HW2 Yuchen Wang

1.

(a) $\mathbb{E}_{\delta\sim N}[\tilde{L}(\theta)] = \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (x^{(i)} + \delta^{(i)})^T\theta)^2]$

$= \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (x^{(i)T}\theta + \delta^{(i)T}\theta))^2]$

$= \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}((y^{(i)} - x^{(i)T}\theta) - \delta^{(i)T}\theta)^2]$

$= \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}[(y^{(i)} - x^{(i)T}\theta)^2 - 2(y^{(i)} - x^{(i)T}\theta)(\delta^{(i)T}\theta) + (\delta^{(i)T}\theta)^2]]$

$= \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - x^{(i)T}\theta)^2 + \frac{1}{N}\sum_{i=1}^{N}[-2(y^{(i)} - x^{(i)T}\theta)(\delta^{(i)T}\theta) + (\delta^{(i)T}\theta)^2]]$

$= L(\theta) + \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}[-2(y^{(i)} - x^{(i)T}\theta)(\delta^{(i)T}\theta) + (\delta^{(i)T}\theta)^2]]$

Based on the linearity property of expectation, we can divide this into several parts

$\mathbb{E}[-2(y^{(i)} - x^{(i)T}\theta)(\delta^{(i)T}\theta)] = \mathbb{E}[(y^{(i)}\delta^{(i)T}\theta) - (x^{(i)T}\theta\delta^{(i)T}\theta)]$

$= y^{(i)}\mathbb{E}[\delta^{(i)T}]\theta - x^{(i)T}\theta\,\mathbb{E}[\delta^{(i)T}]\theta$

Since $\mathbb{E}[\delta^{(i)T}] = 0$, $\mathbb{E}[-2(y^{(i)} - x^{(i)T}\theta)(\delta^{(i)T}\theta)]$ will be 0.

$\mathbb{E}[(\delta^{(i)T}\theta)^2] = \mathbb{E}[(\delta^{(i)T}\theta)^T(\delta^{(i)T}\theta)]$

$= \mathbb{E}[\theta^T\delta^{(i)}\delta^{(i)T}\theta]$

$= \theta^T\mathbb{E}[\delta^{(i)}\delta^{(i)T}]\theta$

$= \theta^T\sigma^2 I\theta$

$= \sigma^2\theta^T\theta$

$= \sigma^2\|\theta\|_2^2$

So, we have $\mathbb{E}_{\delta\sim N}[\tilde{L}(\theta)] = L(\theta) + \mathbb{E}_{\delta\sim N}[\frac{1}{N}\sum_{i=1}^{N}[-2(y^{(i)} - x^{(i)T}\theta)(\delta^{(i)T}\theta) + (\delta^{(i)T}\theta)^2]]$

$= L(\theta) + 0 + \sigma^2\|\theta\|_2^2$

$= L(\theta) + \sigma^2\|\theta\|_2^2$

(b) Adding noise to the model can prevent overfitting. By adding noise to the training data, it forces the model to be more generalized.

(c) When $\sigma \to 0$, $\tilde{L}(\theta) = L(\theta)$, this might overfit the data since no regularization.

(d) When $\sigma \to \infty$, $\tilde{L}(\theta) = \sigma^2\|\theta\|_2^2$. When you want to minimize this loss, $\theta = 0$. No learning here, might cause underfit.

3.

First, we will calculate the log-likelihood L.

Let $\tilde{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$, $\tilde{w}_i = \begin{bmatrix} w_i \\ b_i \end{bmatrix}$, then we will have $a_i(x) = w_i^T x + b_i = \tilde{w}_i^T \tilde{x}$

So, the softmax function could be represented as $\text{softmax}_i(x) = \frac{e^{\tilde{w}_i^T \tilde{x}}}{\sum_{k=1}^{K} e^{\tilde{w}_k^T \tilde{x}}}$

The goal is to find the parameter θ as to maximize the likelihood of having seen the data. This corresponds to maximizing

$$p(x^{(1)}, \cdots x^{(m)}, y^{(1)}, \cdots, y^{(m)} | \theta) = \prod_{j=1}^{m} p(x^{(j)}, y^{(j)} | \theta)$$
$$= \prod_{j=1}^{m} p(x^{(j)} | \theta)\, p(y^{(j)} | x^{(j)}, \theta)$$

We want $\arg\max_\theta \prod_{j=1}^{m} p(x^{(j)} | \theta)\, p(y^{(j)} | x^{(j)}, \theta) = \arg\max_\theta \prod_{j=1}^{m} p(y^{(j)} | x^{(j)}, \theta)$

$$= \arg\max_\theta \sum_{j=1}^{m} \log \text{softmax}_{y^{(j)}}(x^{(j)})$$
$$= \arg\max_\theta \sum_{j=1}^{m} \log \frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}}$$

So, $L = \sum_{j=1}^{m} \log \frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}}$

Let's denote $r_j = \frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}}$ and $l_j = \log(r_j)$

For w_i,

(i) $i = j$:

$$\frac{dL}{dw_i} = \sum_{j=1}^{m} \frac{dL}{dw_i} = \sum_{j=1}^{m} \frac{dr_j}{dw_i} \cdot \frac{dL}{dr_j} = \sum_{j=1}^{m} \frac{da_j}{dw_i} \cdot \frac{dr_j}{da_j} \cdot \frac{dL}{dr_j}$$

Now, $\frac{da_j}{dw_i} = \tilde{x}^{(j)}$

$$\frac{dr_j}{da_j} = \frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}} - \left[\frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}}\right]^2 = r_j(1 - r_j)$$

$$\frac{dL}{dr_j} = \frac{1}{r_j}$$

Putting it all together, $\frac{dL}{dw_i} = \sum_{j=1}^{m} \tilde{x}^{(j)} \cdot (1 - r_j)$

$$= \sum_{j=1}^{m} \tilde{x}^{(j)} \cdot \left(1 - \frac{e^{a_j(x)}}{\sum_{k=1}^{K} e^{a_k(x)}}\right)$$

$$\frac{\partial L}{\partial b_i} = \sum_{j=1}^{N_3} \frac{\partial L_j}{\partial b_i} = \sum_{j=1}^{N_3} \frac{\partial a_j}{\partial b_i} \cdot \frac{\partial r_j}{\partial a_j} \cdot \frac{\partial L_j}{\partial r_j}$$

Now, $\frac{\partial a_j}{\partial b_i} = 1$, where $\frac{\partial r_j}{\partial a_j}$ and $\frac{\partial L_j}{\partial r_j}$ are same.

So, $\frac{\partial L}{\partial b_i} = \sum_{j=1}^{N_3} \left(1 - \frac{e^{a_j(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right)$

(ii) $i \neq j$

$$\frac{\partial L}{\partial w_i} = \sum_{j=1}^{N_3} \frac{\partial L_j}{\partial w_i} = \sum_{j=1}^{N_3} \frac{\partial r}{\partial w_i} \cdot \frac{\partial L_j}{\partial r} = \sum_{j=1}^{N_3} \frac{\partial a_j}{\partial w_i} \cdot \frac{\partial r_j}{\partial a_j} \cdot \frac{\partial L_j}{\partial r_j}$$

$$\frac{\partial r_j}{\partial a_i} = \frac{-e^{a_j(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}} \cdot \frac{e^{a_i(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}} = -r_i r_j$$

$$\frac{\partial a_j}{\partial w_i} = x^{(i)}$$

$$\frac{\partial L_j}{\partial r_j} = \frac{1}{r_j}$$

Putting it all together, $\frac{\partial L}{\partial w_i} = \sum_{j=1}^{N_3} x^{(i)}(-r_i)$

$$= \sum_{j=1}^{N_3} x^{(i)} \left(-\frac{e^{a_i(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right)$$

$$\frac{\partial L}{\partial b_i} = \sum_{j=1}^{N_3} \frac{\partial L_j}{\partial b_i} = \sum_{j=1}^{N_3} \frac{\partial a_j}{\partial b_i} \cdot \frac{\partial r_j}{\partial a_j} \cdot \frac{\partial L_j}{\partial r_j}$$

Now, $\frac{\partial a_j}{\partial b_i} = 1$

$$\frac{\partial L}{\partial b_i} = \sum_{j=1}^{N_3} \left(-\frac{e^{a_i(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right)$$

In summary,

$$\nabla_{w_i} L = \begin{cases} \sum_{j=1}^{N_3} x^{(i)} \cdot \left(1 - \frac{e^{a_j(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right), & \text{when } i = j \\ \sum_{j=1}^{N_3} x^{(i)} \left(-\frac{e^{a_i(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right), & \text{when } i \neq j \end{cases}$$

$$\nabla_{b_i} L = \begin{cases} \sum_{j=1}^{N_3} \left(1 - \frac{e^{a_j(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right), & \text{when } i = j \\ \sum_{j=1}^{N_3} \left(-\frac{e^{a_i(x)}}{\sum_{k=1}^{N_3} e^{a_k(x)}}\right), & \text{when } i \neq j \end{cases}$$

4.

① $\nabla_w L = \frac{1}{k}\sum_{i=1}^{k} \nabla_w \text{hinge}_{y^{(i)}}(x^{(i)})$

The hinge function can be represented as

$$\text{hinge}_{y^{(i)}}(x^{(i)}) = \begin{cases} 0 & , y^{(i)}(w^T x^{(i)} + b) > 1 \\ 1 - y^{(i)}(w^T x^{(i)} + b) & , \text{otherwise} \end{cases}$$

If $y^{(i)}(w^T x^{(i)} + b) > 1$, $\nabla_w \text{hinge}_{y^{(i)}}(x^{(i)}) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^d$

Otherwise, $\nabla_w \text{hinge}_{y^{(i)}}(x^{(i)}) = -y^{(i)} x^{(i)} \in \mathbb{R}^d$

Together, $\nabla_w \text{hinge}_{y^{(i)}}(x^{(i)}) = \mathbb{I}\{y^{(i)}(w^T x^{(i)} + b) < 1\} \odot (-y^{(i)} x^{(i)})$, where $\mathbb{I}\{y^{(i)}(w^T x^{(i)} + b) < 1\}$ is an all ones vector when $y^{(i)}(w^T x^{(i)} + b) < 1$ or an all zeros vector if $y^{(i)}(w^T x^{(i)} + b) < 1$.

Therefore, $\nabla_w L = \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}\{y^{(i)}(w^T x^{(i)} + b) < 1\} \odot (-y^{(i)} x^{(i)})$

② $\nabla_b L = \frac{1}{k}\sum_{i=1}^{k} \nabla_b \text{hinge}_{y^{(i)}}(x^{(i)})$

Similarly, if $y^{(i)}(w^T x^{(i)} + b) > 1$, $\nabla_b \text{hinge}_{y^{(i)}}(x^{(i)}) = 0 \in \mathbb{R}$

Otherwise, $\nabla_b \text{hinge}_{y^{(i)}}(x^{(i)}) = -y^{(i)} \in \mathbb{R}$

Together, $\nabla_b \text{hinge}_{y^{(i)}}(x^{(i)}) = \mathbb{I}\{y^{(i)}(w^T x^{(i)} + b < 1\} \odot (-y^{(i)})$.

Therefore, $\nabla_b L = \frac{1}{k}\sum_{i=1}^{k} \mathbb{I}\{y^{(i)}(w^T x^{(i)} + b < 1\} \odot (-y^{(i)})$.

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.

Import the appropriate libraries

In [21]:

```
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.

# 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
```

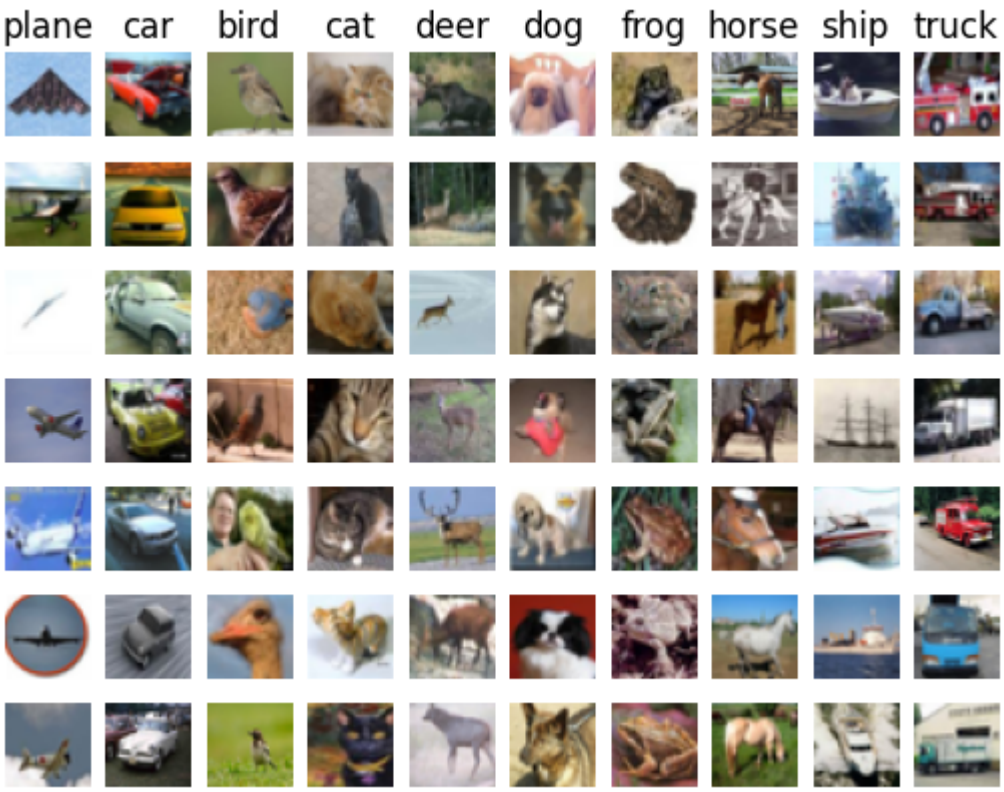
In [22]:

```
# Set the path to the CIFAR-10 data
cifar10_dir = '/Users/wangyuchen/desktop/COM SCI 247/HW/HW2/hw2_Questions/code/cifar
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,)
```

In [23]:

```
# 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', 't
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 [24]:

```
# 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)
```

K-nearest neighbors

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

In [54]:

```
# Import the KNN class

from nndl import KNN
```

In [55]:

```
# 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?

Answers

(1) This function reads in and stores the whole training dataset.

(2) Pros: This is very simple to understand and implement. Also, it is fast. Cons: This is memory-intensive since we have to memorize all pictures in this dataset.

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 [56]:

```
# Implement the function compute_distances() in the KNN class.
# Do not worry about the input 'norm' for now; use the default definition of the nor
# 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: 22.409882068634033
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

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 [57]:

```
# 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 lo
# 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): {}
```

```
Time to run code: 0.31713294982910156
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.

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 [59]:

```
# 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.
# ================================================================ #

predicatey = knn.predict_labels(dists_L2_vectorized, 1)
error = np.count_nonzero(predicatey != y_test) / len(y_test) #find the number of
                                                             #predicated y that
                                                             #is different
                                                             #from real y

# ================================================================ #
# 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%.

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.

Create training and validation folds

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

In [44]:

```
# 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]
# ================================================================ #

X_train_folds = np.array_split(X_train, num_folds)
Y_train_folds = np.array_split(y_train, num_folds)


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

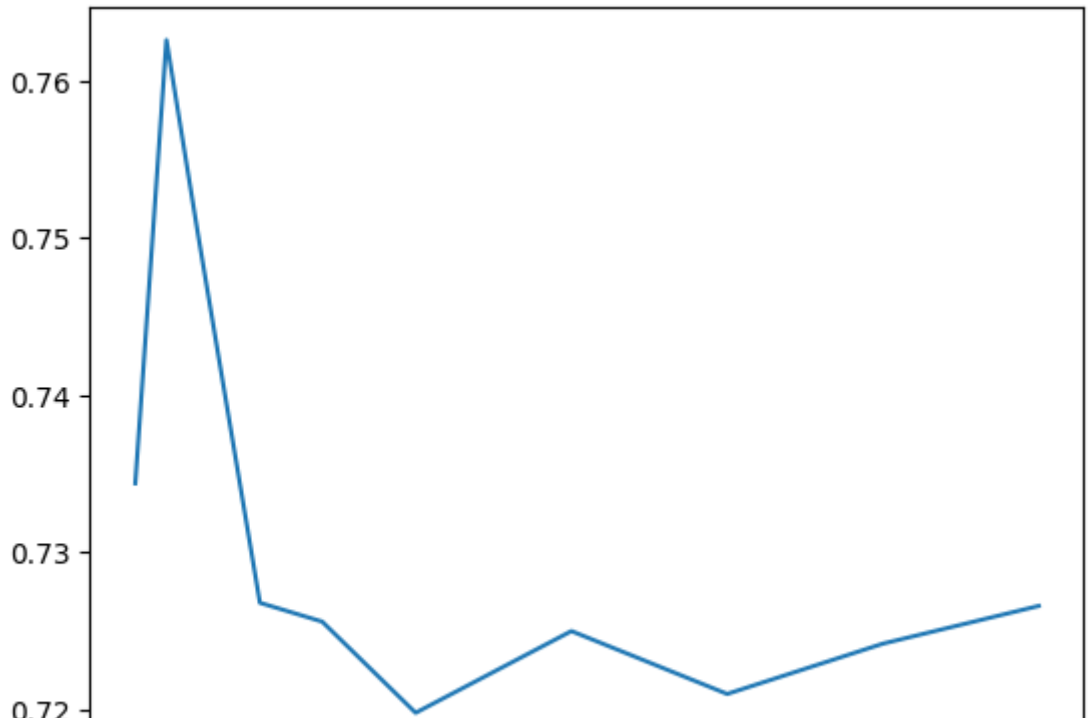
Optimizing the number of nearest neighbors hyperparameter.

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

In [47]:

```
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.
# ================================================================ #

crossv_error = []

for k in ks: #for each k

    curerror = 0

    for i in range (0, num_folds):
        xtraink = []
        ytraink = []
        xtraink = np.concatenate(X_train_folds[:i] + X_train_folds[i+1:]) #training
        ytraink = np.concatenate(Y_train_folds[:i] + Y_train_folds[i+1:]) #training
        knn.train(X=xtraink, y=ytraink) #train the model based on the training fold
        distsL2_vectorized = knn.compute_L2_distances_vectorized(X=np.array(X_train_
        predy = knn.predict_labels(distsL2_vectorized, k)
        curerror += np.count_nonzero(predy != Y_train_folds[i]) / predy.shape[0]
        aveerror = curerror / num_folds #average the error
    print(k, ":", aveerror)

    crossv_error.append(aveerror)


plt.plot(ks, crossv_error) #plot of k vs. cross-validation error
plt.show()



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

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

```
1 : 0.7344
2 : 0.7626000000000002
3 : 0.7504000000000001
5 : 0.7267999999999999
7 : 0.7256
10 : 0.7198
15 : 0.725
20 : 0.721
25 : 0.7242
30 : 0.7266
```




```
Computation time: 26.12
```

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?

Answers:

(1) 10 is the value of k which is best amongst the tested k's.

(2) The cross-validation error for k = 10 is 0.7198

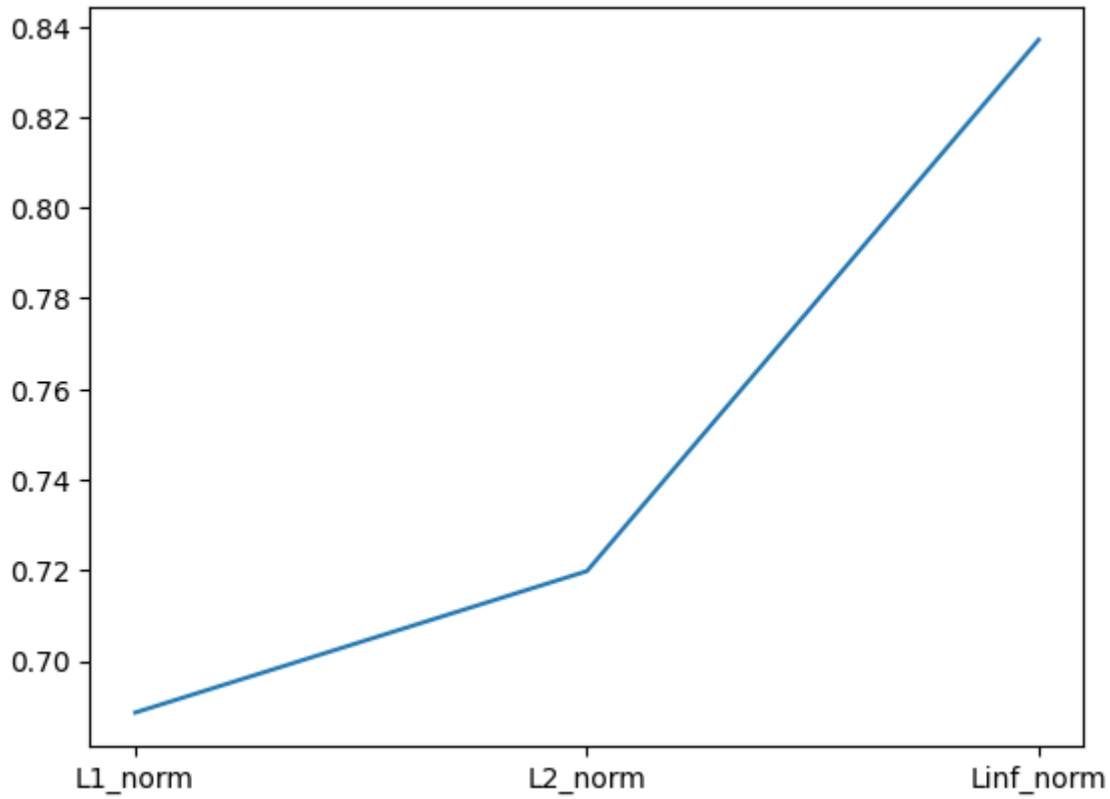
Optimizing the norm

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

In [53]:

```
time_start =time.time()

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.
# ================================================================ #

normerror = []

for n in norms: #for each norm

    curerror = 0

    for i in range (0, num_folds):
        xtraink = []
        ytraink = []
        xtraink = np.concatenate(X_train_folds[:i] + X_train_folds[i+1:]) #training
        ytraink = np.concatenate(Y_train_folds[:i] + Y_train_folds[i+1:]) #training
        knn.train(X=xtraink, y=ytraink) #train the model based on the training fold

        distsL2 = knn.compute_distances(X=np.array(X_train_folds[i]), norm = n)
        prednewy = knn.predict_labels(distsL2, 10)
        curerror += np.count_nonzero(prednewy != Y_train_folds[i]) / prednewy.shape[
        aveerror = curerror / num_folds #average the error

    print(n, ":", aveerror)

    normerror.append(aveerror)


plt.plot(["L1_norm", "L2_norm", "Linf_norm"], normerror) #plot of norms
                                                         #vs. cross-validation erro
plt.show()

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

```
<function <lambda> at 0x7fc888bfc0d0> : 0.6886000000000001
<function <lambda> at 0x7fc888bfcc10> : 0.7198
<function <lambda> at 0x7fc8bc637040> : 0.8370000000000001
```




```
Computation time: 445.57
```

Questions:

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

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

Answers:

(1) L1_norm has the best cross-validation error.

(2) Under L1_norm and k = 10, the cross-validation error is 0.6886000000000001.

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.

In [61]:

```
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.
# ================================================================ #

knn.train(X=X_train, y=y_train)
dists_now = knn.compute_distances(X=X_test, norm = L1_norm) #using the best norm
predynow = knn.predict_labels(dists_now, 10) #using the optimal k
error = np.count_nonzero(predynow != y_test) / len(y_test)

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

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

```
Error rate achieved: 0.722
```

Question:

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

Answer:

The error was 0.726 under k = 1 and the L2-norm. It has decreased by 0.004 by using L1_norm and k = 10.

```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(X[i] - self.X_train[j])

        # ================================================================ #
        # 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.
    # ========================================================== #

    dists = np.sqrt( (np.sum(X**2, axis = 1).reshape(X.shape[0], 1)) + (np.sum(self.X_train**2, axis = 1)) - (2 * X.dot(self.X_train.T)) )

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

```
92        return dists
93
94
95      def predict_labels(self, dists, k=1):
96        """
97        Given a matrix of distances between test points and training points,
98        predict a label for each test point.
99
100       Inputs:
101       - dists: A numpy array of shape (num_test, num_train) where dists[i, j]
102         gives the distance betwen the ith test point and the jth training point.
103
104       Returns:
105       - y: A numpy array of shape (num_test,) containing predicted labels for the
106         test data, where y[i] is the predicted label for the test point X[i].
107       """
108       num_test = dists.shape[0]
109       y_pred = np.zeros(num_test)
110       for i in np.arange(num_test):
111         # A list of length k storing the labels of the k nearest neighbors to
112         # the ith test point.
113         closest_y = []
114         # ========================================================================= #
115         # YOUR CODE HERE:
116         #   Use the distances to calculate and then store the labels of
117         #   the k-nearest neighbors to the ith test point.  The function
118         #   numpy.argsort may be useful.
119         #
120         #   After doing this, find the most common label of the k-nearest
121         #   neighbors.  Store the predicted label of the ith training example
```

```
        #   as y_pred[i].  Break ties by choosing the smaller label.
        # ================================================================ #

        neighborsy = sorted(self.y_train[np.argsort(dists[i])[:k]]) #get the nearest k neighbors based on the distance sort
        y_pred[i] = max(set(neighborsy), key = neighborsy.count)

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

    return y_pred
```

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.

In [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
```

In [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 = '/Users/wangyuchen/desktop/COM SCI 247/HW/HW2/hw2_Questions/code/cifar-10-batches-py' # You ne
    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,)
```

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.

In [216]:
```
from nndl import Softmax
```

In [217]:
```
# 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

In [218]:
```
## Implement the loss function of the softmax using a for loop over
#  the number of examples

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

In [219]:
```
print(loss)
```

```
2.327760702804897
```

Question:

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

Answer:

Currently, there is no regularization or trained model, and we are using the random weights. This might cause the large error.

Softmax gradient

In [220]:
```
## 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 correct.
softmax.grad_check_sparse(X_dev, y_dev, grad)
```

```
numerical: -0.969757 analytic: -0.969757, relative error: 2.962114e-08
numerical: -0.541608 analytic: -0.541608, relative error: 2.532321e-08
numerical: 0.059517 analytic: 0.059517, relative error: 2.484896e-07
numerical: 2.379901 analytic: 2.379901, relative error: 1.451935e-09
numerical: -0.471978 analytic: -0.471978, relative error: 4.772627e-08
numerical: 1.905349 analytic: 1.905349, relative error: 2.032549e-09
numerical: 0.668743 analytic: 0.668743, relative error: 1.007004e-07
numerical: -0.524710 analytic: -0.524710, relative error: 1.791734e-08
numerical: 1.508797 analytic: 1.508797, relative error: 9.542764e-10
numerical: -4.177994 analytic: -4.177995, relative error: 9.838742e-09
```

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.

In [221]:
```
import time
```

In [222]:
```
## 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,

# 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.
```

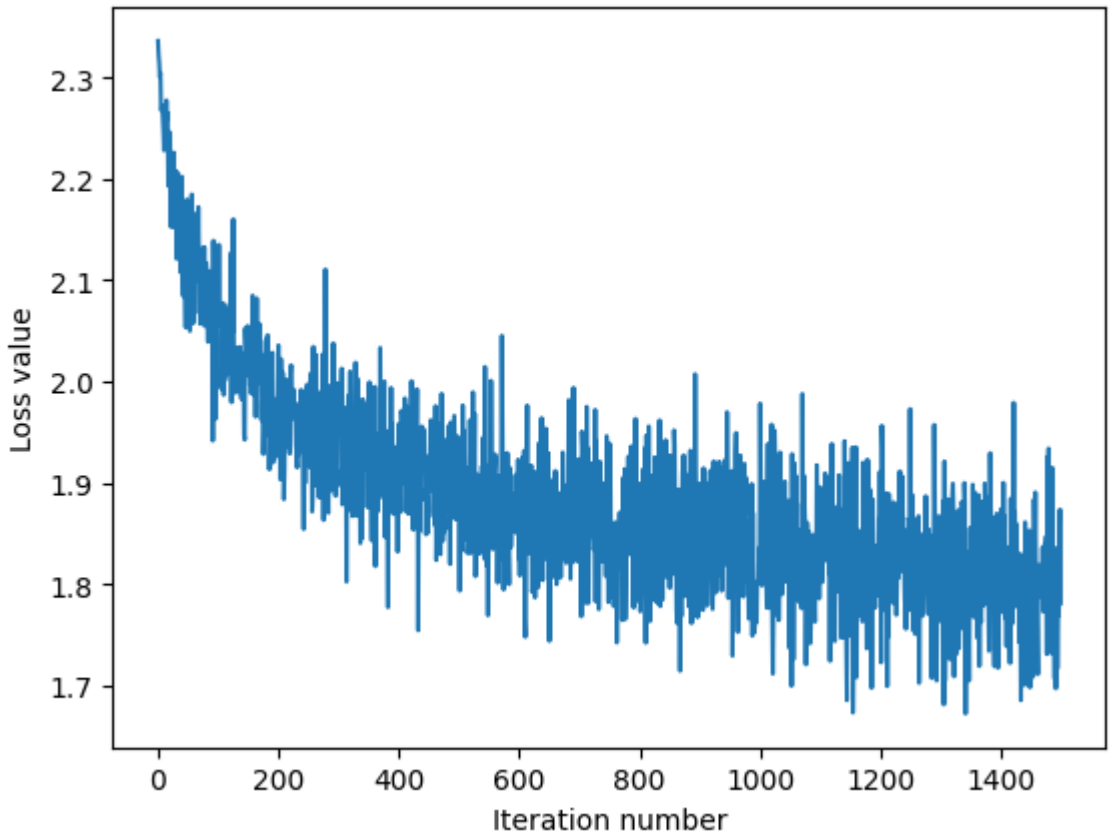
```
Normal loss / grad_norm: 2.3431081152133433 / 337.4978602358297 computed in 0.09418487548828125s
Vectorized loss / grad: 2.3431081152133424 / 337.4978602358297 computed in 0.016119003295898438s
difference in loss / grad: 8.881784197001252e-16 /2.3101392126522646e-13
```

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).

In [223]:
```
# 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 2.336592660663754
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 1.835306222372583
iteration 800 / 1500: loss 1.8293892468827635
iteration 900 / 1500: loss 1.8992158530357484
iteration 1000 / 1500: loss 1.9783503540252299
iteration 1100 / 1500: loss 1.8470797913532635
iteration 1200 / 1500: loss 1.8411450268664082
iteration 1300 / 1500: loss 1.7910402495792102
iteration 1400 / 1500: loss 1.8705803029382257
That took 5.400985240936279s
```



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

In [211]:
```
## 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
```

Optimize the softmax classifier

In [212]:
```
np.finfo(float).eps
```

Out[212]:
```
2.220446049250313e-16
```

In [225]:

```
# ================================================================ #
# 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.
# ================================================================ #

rates = [1e-7, 1e-6, 1e-5, 1e-4]
for learningrate in rates: #for each learning rate, run the train function
    print("learning rate:", learningrate)
    loss_histnew = softmax.train(X_train, y_train, learning_rate=learningrate,
                          num_iters=1500, verbose=True)
    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)), ))

softmax.train(X_train, y_train, learning_rate=1e-6, num_iters=1500)
y_test_pred = softmax.predict(X_test)
test_accuracy = np.mean(np.equal(y_test, y_test_pred))
errorrate = 1 - test_accuracy
print("error rate:", errorrate)

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

```
learning rate: 1e-07
iteration 0 / 1500: loss 2.352610559430411
iteration 100 / 1500: loss 2.0532315368635734
iteration 200 / 1500: loss 2.0250842852711877
iteration 300 / 1500: loss 1.8768439366967846
```

```
iteration 1300 / 1500: loss 1.9240746212608146
iteration 1400 / 1500: loss 1.7846918635831293
training accuracy: 0.37881632653061226
validation accuracy: 0.39
learning rate: 1e-06
iteration 0 / 1500: loss 2.4615346985497166
iteration 100 / 1500: loss 1.7515308294429426
iteration 200 / 1500: loss 1.8653151657870888
iteration 300 / 1500: loss 1.7068279724663449
iteration 400 / 1500: loss 1.6919412980959523
iteration 500 / 1500: loss 1.7445602534086055
iteration 600 / 1500: loss 1.9070927441191992
iteration 700 / 1500: loss 1.6266282009657487
iteration 800 / 1500: loss 1.7137070023201975
iteration 900 / 1500: loss 1.6755856266246776
iteration 1000 / 1500: loss 1.8076485575084922
iteration 1100 / 1500: loss 1.7256661402410827
iteration 1200 / 1500: loss 1.698554182531272
iteration 1300 / 1500: loss 1.792082948906467
iteration 1400 / 1500: loss 1.6508555418428208
training accuracy: 0.42051020408163264
validation accuracy: 0.415
learning rate: 1e-05
iteration 0 / 1500: loss 2.3790388831757823
iteration 100 / 1500: loss 2.3921785131993487
iteration 200 / 1500: loss 3.271633722357008
iteration 300 / 1500: loss 2.5266749618188404
iteration 400 / 1500: loss 2.51713517924201
iteration 500 / 1500: loss 2.934688054749696
iteration 600 / 1500: loss 1.9232171018196902
iteration 700 / 1500: loss 3.4421089096587862
iteration 800 / 1500: loss 2.2737674945475397
iteration 900 / 1500: loss 2.7452588111633127
iteration 1000 / 1500: loss 2.5895477769472315
iteration 1100 / 1500: loss 2.737909522402885
iteration 1200 / 1500: loss 2.7475945671971704
iteration 1300 / 1500: loss 2.573128824801604
iteration 1400 / 1500: loss 3.1635751023691183
training accuracy: 0.28981632653061223
validation accuracy: 0.27
learning rate: 0.0001
iteration 0 / 1500: loss 2.338799247622096
```

```
learning rate: 0.0001
iteration 0 / 1500: loss 2.370051472543467
iteration 100 / 1500: loss 24.730825037678006
iteration 200 / 1500: loss 35.46560148428093
iteration 300 / 1500: loss 20.31132010492675
iteration 400 / 1500: loss 24.54293285481195
iteration 500 / 1500: loss 21.602313494150074
iteration 600 / 1500: loss 19.500515454948275
iteration 700 / 1500: loss 28.42739725098191
iteration 800 / 1500: loss 32.797689386402496
iteration 900 / 1500: loss 29.220960515114186
iteration 1000 / 1500: loss 32.70864966580386
iteration 1100 / 1500: loss 21.57436396682263
iteration 1200 / 1500: loss 27.15909944628512
iteration 1300 / 1500: loss 26.318795208646534
iteration 1400 / 1500: loss 30.874349296588235
training accuracy: 0.2852448979591837
validation accuracy: 0.3
error rate: 0.603
```

Report:

The best learning rate is 1e-06, where the training accuracy: 0.42051020408163264 and validation accuracy: 0.415. And the error rate is 0.603.

```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:
```

```
    # ============================================================= #
    # 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.)
    # ============================================================= #

    row = X.shape[0]
    for i in range (0, row):
        loss += np.log(np.sum(np.exp(X[i].dot(self.W.transpose())))) - X[i].dot(self.W[y[i]]) #follow the softmax loss function

    loss = loss / row

    # ============================================================= #
    # 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.
    """

    # 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.
    # ============================================================= #
```

```
    row = X.shape[0]
    for i in range (0, row): #for each training sample

        loss += np.log(np.sum(np.exp(X[i].dot(self.W.transpose())))) - X[i].dot(self.W[y[i]]) #calculate the loss

        sumexp = (np.sum(np.exp(X[i].dot(self.W.transpose()))))


        for j in range (0, 10): #for each class

            ai = X[i].dot(self.W[j])
            expai = np.exp(ai)

            if j == y[i]: #when the index is the "correct class" index for the sample

                grad[j] += X[i] * (-1 + expai / sumexp)

            else:

                grad[j] += X[i] * expai / sumexp

    loss = loss / row
    grad = grad / row


    # ============================================================ #
    # 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.
    # ============================================================= #

    #calculate the loss using the softmax loss function
    row = X.shape[0]

    loss = np.sum( np.log( np.sum( np.exp(X.dot(self.W.transpose())), axis = 1 ) ) - np.sum(self.W[y] * X, axis = 1) )

    loss = loss / row


```

```
    #calculate the gradient
    avalue = np.dot(self.W, X.transpose())
    avalue -= np.max(avalue, axis = 0, keepdims = True)
    expavalue = np.exp(avalue)
    softmaxfunction = expavalue / np.sum(expavalue, axis=0, keepdims=True)
    softmaxfunction[y, range(row)] -= 1
    grad = np.dot(softmaxfunction, X)

    grad /= row

    # ============================================================= #
    # 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

```

```
184     self.init_weights(dims=[np.max(y) + 1, X.shape[1]])# initializes the weights of self.W
185
186     # Run stochastic gradient descent to optimize W
187     loss_history = []
188
189     for it in np.arange(num_iters):
190       X_batch = None
191       y_batch = None
192
193       # ========================================================== #
194       # YOUR CODE HERE:
195       #   Sample batch_size elements from the training data for use in
196       #     gradient descent.  After sampling,
197       #     - X_batch should have shape: (batch_size, dim)
198       #     - y_batch should have shape: (batch_size,)
199       #   The indices should be randomly generated to reduce correlations
200       #   in the dataset.  Use np.random.choice.  It's okay to sample with
201       #   replacement.
202       # ========================================================== #
203
204       randomindices = np.random.choice(np.arange(num_train), batch_size)
205       X_batch = X[randomindices]
206       y_batch = y[randomindices]
207
208
209       # ========================================================== #
210       # END YOUR CODE HERE
211       # ========================================================== #
212
213       # evaluate loss and gradient
214       loss, grad = self.fast_loss_and_grad(X_batch, y_batch)
215       loss_history.append(loss)
216
217       # ========================================================== #
218       # YOUR CODE HERE:
219       #   Update the parameters, self.W, with a gradient step
220       # ========================================================== #
```

```
      self.W -= grad * learning_rate

      # ========================================================== #
      # 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.
    # ========================================================== #

    y_pred = np.argmax(X.dot(self.W.transpose()), axis = 1) #we want the maximum value for each row

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

    return y_pred
```