(output_size, hidden_size)
        self.params['b2'] = np.zeros(output_size)

    def loss(self, X, y=None, reg=0.0):
        """
        Compute the loss and gradients for a two layer fully connected neural
        network.

        Inputs:
        - X: Input data of shape (N, D). Each X[i] is a training sample.
        - y: Vector of training labels. y[i] is the label for X[i], and each y[i] is
          an integer in the range 0 <= y[i] < C. This parameter is optional; if it
```

```
53            is not passed then we only return scores, and if it is passed then
      we
54            instead return the loss and gradients.
55          - reg: Regularization strength.
56
57          Returns:
58          If y is None, return a matrix scores of shape (N, C) where scores[i,
      c] is
59          the score for class c on input X[i].
60
61          If y is not None, instead return a tuple of:
62          - loss: Loss (data loss and regularization loss) for this batch of
      training
63            samples.
64          - grads: Dictionary mapping parameter names to gradients of those
      parameters
65            with respect to the loss function; has the same keys as
      self.params.
66          """
67          # Unpack variables from the params dictionary
68          W1, b1 = self.params['W1'], self.params['b1']
69          W2, b2 = self.params['W2'], self.params['b2']
70          N, D = X.shape
71
72          # Compute the forward pass
73          scores = None
74
75          # ============================================================== #
76          # START YOUR CODE HERE
77          # ============================================================== #
78          #   Calculate the output scores of the neural network.  The result
79          #   should be (N, C). As stated in the description for this class,
80          #   there should not be a ReLU layer after the second fully-connected
81          #   layer.
82          #   The code is partially given
83          #   The output of the second fully connected layer is the output
      scores.
84          #   Do not use a for loop in your implementation.
85          #   Please use 'h1' as input of hidden layers, and 'a2' as output of
86          #   hidden layers after ReLU activation function.
87          #   [Input X] --W1,b1--> [h1] -ReLU-> [a2] --W2,b2--> [scores]
88          #   You may simply use np.maximun for implementing ReLU.
89          #   Note that there is only one ReLU layer.
90          #   Note that plase do not change the variable names (h1, h2, a2)
91          # ============================================================== #
92
93          h1 = np.dot(X,W1.T) + b1
94          a2 = np.zeros(h1.shape)
95          a2 = np.maximum(a2, h1) # activation with input of h1
96          h2 = np.dot(a2,W2.T) + b2
97          scores = h2
98
99          # ============================================================== #
100         # END YOUR CODE HERE
101         # ============================================================== #
102
103
104         # If the targets are not given then jump out, we're done
105         if y is None:
106             return scores
```

```
107
108         # Compute the loss
109         loss = None
110
111         # scores is num_examples by num_classes (N, C)
112         def softmax_loss(x, y):
113             loss, dx = 0,0
114             #
    ============================================================= #
115             # START YOUR CODE HERE (BONUS QUESTION)
116             #
    ============================================================= #
117             #   Calculate the cross entropy loss after softmax output layer.
118             #   The format are provided in the notebook.
119             #   This function should return loss and dx, same as MSE loss
    function.
120             #
    ============================================================= #
121
122             pass
123
124             #
    ============================================================= #
125             # END YOUR CODE HERE
126             #
    ============================================================= #
127             return loss, dx
128
129
130         def MSE_loss(x, y):
131             loss, dx = 0,0
132             #
    ============================================================= #
133             # START YOUR CODE HERE
134             #
    ============================================================= #
135             #   This function should return loss and dx (gradients ready for
    back prop).
136             #   The loss is MSE loss between network ouput and one hot vector
    of class
137             #   labels is required for backpropogation.
138             #
    ============================================================= #
139             # Hint: Check the type and shape of x and y.
140             #       e.g. print('DEBUG:x.shape, y.shape', x.shape, y.shape)
141
142             # x is our y_pred, and y is the y_target
143
144             num_samples = x.shape[0]
145             num_attrs = x.shape[1]
146             y_target = np.zeros((num_samples, num_attrs))
147             for i in range(num_samples):
148               y_target[i][y[i]] = 1
149             diff = x - y_target
150             dx = diff / num_samples
151             loss = 0.5 * np.sum(np.square(diff)) / num_samples
152
153             #
    ============================================================= #
154             # END YOUR CODE HERE
```

```
155             #
    ============================================================ #
156             return loss, dx
157
158         # data_loss, dscore = softmax_loss(scores, y)
159         # The above line is for bonus question. If you have implemented
    softmax_loss, de-comment this line instead of MSE error.
160
161         data_loss, dscore = MSE_loss(scores, y) # "comment" this line if you
    use softmax_loss
162         # ============================================================ #
163         # START YOUR CODE HERE
164         # ============================================================ #
165         #   Calculate the regularization loss. Multiply the regularization
166         #   loss by 0.5 (in addition to the factor reg).
167         # ============================================================ #
168         reg_loss = 0.5 * reg * (np.sum(W1*W1) + np.sum(W2*W2))
169
170         # ============================================================ #
171         # END YOUR CODE HERE
172         # ============================================================ #
173         loss = data_loss + reg_loss
174
175         grads = {}
176
177         # ============================================================ #
178         # START YOUR CODE HERE
179         # ============================================================ #
180         # Backpropogation: (You do not need to change this!)
181         #   Backward pass is implemented. From the dscore error, we calculate
182         #   the gradient and store as grads['W1'], etc.
183         # ============================================================ #
184         grads['W2'] = a2.T.dot(dscore).T + reg * W2
185         grads['b2'] = np.ones(N).dot(dscore)
186
187         da_h = np.zeros(h1.shape)
188         da_h[h1>0] = 1
189         dh = (dscore.dot(W2) * da_h)
190
191         grads['W1'] = np.dot(dh.T,X) + reg * W1
192         grads['b1'] = np.ones(N).dot(dh)
193         # ============================================================ #
194         # END YOUR CODE HERE
195         # ============================================================ #
196
197         return loss, grads
198
199     def train(self, X, y, X_val, y_val,
200             learning_rate=1e-3, learning_rate_decay=0.95,
201             reg=1e-5, num_iters=100,
202             batch_size=200, verbose=False):
203         """
204         Train this neural network using stochastic gradient descent.
205
206         Inputs:
207         - X: A numpy array of shape (N, D) giving training data.
208         - y: A numpy array f shape (N,) giving training labels; y[i] = c
    means that
209           X[i] has label c, where 0 <= c < C.
210         - X_val: A numpy array of shape (N_val, D) giving validation data.
```

```
        - y_val: A numpy array of shape (N_val,) giving validation labels.
        - learning_rate: Scalar giving learning rate for optimization.
        - learning_rate_decay: Scalar giving factor used to decay the learning rate
          after each epoch.
        - reg: Scalar giving regularization strength.
        - num_iters: Number of steps to take when optimizing.
        - batch_size: Number of training examples to use per step.
        - verbose: boolean; if true print progress during optimization.
        """
        num_train = X.shape[0]
        iterations_per_epoch = max(num_train / batch_size, 1)

        # Use SGD to optimize the parameters in self.model
        loss_history = []
        train_acc_history = []
        val_acc_history = []

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

            #   Create a minibatch (X_batch, y_batch) by sampling batch_size
            #   samples randomly.

            b_index = np.random.choice(num_train, batch_size)
            X_batch = X[b_index]
            y_batch = y[b_index]

            # Compute loss and gradients using the current minibatch
            loss, grads = self.loss(X_batch, y=y_batch, reg=reg)
            loss_history.append(loss)

            #
=========================================================== #
            # START YOUR CODE HERE
            #
=========================================================== #
            #   Perform a gradient descent step using the minibatch to update
            #   all parameters (i.e., W1, W2, b1, and b2).
            #   The gradient has been calculated as grads['W1'], grads['W2'],
            #   grads['b1'], grads['b2']
            #   For example,
            #   W1(new) = W1(old) - learning_rate * grads['W1']
            #   (this is not the exact code you use!)
            #
=========================================================== #

            self.params['W1'] = self.params['W1'] - learning_rate *
grads['W1']
            self.params['b1'] = self.params['b1'] - learning_rate *
grads['b1']
            self.params['W2'] = self.params['W2'] - learning_rate *
grads['W2']
            self.params['b2'] = self.params['b2'] - learning_rate *
grads['b2']

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

```
            #
========================================================== #

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

            # Every epoch, check train and val accuracy and decay learning
rate.
            if it % iterations_per_epoch == 0:
                # Check accuracy
                train_acc = (self.predict(X_batch) == y_batch).mean()
                val_acc = (self.predict(X_val) == y_val).mean()
                train_acc_history.append(train_acc)
                val_acc_history.append(val_acc)

                # Decay learning rate
                learning_rate *= learning_rate_decay

        return {
          'loss_history': loss_history,
          'train_acc_history': train_acc_history,
          'val_acc_history': val_acc_history,
        }

    def predict(self, X):
        """
        Use the trained weights of this two-layer network to predict labels
for
        data points. For each data point we predict scores for each of the C
        classes, and assign each data point to the class with the highest
score.

        Inputs:
        - X: A numpy array of shape (N, D) giving N D-dimensional data points
to
          classify.

        Returns:
        - y_pred: A numpy array of shape (N,) giving predicted labels for
each of
          the elements of X. For all i, y_pred[i] = c means that X[i] is
predicted
          to have class c, where 0 <= c < C.
        """
        y_pred = None

        # ========================================================== #
        # START YOUR CODE HERE
        # ========================================================== #
        #   Predict the class given the input data.
        # ========================================================== #
        scores = self.loss(X)
        y_pred = np.argmax(scores,axis=1)

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

        return y_pred
```

```
314
315
316
```