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 the notebook
# rather than in a new window.
%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# Some more magic so that the notebook will reload external python modules;
# 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 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 [3]:

```python
# 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]:

```python
# 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]:

```python
# 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]:

```python
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 processing
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.

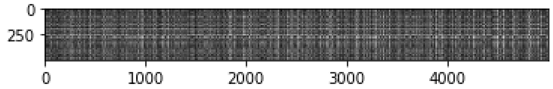
In [7]:

```
# 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 example 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:

1. The distinctly bright rows are test examples which are outliers with respect to the training set distribution - very far (having large l2 distances) from all training examples.
2. The distinctly bright columns are training examples which are outliers with respect to the test set distributions - very far (having large l2 distances) from all the test examples.

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. and 2.

Your explanation: \ In option 1, we add a constant (-mean) to data which would get cancelled out while computing the L1 distances \ L1' = x1-m - (x2-m) = x1-x2 = L1.

In option 2, we add a constant (-mean) as well as multiply by a constant (1/std_dev). The mean gets cancelled out while computing the L1 distance while all the distances get scaled by a factor of (1/std_dev). This affects all the distances in the same manner and does not affect their ordering which means we get the same set of nearest neighbors.\ L1' = (x1-m)/std_dev - (x2-m)/std_dev = (x1-x2)/std_dev = L1/std_dev

For option 3, lets say our points are (r1 cos(a), r1 sin(a)) and (r2 cos(b), r2 sin(b)). If we rotate the axes by d, the points become (r1 cos(a-d), r1 sin(a-d)) and (r2 cos(b-d), r2 sin(b-d))\ L1' = r1 (cos(a)cos(d) + sin(a)sin(d)) - r2 (cos(b)cos(d) + sin(b)sin(d)) + r1 (sin(a)cos(d) - cos(a)sin(d)) - r2 (sin(b)cos(d) - cos(b)sin(d))\ = (cos(d) - sin(d)) (r1 cos(a) - r2 cos(b) + r1 sin(a) - r2 sin(b)) + 2*r1 sin(a)sin(d) - 2*r2 sin(b)sin(d)\ = (cos(d) - sin(d))*L1 + 2*sin(d)*(r1 sin(a) - s2 sin(b))\ The above shows the L1 distances would change by a variable value which would affect the relative distances and the nearest k neighbors need not remain the same.

In [11]:

```
# Now lets speed up distance matrix computation by using partial vectorization
# 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 sure 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 the square
# root of the squared sum of differences of all elements; in other words, 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 before:
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 vectorized implementation
```

```
Two loop version took 30.059655 seconds
One loop version took 33.482554 seconds
No loop version took 0.173258 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.

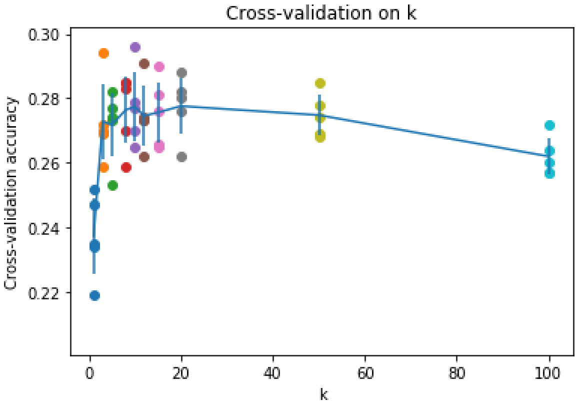
In [14]:

```
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_folds 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.array_split(X_train, num_folds)
# print(np.array(X_train_folds).shape)
y_train_folds = np.array_split(y_train, num_folds)
# print(np.array(y_train_folds).shape)
################################################################################
#                                 END OF YOUR CODE                             #
################################################################################

# A dictionary holding the accuracies for different values of k that we find
# when running cross-validation. After running cross-validation,
# k_to_accuracies[k] should be a list of length num_folds giving the different
# 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 each        #
# 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 k in k_choices:
#     print (k, "\n")
    accuracies = []
    for validation_fold in range(num_folds):
#         print(validation_fold, "\n")
        num_validate = X_train_folds[validation_fold].shape[0];
        # Select all values from folds where the idx != validation_fold
        X_train_cv = np.array([X_fold for idx,X_fold in enumerate(X_train_folds) if idx!=validation_fo
ld])
        X_train_cv = X_train_cv.reshape(-1, X_train_cv.shape[-1])
        y_train_cv = np.array([y_fold for idx,y_fold in enumerate(y_train_folds) if idx!=validation_fo
ld])
        y_train_cv = y_train_cv.flatten()
#         print(X_train_cv.shape)
#         print(y_train_cv.shape)
        classifier.train(X_train_cv, y_train_cv)

        dists = classifier.compute_distances_no_loops(X_train_folds[validation_fold])
        y_valdiate_pred = classifier.predict_labels(dists, k=k+1)
        num_correct = np.sum(y_valdiate_pred == y_train_folds[validation_fold])
        accuracy = float(num_correct) / num_validate
        accuracies.append(accuracy)
    k_to_accuracies[k] = accuracies
################################################################################
#                                 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 = 0.235000
k = 1, accuracy = 0.219000
k = 1, accuracy = 0.234000
k = 1, accuracy = 0.247000
k = 1, accuracy = 0.252000
k = 3, accuracy = 0.259000
k = 3, accuracy = 0.270000
k = 3, accuracy = 0.269000
k = 3, accuracy = 0.294000
k = 3, accuracy = 0.272000
k = 5, accuracy = 0.253000
k = 5, accuracy = 0.277000
k = 5, accuracy = 0.274000
k = 5, accuracy = 0.273000
k = 5, accuracy = 0.282000
k = 8, accuracy = 0.259000
k = 8, accuracy = 0.283000
k = 8, accuracy = 0.270000
k = 8, accuracy = 0.285000
k = 8, accuracy = 0.285000
k = 10, accuracy = 0.265000
k = 10, accuracy = 0.296000
k = 10, accuracy = 0.277000
k = 10, accuracy = 0.279000
k = 10, accuracy = 0.270000
k = 12, accuracy = 0.262000
k = 12, accuracy = 0.291000
k = 12, accuracy = 0.273000
k = 12, accuracy = 0.274000
k = 12, accuracy = 0.273000
k = 15, accuracy = 0.265000
k = 15, accuracy = 0.276000
k = 15, accuracy = 0.290000
k = 15, accuracy = 0.281000
k = 15, accuracy = 0.266000
k = 20, accuracy = 0.262000
k = 20, accuracy = 0.280000
k = 20, accuracy = 0.282000
k = 20, accuracy = 0.276000
k = 20, accuracy = 0.288000
k = 50, accuracy = 0.274000
k = 50, accuracy = 0.285000
k = 50, accuracy = 0.278000
k = 50, accuracy = 0.269000
k = 50, accuracy = 0.268000
k = 100, accuracy = 0.257000
k = 100, accuracy = 0.272000
k = 100, accuracy = 0.264000
k = 100, accuracy = 0.257000
k = 100, accuracy = 0.260000
```

In [15]:

```python
# 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 deviation
accuracies_mean = np.array([np.mean(v) for k,v in sorted(k_to_accuracies.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 [21]:

```python
# Based on the cross-validation results above, choose the best value for k,
# retrain the classifier using all the training data, and test it on the test
# data. You should be able to get above 28% accuracy on the test data.
best_k = 6

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: 4

Your explanation:\ Option 1 is incorrect because each of the training examples in a 1-NN classifier would have the 0 distance to themselves and hence would always get correctly classified. Thus the training error for a 1-NN classifier would be 0. Depending on the distribution of the data, a 5-NN classifier could also achieve 0 training error if the data from different classes is very well separated such that the 5 closest points to a data point are always from the same class as the data point.

Option 2 is incorrect because the test error is a function of the distribution of test and train datasets and the value of k doesn't necessarily imply better or worse performance.

Option 3 is incorrect because a k-NN classifier could have a non-linear decision boundary. For example, consider a data set consisting of data points of the same class distributed in concentric circles. The decision boundary would also be a collection of concentric circles if the data is well separated.

Option 4 is correct as the classification step requires computing distances from all points in the dataset and then computing the closest k points. Thus, the time to classify is proportional to the size of the training set.

In [ ]: