CS 189/289A Introduction to Machine Learning
Spring 2025 Jonathan Shewchuk HW1

Due: Wednesday, January 29 at 11:59 pm

This homework comprises a set of coding exercises and a few math problems. While we have you train models across three datasets, the code for this entire assignment can be written in under 250 lines. Start this homework early! You can submit to Kaggle only twice a day.

Deliverables:

1. Submit your predictions for the test sets to Kaggle as early as possible. Include your Kaggle scores in your write-up (see below).

2. Submit a **PDF** of your homework, **with an appendix listing all your code**, to the Gradescope assignment entitled "HW1 Write-Up". You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.

   - On the first page of your write-up, please list students who helped you or whom you helped on the homework. (Note that sending each other code is not allowed.)
   - On the first page of your write-up, please copy the following statement and sign your signature next to it. (Mac Preview, PDF Expert, and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make *extra* clear the consequences of cheating.

     "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."

3. Submit all the code needed to reproduce your results to the Gradescope assignment entitled "HW1 Code". You must submit your code twice: once in your PDF write-up (above) so the readers can easily read it, and again in compilable/interpretable form so the readers can easily run it. **Do NOT include any data files we provided.** Please include a short file named README listing your name, student ID, and instructions on how to reproduce your results. Please take care that your code doesn't take up inordinate amounts of time or memory.

   The Kaggle score will not be accepted if the code provided a) does not compile or b) compiles but does not produce the file submitted to Kaggle.

Python Configuration and Data Loading

This section is only setup and requires no submitted solution. Please follow the instructions below to ensure that your Python environment is configured properly, and you are able to successfully load the data provided with this homework. For all coding questions, we recommend using Anaconda for Python 3.

(a) Either install Anaconda for Python 3, or ensure you're using Python 3. To ensure you're running Python 3, open a terminal in your operating system and execute the following command:

```
python --version
```

Do not proceed until you're running Python 3.

(b) Install the following dependencies required for this homework by executing the following command in your operating system's terminal:

```
pip install scikit-learn scipy numpy matplotlib
```

Please use Python 3 with the modules specified above to complete this homework.

(c) You will be running out-of-the-box implementations of Support Vector Machines to classify three datasets. You will find a set of `.npz` files in the `data` folder for this homework. Each `.npz` file will load as a Python dictionary. Each dictionary contains three fields:

- **training_data**, the training set features. Rows are sample points and columns are features.
- **training_labels**, the training set labels. Rows are sample points. There is one column: the labels corresponding to rows of training_data above.
- **test_data**, the test set features. Rows are sample points and columns are features. You will fit a model to predict the labels for this test set, and submit those predictions to Kaggle.

The three datasets for the coding portion of this assignment are described below.

- **toy-data.npz** is a synthetic dataset with two features (2-dimensional) and two classes. The training set has 1,000 examples, and no test set is provided. This dataset is only used in Section 2 of this homework.
- **mnist-data.npz** contains data from the MNIST dataset. There are 60,000 labeled images of handwritten digits for training, and 10,000 for testing. The images are grayscale and provided by 28×28 pixels. There are 10 possible labels for each image: the digits 0–9. We will use the simplest of features for classification: raw pixel brightness values. In other words, our feature vector for an image will be a row vector with all the pixel values concatenated in a row major (or column major) order.

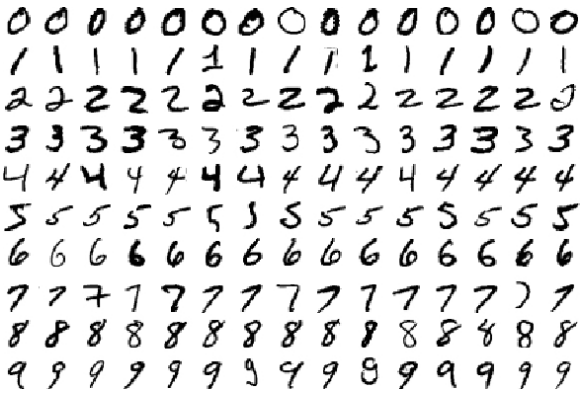

Figure 1: Examples from the MNIST dataset.

- **spam-data.npz** contains email data. The labels are 1 for spam and 0 for ham (not spam). We provide the raw email data in the subfolders **spam**, **ham**, and **test** (unlabeled test data) in the `data` folder. We also provide the script **featurize.py** to compute the frequency of certain words within the email text and generate a corresponding feature vector. You may modify **featurize.py** to generate new features for the email data.

To check whether your Python environment is configured properly for this homework, run the `load` script (copied below) from within the scripts folder. This script should load each dataset and print the shapes of the training data, training labels, and test data. Confirm that the shapes align with your expectations based on the descriptions above.

Pay attention to errors raised when attempting to import any dependencies. Resolve such errors by manually installing the required dependency (e.g. execute `pip install numpy` for import errors relating to the numpy package). You must have your environment configured properly before moving on.

```
import sys
if sys.version_info[0] < 3:
  raise Exception("Python 3 not detected.")
import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm
from scipy import io

if __name__ == "__main__":
    for data_name in ["toy", "mnist", "spam"]:
        data = np.load(f"../data/{data_name}-data.npz")
        print("\nloaded %s data!" % data_name)
        fields = "training_data", "training_labels", "test_data"
        for field in fields:
            print(field, data[field].shape)
```

1 Honor Code

Declare and sign the following statement:

"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

Signature : _____

While discussions are encouraged, *everything* in your solution must be your (and only your) creation. Furthermore, all external material (i.e., *anything* outside lectures and assigned readings, including figures and pictures) should be cited properly. We wish to remind you that consequences of academic misconduct are *particularly severe*!

This section provides background information on Support Vector Machines (SVMs) used in this homework. You can choose to focus on the coding sections first and revisit this section later, but make sure that this section precedes the coding questions in your write-up.

2 Theory of Hard-Margin Support Vector Machines

A *decision rule* (or *classifier*) is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to $+1$ ("in class") or -1 ("not in class"). The decision rule for linear SVMs is of the form

$$r(x) = \begin{cases} +1 & \text{if } w \cdot x + \alpha \geq 0, \\ -1 & \text{otherwise,} \end{cases} \tag{1}$$

where $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ are the parameters of the SVM. To select the best parameters for this decision rule, suppose we have a set of n training points and correponding labels. Let $X_i \in \mathbb{R}^d$ be the ith training point with corresponding label $y_i \in \{+1, -1\}$. The primal hard-margin SVM optimization problem (used to find the optimal decision rule parameters) is

$$\min_{w,\alpha} \ \|w\|^2 \ \text{ subject to } \ y_i(X_i \cdot w + \alpha) \geq 1, \ \ \forall i \in \{1, \ldots, n\}, \tag{2}$$

where $\|\cdot\|$ denotes the ℓ_2 norm (i.e., $\|w\| = \sqrt{w \cdot w}$).

We can rewrite this optimization problem by using Lagrange multipliers to eliminate the constraints. (If you're curious to know what Lagrange multipliers are, we recommended taking a look at this Wikipedia page, but you don't need to understand them to do this problem.) We thereby obtain the equivalent optimization problem

$$\max_{\lambda_i \geq 0} \min_{w,\alpha} \ \|w\|^2 - \sum_{i=1}^{n} \lambda_i (y_i(X_i \cdot w + \alpha) - 1). \tag{3}$$

Note: λ_i must be greater than or equal to 0.

(a) Show that equation (3) can be rewritten as the *dual optimization problem*

$$\max_{\lambda_i \geq 0} \ \sum_{i=1}^{n} \lambda_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j X_i \cdot X_j \ \text{ subject to } \ \sum_{i=1}^{n} \lambda_i y_i = 0. \tag{4}$$

Hint: Use calculus to determine and prove what values of w and α optimize equation (3). Explain where the new constraint comes from.

(b) Suppose we know the values λ_i^* and α^* that optimize equation (3). Show that the decision rule specified by equation (1) can be written

$$r(x) = \begin{cases} +1 & \text{if } \alpha^* + \frac{1}{2} \sum_{i=1}^{n} \lambda_i^* y_i X_i \cdot x \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{5}$$

(c) Applying Karush–Kuhn–Tucker (KKT) conditions (see this Wikipedia page for more information), any pair of optimal primal and dual solutions w^*, α^*, λ^* for a linear, hard-margin SVM must satisfy the following condition:

$$\lambda_i^*(y_i(X_i \cdot w^* + \alpha^*) - 1) = 0 \quad \forall i \in \{1, \ldots, n\}$$

This condition is called *complementary slackness*. Explain what this implies for points corresponding to $\lambda_i^* > 0$. That is, what relationship must exist between the data points, labels, and decision rule parameters? (You can provide a one sentence answer.)

(d) The training points X_i for which $\lambda_i^* > 0$ are called the *support vectors*. In practice, we frequently encounter training data sets for which the support vectors are a small minority of the training points, especially when the number of training points is much larger than the number of features. Explain why the support vectors are the only training points needed to use the decision rule.

(e) The obtained parameters when fitting the linear SVM to the 2D synthetic dataset found in **toy-data.npz** approximately correspond to

$$w = \begin{bmatrix} -0.4528 \\ -0.5190 \end{bmatrix} \quad \text{and} \quad \alpha = 0.1471. \tag{6}$$

Using only matplotlib basic plotting functions, produce a plot of

- the data points,
- the decision boundary,
- the margins, defined as $\{x \in \mathbb{R}^2 : w \cdot x + \alpha = \pm 1\}$.

Clearly indicate the points in your plot that are support vectors.

Hint: You can use the functions in the code snippet below, which plot the data points and decision boundary but not the margins.

```
# Plot the data points
def plot_data_points(data, labels):
    plt.scatter(data[:, 0], data[:, 1], c=labels)

# Plot the decision boundary
def plot_decision_boundary(w, b):
    x = np.linspace(-5, 5, 100)
    y = -(w[0] * x + b) / w[1]
    plt.plot(x, y, 'k')

# Plot the margins
def plot_margins(w, b):
    ## TODO
```

(f) Assume the training points are linearly separable. Using the original SVM formulation in equation 2, prove that there is at least one support vector for each class, +1 and −1.

Hint: Use contradiction. Construct a new weight vector $w' = w/(1 + \epsilon/2)$ and corresponding bias α' where $\epsilon > 0$. It is up to you to determine what ϵ should be based on the contradiction. If you provide a symmetric argument, you need only provide a proof for one of the two classes.

For the entire assignment, you may use sklearn only for the SVM model. Everything else must be done without the use of sklearn.

3 Data Partitioning and Evaluation Metrics

In machine learning, it is typical to rely on a set of held-out data points, referred to as the *validation* set, to evaluate the performance of various models and ultimately select the best performing one, while using the rest of the data, or *training* set, to train the models. In its simplest form, evaluating a trained model requires you to (i) set aside a validation set and (ii) select a reasonable metric to evaluate model performance.

In this question, you will implement these components that will be useful for the rest of the assignment. **Please do not use any sklearn functions in this section**.

(a) **Data partitioning**: Rarely will you receive "training" data and "validation" data; usually you will have to partition available labeled data yourself. In this question, you will *shuffle and partition* each of the datasets in the assignment[1]. Shuffling prior to splitting crucially ensures that all classes are represented in your partitions. For this question, please do not use any functions available in sklearn. For the MNIST dataset, write code that sets aside 10,000 training images as a validation set. For the spam dataset, write code that sets aside 20% of the training data as a validation set.

(b) **Evaluation metric**: There are several ways to evaluate models. We will use *classification accuracy*, or the percent of examples classified correctly, as a measure of the classifier performance. Error rate, or one minus the accuracy, is another common metric. Suppose you have a set of n input observations. Write a function that takes as inputs the set of true labels, $y \in \mathbb{R}^n$, and the set of predicted labels, $\hat{y} \in \mathbb{R}^n$, and computes the classification accuracy score

$$s = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[y_i = \hat{y}_i\right]. \tag{7}$$

In the accuracy score, $\mathbb{I}\left[y_i = \hat{y}_i\right]$ is an indicator function defined as

$$\mathbb{I}\left[y_i = \hat{y}_i\right] = \begin{cases} 1 & \text{if } y_i = \hat{y}_i \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

Here, y_i and $\hat{y}_i$ respectively denote the ground-truth and predicted label for observation i.

Deliverable: Attach a copy of your data partitioning and evaluation metric code to your homework report under question 3.

[1]Make sure that you shuffle the labels with the training images. It's a very common error to mislabel the training images by forgetting to permute the labels with the images!

4 Support Vector Machines: Coding

We will use linear Support Vector Machines (SVMs) to classify our datasets. Train a linear SVM on the MNIST and spam datasets that you divided into training and validation sets in the previous problem. For each dataset, plot the classification accuracy (as defined in the previous problem) on the training and validation sets versus the number of training examples that you used to train your classifier. The number of training examples to use are listed for each dataset in the following parts. To evaluate validation accuracy, you should always use the same validation set obtained in the previous problem. To evaluate training accuracy, you should use the portion of the full training set that is actually used to train the model.

You may use `sklearn` **only** for the SVM model. Everything else must be done without the use of `sklearn`. **Note:** You can use either `SVC(kernel='linear')` or `LinearSVC` as your SVM model, though they each solve slightly different optimization problems using different libraries.

(a) For the **MNIST** dataset, use raw pixels as features. Train your model with the following numbers of training examples: 100, 200, 500, 1,000, 2,000, 5,000, 10,000. You should expect validation accuracies between 70% and 90%. [2]

(b) For the **spam** dataset, use the provided word frequencies as features. Train your model with the following numbers of training examples: 100, 200, 500, 1,000, 2,000, **ALL**. (**ALL** is all of the data we already set aside for training.) Performance may vary some for this dataset, but you should expect validation accuracies between 70% and 90%.

Deliverable: For this question, you should include two plots showing training and validation accuracy versus number of examples for each of the datasets. Additionally, be sure to include your code in the "Code Appendix" portion of your write-up.

[2] Hint: Be consistent with any preprocessing you do. Use either integer values between 0 and 255 or floating-point values between 0 and 1. Training on floats and then testing with integers is bound to cause trouble.

5 Hyperparameter Tuning

In the previous problem, by training SVM models with varying amounts of training examples, you learned parameters for a model that classifies the data. Many classifiers also have *hyperparameters* that you can tune to influence the parameters. In this problem, we'll determine good values for the regularization parameter C in the soft-margin SVM algorithm. The interpretation of this parameter, as well as the functioning of the soft-margin SVM will be covered in lecture. For now, consider C as a parameter of a black-box algorithm that we aim to optimize.

When we are trying to choose a hyperparameter value, we train the model repeatedly with different hyperparameters. We select the hyperparameter that gives the model the highest *validation* accuracy. Before generating predictions for the test set, the model should be retrained using all the labeled data (including the validation data) and the previously-determined hyperparameter. **The use of automatic hyperparameter optimization libraries is strictly prohibited.**

For the MNIST dataset, train a linear SVM model with various values of C to determine the best value for this hyperparameter. You should try different orders of magnitude of C values (e.g., 0.1, 1, 10, etc.). For performance reasons, you are required to train with at least 10,000 training examples. You can train on more if you like, but it is not required.

Deliverable: In your report, list at least 8 C values you tried, the corresponding accuracies, and the value corresponding to the highest validation accuracy. Reference any code you used to perform a hyperparameter sweep in the code appendix.

6 K-Fold Cross-Validation

For smaller datasets (like the spam dataset), the validation set contains fewer examples, which means that our estimate of accuracy has high variance and thus might not be accurate. A way to combat this is to use *k-fold cross-validation*.

In k-fold cross-validation, the training data is shuffled and partitioned into k disjoint sets. Then the model is trained on $k-1$ sets and validated on the k^{th} set. This process is repeated k times with each set chosen as the validation set once. The cross-validation accuracy we report is the accuracy averaged over the k iterations. **The use of automatic cross-validation libraries is prohibited.**

For the spam dataset, use 5-fold cross-validation to find and report the best C value. As in problem 5, you should try different orders of magnitude of C. **Hint:** Effective cross-validation requires choosing from *random* partitions. This is best implemented by randomly shuffling your training examples and labels, and then partitioning them by their indices.

Deliverable: In your report, list at least 8 C values you tried, the corresponding accuracies, and the best C value. Reference any code you used to perform cross validation in the code appendix.

7 Kaggle

- MNIST Competition: https://www.kaggle.com/t/4bd3a8ef94dc4b3e88b57d45251742a6
- SPAM Competition: https://www.kaggle.com/t/dd35c1b3d87346458938b74689a3a5d6

Using an SVM model, generate predictions for the two test sets we provide and save those predictions to CSV files. The CSV file should have two columns–`Id` and `Category`–with a comma as a delimiter between them. The Id column should start at the integer 1 and end at the number of elements in the test set. The category label should be a dataset-dependent integer (one of $\{0, \ldots, 9\}$ for MNIST and one of $\{0, 1\}$ for spam. Be sure to use integer labels (not floating-point), and ensure that there are no spaces (not even after the commas). You can use the function below (copied from `save_csv`) to generate the CSV files:

```
def results_to_csv(y_test, file_name):
    y_test = y_test.astype(int)
    df = pd.DataFrame({'Category': y_test})
    df.index += 1
    df.to_csv(file_name, index_label='Id')
```

To check that your CSV files are formatted correctly, use `scripts/check.py` as a sanity check:

```
python check.py <competition name, eg. mnist> <submission csv file>
```

Once you have properly formatted CSV files with your predictions, upload your predictions to the Kaggle leaderboards and view the accuracy of your models. **Note that Kaggle only permits two submissions per leaderboard per day, so please start early!**

General comments about the Kaggle sections of homeworks: Most or all of the coding homeworks will include a Kaggle section. Whereas other parts of the homework might impose strict limits on what methods and libraries you are permitted to use, the Kaggle portions permit you to apply your creativity and find clever ways to improve your leaderboard performance. (Although extensive creativity is not generally necessary to get full points on an assignment, topping the Kaggle leaderboard gives your professor good material for letters of recommendation.)

The main restriction for Kaggle competitions is that you cannot use an entirely different learning technique. For example, this is an SVM homework, so you must use an SVM. (You are not permitted to use a neural network or a decision tree instead of, or in addition to, an SVM.) You are also absolutely not allowed to search for the labeled test data and submit that to Kaggle.

There are many other allowable ways to achieve higher positions on the Kaggle leaderboards. For example, you may use a nonlinear SVM kernel or add/remove features to/from the data. (For reasons we will learn later this semester, dropping features that have little or no predictive power will often improve your test performance as much as adding the right new features.) Spam is a particularly good dataset for playing with feature engineering. One easy way to perform better in the spam competetion is to add extra features with `featurize.py`. (We describe how you might do this below.) For this dataset, you could also look into using a bag-of-words model. For image data (i.e., MNIST), you could explore SIFT and HOG features.

Whatever creative ideas you apply, please explain what you did in your write-up. Cite any external sources you used to get ideas. If you have any questions about whether something is allowed or not, please ask on Ed Dicussion to avoid being penalized.

Deliverable: Your deliverable for this question has three parts. First, include a screenshot of your place on the Kaggle leaderboard for each of the datasets. (Be sure to include your Kaggle name, your score, and the time of your last submission. The time of your submission must be before your submission of the homework on Gradescope.) Second, include an explanation of what you tried, what worked, and what did not work to improve your accuracy. Finally, make sure to include all your code in the code appendix and provide a reference to it.

Modifying features for spam: The Python script `scripts/featurize.py` extracts features from the original emails in the spam dataset. The spam emails can be found in `data/spam/`, the ham emails can be found in `data/ham/`, and the emails for the test set can be found in `data/test/`. You are encouraged to look at the emails and try to think of features you think would be useful in classifying an email as spam or ham.

To add a new feature, modify `featurize.py`. You are free to change the structure of the code provided, but if you are following the given structure, you need to do two things:

- Define a function (e.g., `my_feature(text, freq)`) that computes the value of your feature for a given email. The argument `text` contains the raw text of the email, and `freq` is a dictionary containing the counts of each word in the email (or 0 if the word is not present). The value you return should be an integer or a float.
- Modify `generate_feature_vector` to append your feature to the feature vector.

Once you are done modifying `scripts/featurize.py`, re-generate the training and test data by running this script from within the `scripts` directory.