k-Nearest Neighbor (kNN) exercise

Complete and hand in this completed worksheet (including its outputs and any supporting code outside of the worksheet) with your assignment submission. For more details see the [assignments page (https://compsci682-fa19.github.io/assignments2019/assignment1)](https://compsci682-fa19.github.io/assignments2019/assignment1) on the course website.

The kNN classifier consists of two stages:

- During training, the classifier takes the training data and simply remembers it
- During testing, kNN classifies every test image by comparing to all training images and transfering the labels of the k most similar training examples
- The value of k is cross-validated

In this exercise you will implement these steps and understand the basic Image Classification pipeline, cross-validation, and gain proficiency in writing efficient, vectorized code.

In [1]:
```
# Run some setup code for this notebook.
from __future__ import print_function

import random
import numpy as np
from cs682.data_utils import load_CIFAR10
import matplotlib.pyplot as plt


# This is a bit of magic to make matplotlib figures appear inline in t
he notebook
# rather than in a new window.
%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plo
ts
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# Some more magic so that the notebook will reload external python mod
ules;
# see http://stackoverflow.com/questions/1907993/autoreload-of-modules
-in-ipython
%load_ext autoreload
%autoreload 2
```

In [2]:
```
# Load the raw CIFAR-10 data.
cifar10_dir = 'cs682/datasets/cifar-10-batches-py'

# Cleaning up variables to prevent loading data multiple times (which
may cause memory issue)
try:
   del X_train, y_train
   del X_test, y_test
   print('Clear previously loaded data.')
except:
   pass

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 da
ta.
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 [3]:

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

In [4]:

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

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

In [5]:
```
# 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)
```

In [6]:
```
from cs682.classifiers import KNearestNeighbor

# Create a kNN classifier instance.
# Remember that training a kNN classifier is a noop:
# the Classifier simply remembers the data and does no further process
ing
classifier = KNearestNeighbor()
classifier.train(X_train, y_train)
```

We would now like to classify the test data with the kNN classifier. Recall that we can break down this process into two steps:

1. First we must compute the distances between all test examples and all train examples.
2. Given these distances, for each test example we find the k nearest examples and have them vote for the label

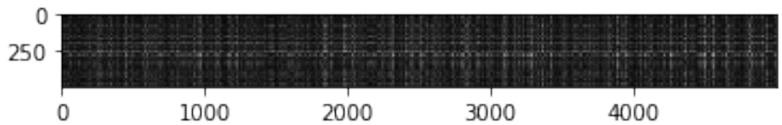
Lets begin with computing the distance matrix between all training and test examples. For example, if there are **Ntr** training examples and **Nte** test examples, this stage should result in a **Nte x Ntr** matrix where each element (i,j) is the distance between the i-th test and j-th train example.

First, open `cs682/classifiers/k_nearest_neighbor.py` and implement the function `compute_distances_two_loops` that uses a (very inefficient) double loop over all pairs of (test, train) examples and computes the distance matrix one element at a time.

In [7]:
```
# Open cs682/classifiers/k_nearest_neighbor.py and implement
# compute_distances_two_loops.

# Test your implementation:
dists = classifier.compute_distances_two_loops(X_test)
print(dists.shape)
```

```
(500, 5000)
```

In [8]:
```
# We can visualize the distance matrix: each row is a single test exam
ple and
# its distances to training examples
plt.imshow(dists, interpolation='none')
plt.show()
```



Inline Question #1: Notice the structured patterns in the distance matrix, where some rows or columns are visible brighter. (Note that with the default color scheme black indicates low distances while white indicates high distances.)

- What in the data is the cause behind the distinctly bright rows?
- What causes the columns?

Your Answer: *fill this in.*

- first question:

```
 a bright row indicates that one test picture is a outlier in training
pictures, maybe it gets some special background color. The prediction
of this picture will not be accurate.
```

- second question:

```
 a bright column indicates that one training picture does not match wi
th test pictures. Maybe, this training picture is a outlier, so that w
e can remove it from training data. Another possibility is that there
are few pictures in test set fit in the category of this training pict
ure.
```

In [9]:
```
# Now implement the function predict_labels and run the code below:
# We use k = 1 (which is Nearest Neighbor).
y_test_pred = classifier.predict_labels(dists, k=1)

# Compute and print the fraction of correctly predicted examples
num_correct = np.sum(y_test_pred == y_test)
accuracy = float(num_correct) / num_test
print('Got %d / %d correct => accuracy: %f' % (num_correct, num_test,
accuracy))
```

```
Got 137 / 500 correct => accuracy: 0.274000
```

You should expect to see approximately `27%` accuracy. Now lets try out a larger `k`, say `k = 5`:

In [10]:
```
y_test_pred = classifier.predict_labels(dists, k=5)
num_correct = np.sum(y_test_pred == y_test)
accuracy = float(num_correct) / num_test
print('Got %d / %d correct => accuracy: %f' % (num_correct, num_test,
accuracy))
```

```
Got 139 / 500 correct => accuracy: 0.278000
```

You should expect to see a slightly better performance than with `k = 1`.

Inline Question 2 We can also other distance metrics such as L1 distance. The performance of a Nearest Neighbor classifier that uses L1 distance will not change if (Select all that apply.):

1. The data is preprocessed by subtracting the mean.
2. The data is preprocessed by subtracting the mean and dividing by the standard deviation.
3. The coordinate axes for the data are rotated.
4. None of the above.

Your Answer: 1,2 *Your explanation*: L1=||y1-x1||+||y2-x2||... in 1, L1_1_2=||(x1-mean_x) - (x2-mean_x)||=||x1-x2|| in 2, L1_1_2=||(x1-mean_x)/varient - (x2-mean_x)/varient||=||x1-x2||/varient so, the ordering of distences will not change in 3, if coordinate axes rotate alfa degrees, then x_prime=x*cos(alfa)*+y*sin(alfa), y_prime=y*cos(alfa)*-x*sin(alfa) so that , if we rotate 30 degrees, (1,0)--> (√3/2,-1/2) (0,1)-->(1/2,√3/2) (1,2)--> (√3/2+1,√3-1/2) ||(1,2)-(1,0)||=2 ||(1,2)-(0,1)||=2 ||(√3/2+1,√3-1/2)-(√3/2,-1/2)||=1+√3 ||(√3/2+1,√3-1/2)-(1/2,√3/2)||=√3/2+1/2+√3/2-1/2=√3 so, the distance relationship changes

In [11]:
```
# Now lets speed up distance matrix computation by using partial vecto
rization
# with one loop. Implement the function compute_distances_one_loop and
run the
# code below:
dists_one = classifier.compute_distances_one_loop(X_test)

# To ensure that our vectorized implementation is correct, we make sur
e that it
# agrees with the naive implementation. There are many ways to decide
whether
# two matrices are similar; one of the simplest is the Frobenius norm.
In case
# you haven't seen it before, the Frobenius norm of two matrices is th
e square
# root of the squared sum of differences of all elements; in other wor
ds, reshape
# the matrices into vectors and compute the Euclidean distance between
them.
difference = np.linalg.norm(dists - dists_one, ord='fro')
print('Difference was: %f' % (difference, ))
if difference < 0.001:
    print('Good! The distance matrices are the same')
else:
    print('Uh-oh! The distance matrices are different')
```

```
Difference was: 0.000000
Good! The distance matrices are the same
```

In [12]:
```
# Now implement the fully vectorized version inside compute_distances_
no_loops
# and run the code
dists_two = classifier.compute_distances_no_loops(X_test)

# check that the distance matrix agrees with the one we computed befor
e:
difference = np.linalg.norm(dists - dists_two, ord='fro')
print('Difference was: %f' % (difference, ))
if difference < 0.001:
    print('Good! The distance matrices are the same')
else:
    print('Uh-oh! The distance matrices are different')
```

```
Difference was: 0.000000
Good! The distance matrices are the same
```

In [13]:
```
# Let's compare how fast the implementations are
def time_function(f, *args):
    """
    Call a function f with args and return the time (in seconds) that
it took to execute.
    """
    import time
    tic = time.time()
    f(*args)
    toc = time.time()
    return toc - tic

two_loop_time = time_function(classifier.compute_distances_two_loops,
X_test)
print('Two loop version took %f seconds' % two_loop_time)

one_loop_time = time_function(classifier.compute_distances_one_loop, X
_test)
print('One loop version took %f seconds' % one_loop_time)

no_loop_time = time_function(classifier.compute_distances_no_loops, X_
test)
print('No loop version took %f seconds' % no_loop_time)

# you should see significantly faster performance with the fully vecto
rized implementation
```

```
Two loop version took 31.885908 seconds
One loop version took 39.287640 seconds
No loop version took 0.169557 seconds
```

Cross-validation

We have implemented the k-Nearest Neighbor classifier but we set the value k = 5 arbitrarily. We will now determine the best value of this hyperparameter with cross-validation.

In [14]:
```
num_folds = 5
k_choices = [1, 3, 5, 8, 10, 12, 15, 20, 50, 100]

X_train_folds = []
y_train_folds = []
######################################################################
##########
# TODO:
#
# Split up the training data into folds. After splitting, X_train_fold
s and    #
```

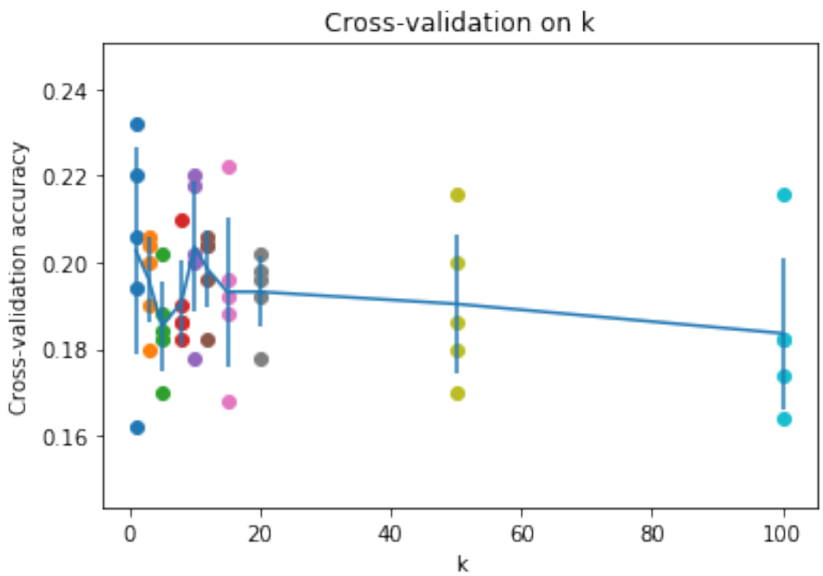
```
# y_train_folds should each be lists of length num_folds, where
#
# y_train_folds[i] is the label vector for the points in X_train_folds
[i].     #
# Hint: Look up the numpy array_split function.
#
########################################################################
##########
X_train_folds = np.split(X_train,num_folds)
y_train_folds = np.split(y_train,num_folds)
########################################################################
##########
#                                 END OF YOUR CODE
#
########################################################################
##########

# A dictionary holding the accuracies for different values of k that w
e find
# when running cross-validation. After running cross-validation,
# k_to_accuracies[k] should be a list of length num_folds giving the d
ifferent
# accuracy values that we found when using that value of k.
k_to_accuracies = {}


########################################################################
##########
# TODO:
#
# Perform k-fold cross validation to find the best value of k. For eac
h        #
# possible value of k, run the k-nearest-neighbor algorithm num_folds
times,   #
# where in each case you use all but one of the folds as training data
and the #
# last fold as a validation set. Store the accuracies for all fold and
all     #
# values of k in the k_to_accuracies dictionary.
#
########################################################################
##########
for possible_k in k_choices:
    for i in range(num_folds):
        classifier = KNearestNeighbor()
        index=list(range(num_folds))
        index.remove(i)
        train=np.array(X_train_folds)[index].reshape(-1 ,X_train.shape
[1])
        classifier.train(train, np.array(y_train_folds)[index,:].resha
```

```
pe(-1))
        y_test_pred = classifier.predict_labels(X_train_folds[i], k=po
ssible_k)
        num_correct = np.sum(y_test_pred == y_train_folds[i])
        accuracy = float(num_correct) / num_test
        if k_to_accuracies.get(possible_k) is None:
            k_to_accuracies[possible_k]=[]
        k_to_accuracies[possible_k].append(accuracy)
######################################################################
##########
#                                 END OF YOUR CODE
#
######################################################################
##########

# Print out the computed accuracies
for k in sorted(k_to_accuracies):
    for accuracy in k_to_accuracies[k]:
        print('k = %d, accuracy = %f' % (k, accuracy))
```

```
k = 1, accuracy = 0.220000
k = 1, accuracy = 0.162000
k = 1, accuracy = 0.194000
k = 1, accuracy = 0.206000
k = 1, accuracy = 0.232000
k = 3, accuracy = 0.204000
k = 3, accuracy = 0.180000
k = 3, accuracy = 0.200000
k = 3, accuracy = 0.206000
k = 3, accuracy = 0.190000
k = 5, accuracy = 0.202000
k = 5, accuracy = 0.170000
k = 5, accuracy = 0.184000
k = 5, accuracy = 0.188000
k = 5, accuracy = 0.182000
k = 8, accuracy = 0.186000
k = 8, accuracy = 0.210000
k = 8, accuracy = 0.186000
k = 8, accuracy = 0.182000
k = 8, accuracy = 0.190000
k = 10, accuracy = 0.178000
k = 10, accuracy = 0.220000
k = 10, accuracy = 0.218000
k = 10, accuracy = 0.202000
k = 10, accuracy = 0.200000
k = 12, accuracy = 0.196000
k = 12, accuracy = 0.182000
k = 12, accuracy = 0.204000
k = 12, accuracy = 0.204000
k = 12, accuracy = 0.206000
```

```
k = 15, accuracy = 0.192000
k = 15, accuracy = 0.168000
k = 15, accuracy = 0.196000
k = 15, accuracy = 0.188000
k = 15, accuracy = 0.222000
k = 20, accuracy = 0.192000
k = 20, accuracy = 0.178000
k = 20, accuracy = 0.198000
k = 20, accuracy = 0.196000
k = 20, accuracy = 0.202000
k = 50, accuracy = 0.170000
k = 50, accuracy = 0.216000
k = 50, accuracy = 0.180000
k = 50, accuracy = 0.200000
k = 50, accuracy = 0.186000
k = 100, accuracy = 0.182000
k = 100, accuracy = 0.182000
k = 100, accuracy = 0.174000
k = 100, accuracy = 0.216000
k = 100, accuracy = 0.164000
```

In [15]:
```
# plot the raw observations
for k in k_choices:
    accuracies = k_to_accuracies[k]
    plt.scatter([k] * len(accuracies), accuracies)

# plot the trend line with error bars that correspond to standard devi
ation
accuracies_mean = np.array([np.mean(v) for k,v in sorted(k_to_accuraci
es.items())])
accuracies_std = np.array([np.std(v) for k,v in sorted(k_to_accuracies
.items())])
plt.errorbar(k_choices, accuracies_mean, yerr=accuracies_std)
plt.title('Cross-validation on k')
plt.xlabel('k')
plt.ylabel('Cross-validation accuracy')
plt.show()
```



In [16]:

```
# Based on the cross-validation results above, choose the best value f
or k,
# retrain the classifier using all the training data, and test it on t
he test
# data. You should be able to get above 28% accuracy on the test data.
mean_accuracy = {}
for k in k_choices:
    accuracy_list = k_to_accuracies.get(k)
    mean_accuracy[k] = sum(accuracy_list) / len(accuracy_list)

best_k = max(mean_accuracy, key=mean_accuracy.get)

classifier = KNearestNeighbor()
classifier.train(X_train, y_train)
y_test_pred = classifier.predict(X_test, k=best_k)

# Compute and display the accuracy
num_correct = np.sum(y_test_pred == y_test)
accuracy = float(num_correct) / num_test
print('Got %d / %d correct => accuracy: %f' % (num_correct, num_test,
accuracy))
```

```
Got 141 / 500 correct => accuracy: 0.282000
```

Inline Question 3 Which of the following statements about k-Nearest Neighbor (k-NN) are true in a classification setting, and for all k? Select all that apply.

1. The training error of a 1-NN will always be better than that of 5-NN.
2. The test error of a 1-NN will always be better than that of a 5-NN.
3. The decision boundary of the k-NN classifier is linear.
4. The time needed to classify a test example with the k-NN classifier grows with the size of the training set.
5. None of the above.

Your Answer: 1, 4 *Your explanation*:

```
1. I think training errors of 1-NN is 0. since if test picture is in taini
ng ,  then the most most close picture is always it self. But in 5-NN, the
re may be other pictures from other category close to this test picture ,
and they are majorities of the most close 5 pictures.
2. In our this assignment, we can see that sometimes, 5-NN is better than
1-NN
3. the boundary can be any shape i guess.
4. a test example need to compare with all the training set , so the time
costs will grow with the size of the training set.
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: