**CS 189 HW 1**

Students Who Helped: Suhrid Saha

*“I certify that all solutions are entirely in my own words and that I have not looked at*

*another student’s solutions. I have given credit to all external sources I consulted.”*

Sign: Shrihan Agarwal

**Q1.**

**def load\_and\_save(filename, num = None, percent = None):**

**dataset = io.loadmat("data/" + filename)**

**if percent:**

**num = int(dataset["training\_labels"].shape[0] \* percent)**

**t\_dat, t\_lbl, v\_dat, v\_lbl = split\_data(dataset["training\_data"],**

**dataset["training\_labels"],**

**num)**

**dataset["training\_data"] = t\_dat**

**dataset["training\_labels"] = t\_lbl**

**dataset["valid\_data"] = v\_dat**

**dataset["valid\_labels"] = v\_lbl**

**io.savemat("data/prep\_" + filename, dataset)**

**def split\_data(train, labels, num\_valid):**

**num\_data = train.shape[0]**

**assert num\_valid <= len(train)**

**assert num\_data == labels.shape[0]**

**idx = np.arange(num\_data)**

**np.random.shuffle(idx)**

**train\_shf = train[idx]**

**lbl\_shf = labels[idx]**

**valid\_dat = train\_shf[:num\_valid]**

**valid\_lbl = lbl\_shf[:num\_valid]**

**train\_dat = train\_shf[num\_valid:]**

**train\_lbl = lbl\_shf[num\_valid:]**

**return train\_dat, train\_lbl, valid\_dat, valid\_lbl**

**load\_and\_save("mnist\_data.mat", num = 10000)**

**load\_and\_save("spam\_data.mat", percent = 0.2)**

**load\_and\_save("cifar10\_data.mat", num = 5000)**

This section of my code not only shuffles and splits the data, but also saves the data into a new .mat file for easy access. I use this later on in the code to streamline access to the modified data.

Chart, line chart

Description automatically generated

**Q2.**

**(a)**

As expected, validation accuracies are between 70-90%. The training accuracy drops with time as the linear svm has to accommodate more datapoints.

Chart, line chart

Description automatically generated

**(b)**

Similarly, the spam dataset maintains an accuracy between 70-90% as well. However, due to the stochasticity of the results for SPAM, k-fold cross validation is needed.

**A picture containing chart

Description automatically generated**

**(c)**

For CIFAR10, the dataset maintains an accuracy of around 25-35%. Code snippets for this question are in the APPENDIX.

**Q3.**

The values tried are as follows. The best value is in bold, corresponding to a C of 10-2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Value of C | 10-3 | **10-2** | 10-1 | 100 | 101 | 102 | 103 | 104 |
| Val Acc. | 0.8887 | **0.9049** | 0.9023 | 0.8928 | 0.8764 | 0.8527 | 0.8531 | 0.8440 |
| Train Acc. | 0.8986 | **0.9241** | 0.9475 | 0.9634 | 0.9746 | 0.9633 | 0.9633 | 0.9543 |

These values were trained with a training set of 10,000 examples. And they are graphed below:

Chart, line chart

Description automatically generated

**Q4.**

Using 5-fold cross validation to get a more accurate accuracy for SPAM, the values tried are as follows. The best value is in bold, corresponding to a C of 103.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Value of C | 10-3 | 10-2 | 10-1 | 100 | 101 | 102 | **103** | 104 |
| Val Acc. | 0.7791 | 0.7949 | 0.8002 | 0.8009 | 0.8050 | 0.8079 | **0.8157** | 0.7915 |
| Train Acc. | 0.7791 | 0.7994 | 0.8052 | 0.8076 | 0.8103 | 0.8130 | **0.8173** | 0.7960 |

Chart, line chart

Description automatically generatedThe plot of the above data is below. Code for this is provided in the appendix.

**Q5.**

MNIST

Kaggle Score: 0.9833

* First, I tried standardizing the data by normalizing, i.e., dividing my 255. This created a minor difference, but not a significant one.
* Following the professor’s suggestions on piazza, I attempted to standardize the standard deviation of the dataset in addition, but soon realized that that would not help, since the MNIST data doesn’t follow a smooth normal distribution, rather, it has a few very bright points and other very dark points.
* Using this as a basis to go further, I attempted to increase the contrast. That is, I made all image points greater than 50 = 1, and all image points smaller than 50 = 0. This would ensure that only the shape of the number would be highlighted, and reduce complexity. Unfortunately, this change only reduced the performance of the algortihm.
* Finally, I switched to using sklearn’s StandardSolver(), which standardizes the data for you. I did this on both the training and validation data. This boosted the algorithm’s performance significantly.
* I then tried to increase the degree of the solver (polynomial rather than linear), and graphed the results for varying degrees. I also graphed a 2d optimization consisting of two hyperparameters: C and the degree. However, both cases showed that having a linear solver produced the best solution.
* I finally settled on ‘rbf’ rather than ‘poly’ for the solver type, and continued this for SPAM and CIFAR10 as well.
* I optimized C using a simple plot and choose technique. I finally settled on a C value of 1e5.
* I then trained on the entire MNIST dataset, including the validation data, and predicted values for the test data.

SPAM

Kaggle Score: 0.89038

* I added a few features manually, like “click”, “money”, etc., but quickly realized this was not efficient. I decided to find the optimal features using sklearn’s TfidfVectorizer.
* I combed through the entire spam email dataset, searched for words, and added them to the vectorizer. I finally made it output the 100 best features for this dataset. I did the same for the ham dataset and took the intersection of the features by combining them and using a set.
* I copy-pasted this set to the featurize.py code, and coded it so that each of the features were added to the feature vector. The final length of the feature vector was 156.
* I then performed standardization with StandardScaler(), and optimized for the hyperparameter C.
* This got me to a local training accuracy of 0.9535 with no k-fold cross validation. I saved this and sent it to Kaggle, where it unfortunately performed worse due to the stochasticity.

CIFAR10

Kaggle Score: 0.55779

* I did not work too hard for this one. I simply standardized the set with StandardSolver(), optimized for C, and then most importantly, let my computer run for a few hours on the entire CIFAR10 dataset.
* Running on the entire dataset was sufficient to boost the performance from around 33% to 55%.

**Appendix**

**Q1.**

import numpy as np

import matplotlib.pyplot as plt

from sklearn import svm

from sklearn.metrics import accuracy\_score

from scipy import io

# Q1

def load\_and\_save(filename, num = None, percent = None):

dataset = io.loadmat("data/" + filename)

if percent:

num = int(dataset["training\_labels"].shape[0] \* percent)

t\_dat, t\_lbl, v\_dat, v\_lbl = split\_data(dataset["training\_data"],

dataset["training\_labels"],

num)

dataset["training\_data"] = t\_dat

dataset["training\_labels"] = t\_lbl

dataset["valid\_data"] = v\_dat

dataset["vaild\_labels"] = v\_lbl

io.savemat("data/prep\_" + filename, dataset)

def split\_data(train, labels, num\_valid):

num\_data = train.shape[0]

assert num\_valid <= len(train)

assert num\_data == labels.shape[0]

idx = np.arange(num\_data)

np.random.shuffle(idx)

train\_shf = train[idx]

lbl\_shf = labels[idx]

valid\_dat = train\_shf[:num\_valid]

valid\_lbl = lbl\_shf[:num\_valid]

train\_dat = train\_shf[num\_valid:]

train\_lbl = lbl\_shf[num\_valid:]

return train\_dat, train\_lbl, valid\_dat, valid\_lbl

load\_and\_save("mnist\_data.mat", num = 10000)

load\_and\_save("spam\_data.mat", percent = 0.2)

load\_and\_save("cifar10\_data.mat", num = 5000)

**Q2.**

# Q2

def perform\_training\_mnist(data, num):

td = data["training\_data"][:num] / 255

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num] / 255

vl = np.ravel(data["valid\_labels"])[:num]

alg = svm.LinearSVC(max\_iter = 5000)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return num, vacc, tacc

def perform\_training\_spam(data, num):

td = data["training\_data"][:num]

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num]

vl = np.ravel(data["valid\_labels"])[:num]

alg = svm.LinearSVC(max\_iter = 100000)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return num, vacc, tacc

# MNIST

data = io.loadmat("data/prep\_mnist\_data.mat")

qt = [100, 200, 500, 1000, 2000, 5000, 10000]

tot\_acc = np.array([perform\_training\_mnist(data, i) for i in qt])

plt.plot(qt, tot\_acc.T[2], label = "Training")

plt.plot(qt, tot\_acc.T[1], label = "Validation")

plt.title("MNIST Accuracy vs. Number of Datapoints")

plt.legend()

plt.xlabel("Number")

plt.ylabel("Accuracy")

# SPAM

sata = io.loadmat("data/prep\_spam\_data.mat")

qt = [100, 200, 500, 1000, 2000, 4139]

tot\_vacc = np.array([perform\_training\_spam(sata, i) for i in qt])

plt.plot(qt, tot\_vacc.T[2], label = "Training")

plt.plot(qt, tot\_vacc.T[1], label = "Validation")

plt.title("SPAM Accuracy vs. Number of Datapoints")

plt.legend()

plt.xlabel("Number")

plt.ylabel("Accuracy")

# CIFAR

cata = io.loadmat("data/prep\_cifar10\_data.mat")

qt = [100, 200, 500, 1000, 2000, 5000]

tot\_acc = np.array([perform\_training\_mnist(cata, i) for i in qt])

plt.plot(qt, tot\_acc.T[2], label = "Training")

plt.plot(qt, tot\_acc.T[1], label = "Validation")

plt.title("CIFAR10 Accuracy vs. Number of Datapoints")

plt.legend()

plt.xlabel("Number")

plt.ylabel("Accuracy")

**Q3.**

# Q3

def perform\_training\_mnist\_C(data, num, C\_val):

td = data["training\_data"][:num] / 255

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num] / 255

vl = np.ravel(data["valid\_labels"])[:num]

alg = svm.LinearSVC(max\_iter = 5000, C = C\_val)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return num, vacc, tacc

C\_vals = np.logspace(-3, 4, 8)

data = io.loadmat("data/prep\_mnist\_data.mat")

tot\_acc = np.array([perform\_training\_mnist\_C(data, 10000, i) for i in C\_vals])

plt.plot(C\_vals, tot\_acc.T[2], label = "Training")

plt.plot(C\_vals, tot\_acc.T[1], label = "Validation")

plt.title("MNIST Accuracy vs. Value of Hyperparameter C")

plt.legend()

plt.xlabel("C")

plt.ylabel("Accuracy")

print("Validation Accuracy:", tot\_acc.T[1])

print("Training Accuracy:", tot\_acc.T[2])

**Q4.**

# Q4

def split\_data\_kcross(train, labels, num\_valid, n):

num\_data = train.shape[0]

assert num\_valid <= len(train)

assert num\_data == labels.shape[0]

idx = np.arange(num\_data)

np.random.shuffle(idx)

train\_shf = train[idx]

lbl\_shf = labels[idx]

valid\_dat = train\_shf[n \* num\_valid : (n + 1) \* num\_valid]

valid\_lbl = lbl\_shf[n \* num\_valid : (n + 1) \* num\_valid]

train\_dat = np.concatenate([train\_shf[:n \* num\_valid], train\_shf[(n + 1) \* num\_valid:]])

train\_lbl = np.concatenate([lbl\_shf[:n \* num\_valid], lbl\_shf[(n + 1) \* num\_valid:]])

return train\_dat, train\_lbl, valid\_dat, valid\_lbl

def perform\_training\_spam\_cross(data, C\_val):

td = data[0]

tl = np.ravel(data[1])

vd = data[2]

vl = np.ravel(data[3])

alg = svm.LinearSVC(max\_iter = 10000, C = C\_val)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return vacc, tacc

def averageTraining(total\_data, C\_val):

res = np.array([perform\_training\_spam\_cross(total\_data[i], C\_val) for i in range(5)])

vacc = np.mean(res.T[0])

tacc = np.mean(res.T[1])

return vacc, tacc

spam\_dat = io.loadmat("data/spam\_data.mat")

total\_data = [split\_data\_kcross(spam\_dat["training\_data"],

spam\_dat["training\_labels"],

spam\_dat["training\_data"].shape[0] // 5,

i) for i in range(5)]

C\_vals = np.logspace(-3, 4, 8)

res = np.array([averageTraining(total\_data, i) for i in C\_vals])

plt.plot(C\_vals, res.T[1], label = "Training")

plt.plot(C\_vals, res.T[0], label = "Validation")

plt.title("SPAM Accuracy vs. Value of Hyperparameter C")

plt.legend()

plt.semilogx()

plt.xlabel("C")

plt.ylabel("Accuracy")

**Q5.**

# Q5

import numpy as np

import matplotlib.pyplot as plt

from sklearn import svm

from sklearn.metrics import accuracy\_score

from scipy import io

def split\_data(train, labels, num\_valid):

num\_data = train.shape[0]

assert num\_valid <= len(train)

assert num\_data == labels.shape[0]

idx = np.arange(num\_data)

np.random.shuffle(idx)

train\_shf = train[idx]

lbl\_shf = labels[idx]

valid\_dat = train\_shf[:num\_valid]

valid\_lbl = lbl\_shf[:num\_valid]

train\_dat = train\_shf[num\_valid:]

train\_lbl = lbl\_shf[num\_valid:]

return train\_dat, train\_lbl, valid\_dat, valid\_lbl

def load\_and\_save(filename, num = None, percent = None):

dataset = io.loadmat(filename)

if percent:

num = int(dataset["training\_labels"].shape[0] \* percent)

t\_dat, t\_lbl, v\_dat, v\_lbl = split\_data(dataset["training\_data"],

dataset["training\_labels"],

num)

dataset["training\_data"] = t\_dat

dataset["training\_labels"] = t\_lbl

dataset["valid\_data"] = v\_dat

dataset["valid\_labels"] = v\_lbl

io.savemat("prep\_" + filename, dataset)

load\_and\_save("mnist\_kdata.mat", num = 10000)

def perform\_training\_mnist(data, num):

td = data["training\_data"][:num] / 255

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num] / 255

vl = np.ravel(data["valid\_labels"])[:num]

alg = svm.LinearSVC(max\_iter = 5000)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return num, vacc, tacc

from sklearn.preprocessing import StandardScaler

def perform\_training\_mnist\_2(data, num, C\_val):

td = data["training\_data"][:num]

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num]

vl = np.ravel(data["valid\_labels"])[:num]

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(vd)

alg = svm.SVC(max\_iter = 5000, kernel = 'rbf', degree = 1, C = 1e5)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return vacc

def perform\_training\_mnist\_3(data):

td = data["training\_data"]

tl = np.ravel(data["training\_labels"])

vd = data["valid\_data"]

vl = np.ravel(data["valid\_labels"])

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(vd)

alg = svm.SVC(max\_iter = 5000, kernel = 'rbf', degree = 1, C = 1e5)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return vacc

def perform\_training\_mnist\_final(data):

td = data["training\_data"]

tl = np.ravel(data["training\_labels"])

vd = data["test\_data"]

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(vd)

alg = svm.SVC(max\_iter = 5000, kernel = 'rbf', degree = 1, C = 1e5)

alg.fit(td, tl)

vp = alg.predict(vd)

return alg, vp

#data = io.loadmat("prep\_mnist\_kdata.mat")

#C\_vals = np.logspace(-5, 5, 10)

#C\_vals = np.linspace(2, 2.3, 30)

#tot\_acc = np.array([perform\_training\_mnist\_2(data, 1000, i) for i in C\_vals])

# 2D optimization of polynomial degree and C (training was tried with kernel = 'poly')

# deg = np.linspace(1, 10, 10)

# tot\_acc = np.array([[perform\_training\_mnist\_2(data, 10000, i, j) for i in C\_vals] for j in deg])

#plt.plot(C\_vals, tot\_acc.T[2] - 0.86, label = "Training")

plt.plot(C\_vals, tot\_acc.T, label = "Validation")

plt.title("MNIST Accuracy vs. Value of Degree")

plt.legend()

plt.semilogx()

plt.xlabel("C")

plt.ylabel("Accuracy")

data = io.loadmat("mnist\_kdata.mat")

alg, vp = perform\_training\_mnist\_final(data)

idx = list(range(1, len(vp) + 1))

pre\_csv = np.array(list(zip(idx, vp)), dtype = int)

np.savetxt("shri\_pred.csv", pre\_csv, fmt = '%s', delimiter=',')

from collections import defaultdict

import glob

import re

import scipy.io

import numpy as np

spam\_filenames = glob.glob('data/spam/' + '\*.txt')

ham\_filenames = glob.glob('data/ham/' + '\*.txt')

email\_list = []

for filename in spam\_filenames:

with open(filename, 'r', encoding='utf-8', errors='ignore') as f:

try:

text = f.read() # Read in text from file

except Exception as e:

# skip files we have trouble reading.

continue

text = text.replace('\r\n', ' ') # Remove newline character

# Create a feature vector

email\_list.append(text)

from sklearn.feature\_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(max\_features = 100)

X = vectorizer.fit\_transform(email\_list)

a = vectorizer.get\_feature\_names\_out()

email\_list\_ham = []

for filename in ham\_filenames:

with open(filename, 'r', encoding='utf-8', errors='ignore') as f:

try:

text = f.read() # Read in text from file

except Exception as e:

# skip files we have trouble reading.

continue

text = text.replace('\r\n', ' ') # Remove newline character

# Create a feature vector

email\_list\_ham.append(text)

vectorizer = TfidfVectorizer(max\_features = 100)

X = vectorizer.fit\_transform(email\_list\_ham)

b = vectorizer.get\_feature\_names\_out()

c = np.concatenate([a, b])

final\_feature\_list = set(c)

# Feature set was modified as follows

for i in feature\_set:

a = lambda text, freq: float(freq[i])

feature.append(a)

sata = io.loadmat('data/spam\_data.mat')

def split\_data(train, labels, num\_valid):

num\_data = train.shape[0]

assert num\_valid <= len(train)

assert num\_data == labels.shape[0]

idx = np.arange(num\_data)

np.random.shuffle(idx)

train\_shf = train[idx]

lbl\_shf = labels[idx]

valid\_dat = train\_shf[:num\_valid]

valid\_lbl = lbl\_shf[:num\_valid]

train\_dat = train\_shf[num\_valid:]

train\_lbl = lbl\_shf[num\_valid:]

return train\_dat, train\_lbl, valid\_dat, valid\_lbl

def load\_and\_save(filename, num = None, percent = None):

dataset = io.loadmat(filename)

if percent:

num = int(dataset["training\_labels"].shape[0] \* percent)

t\_dat, t\_lbl, v\_dat, v\_lbl = split\_data(dataset["training\_data"],

dataset["training\_labels"],

num)

dataset["training\_data"] = t\_dat

dataset["training\_labels"] = t\_lbl

dataset["valid\_data"] = v\_dat

dataset["valid\_labels"] = v\_lbl

io.savemat("prep\_"+filename, dataset)

def perform\_training\_spam(data, num, C\_val):

td = data["training\_data"][:num]

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num]

vl = np.ravel(data["valid\_labels"])[:num]

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(vd)

alg = svm.SVC(max\_iter = 10000, kernel = 'rbf', C = C\_val)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return vacc

def perform\_training\_spam\_final(data):

td = data["training\_data"]

tl = np.ravel(data["training\_labels"])

test = data["test\_data"]

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(test)

alg = svm.SVC(max\_iter = 10000, kernel = 'rbf', C = 14.3)

alg.fit(td, tl)

vp = alg.predict(test)

return alg, vp

load\_and\_save("spam\_data.mat", percent = 0.2)

sata\_fin = io.loadmat("prep\_spam\_data.mat")

from sklearn.preprocessing import StandardScaler

#C\_vals = np.logspace(-5, 10, 10)

#C\_vals = np.logspace(0.6, 1.4, 10)

C\_vals = np.linspace(10, 16, 10)

res = [perform\_training\_spam(sata\_fin, 5000, i) for i in C\_vals]

res = np.array(res)

plt.plot(C\_vals, res)

plt.semilogx()

data\_final = io.loadmat("spam\_data.mat")

alg, pred = perform\_training\_spam\_final(data\_final)

idx = list(range(1, len(pred) + 1))

pre\_csv = np.array(list(zip(idx, pred)), dtype = int)

np.savetxt("shri\_pred\_spam.csv", pre\_csv, fmt = '%s', delimiter=',')

def split\_data(train, labels, num\_valid):

num\_data = train.shape[0]

assert num\_valid <= len(train)

assert num\_data == labels.shape[0]

idx = np.arange(num\_data)

np.random.shuffle(idx)

train\_shf = train[idx]

lbl\_shf = labels[idx]

valid\_dat = train\_shf[:num\_valid]

valid\_lbl = lbl\_shf[:num\_valid]

train\_dat = train\_shf[num\_valid:]

train\_lbl = lbl\_shf[num\_valid:]

return train\_dat, train\_lbl, valid\_dat, valid\_lbl

def load\_and\_save(filename, num = None, percent = None):

dataset = io.loadmat(filename)

if percent:

num = int(dataset["training\_labels"].shape[0] \* percent)

t\_dat, t\_lbl, v\_dat, v\_lbl = split\_data(dataset["training\_data"],

dataset["training\_labels"],

num)

dataset["training\_data"] = t\_dat

dataset["training\_labels"] = t\_lbl

dataset["valid\_data"] = v\_dat

dataset["valid\_labels"] = v\_lbl

io.savemat("prep\_"+filename, dataset)

def perform\_training\_cifar(data, num, C\_val):

td = data["training\_data"][:num]

tl = np.ravel(data["training\_labels"])[:num]

vd = data["valid\_data"][:num]

vl = np.ravel(data["valid\_labels"])[:num]

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(vd)

alg = svm.SVC(max\_iter = 5000, kernel = 'rbf', C = C\_val)

alg.fit(td, tl)

vp = alg.predict(vd)

tp = alg.predict(td)

vacc = accuracy\_score(vl, vp)

tacc = accuracy\_score(tl, tp)

return vacc

def perform\_training\_cifar\_final(data, C\_val):

td = data["training\_data"]

tl = np.ravel(data["training\_labels"])

vd = data["test\_data"]

scaler = StandardScaler()

scaler.fit\_transform(td)

scaler.transform(vd)

alg = svm.SVC(max\_iter = 5000, kernel = 'rbf', C = C\_val)

alg.fit(td, tl)

vp = alg.predict(vd)

return alg, vp

load\_and\_save("cifar10\_data.mat", num = 5000)

cata = io.loadmat("prep\_cifar10\_data.mat")

#C\_vals = np.logspace(-5, 5, 10)

C\_vals = np.logspace(0, 30, 10)

res = np.array([perform\_training\_cifar(cata, 1000, i) for i in C\_vals])

plt.plot(C\_vals, res)

plt.semilogx()

cata\_fin = io.loadmat("cifar10\_data.mat")

alg, vp = perform\_training\_cifar\_final(cata\_fin, 1e7)

idx = list(range(1, len(vp) + 1))

pre\_csv = np.array(list(zip(idx, vp)), dtype = int)

np.savetxt("shri\_pred\_cifar.csv", pre\_csv, fmt = '%s', delimiter=',')