Homework 1, Machine Learning, Fall 2023

***IMPORTANT* Homework Submission Instructions**

1. All homeworks must be submitted in one PDF file to Gradescope.
2. Please make sure to select the corresponding HW pages on Gradescope for each question
3. For all coding components, complete the solutions with a Jupyter notebook/Google Colab, and export the notebook (including both code and outputs) into a PDF file. Concatenate the theory solutions PDF file with the coding solutions PDF file into one PDF file which you will submit.
4. Failure to adhere to the above submission format may result in penalties.

As a reminder, don't wait until the last minute to ask for help, because the person you need to speak with might be dealing with many other students close to the deadline. Also, the teaching staff has been instructed that it is not mandatory for them to answer questions on the day the homework is due. So get your questions in early!

All homework assignments must be your independent work product, no collaboration is allowed. You may check your assignment with others or the internet after you have done it, but you must actually do it by yourself. **Please copy the following statements at the top of your assignment file**:

Agreement 1) This assignment represents my own work. I did not work on this assignment with others. All coding was done by myself.

Agreement 2) I understand that if I struggle with this assignment that I will reevaluate whether this is the correct class for me to take. I understand that the homework only gets harder.

1 Concepts of Learning - Theory (Stephen)

The goal of this question is to get you familiarized with the conceptual taxonomy of different types of machine learning tasks and to get you thinking about how ML could be applied in real life.
Some common types of machine learning tasks/problems are listed below:

- classification
- regression
- ranking
- clustering
- conditional probability estimation
- density estimation
- pattern mining

You are given a list of situations below. Assign one machine learning task from the list above to each situation below.
The situations are deliberately designed to simulate real-life applications and hence are open-ended. For each situation, there may be more than one answer that could be appropriate, depending on your interpretation of the data and task. **You need only give ONE reasonable answer for each situation to get full credit.** Please limit any justification to at most two sentences.

1. You are developing a multiplayer online real-time strategy game and you need to find a way to match players of similar skill against one another. Ideally, the players will be consistently challenged, not bored by less experienced players or destroyed by more experienced players. You have the players play a few games so you can retrieve features related to their playing ability.

2. You are a medical doctor trying to determine where the different segments of brain tumors are located in MRI images. You have 1,000 images of patients diagnosed with brain tumors. Using your expert knowledge, you hand label 100 images yourself by highlighting the appropriate pixels corresponding to edema, necrosis, or enhancing segments. You would like an ML algorithm to do this work for you.

3. You are a composer building a system that automatically generates ragtime piano pieces. Part of this process involves coming up with a harmonic progression for individual phrases. You have analyzed the harmonic progressions of 20 ragtime pieces. You know that phrases are goal-oriented, ending on either a “I” or “V” harmony, so you decide to generate harmonic progressions backwards from the ends of phrases. You need to find a distribution over possible harmonies for time t given the harmony at time $t + 1$. This way, you can sample from the possible harmonies and avoid using the most common progression every time.

4. You are the Director of Pricing Algorithms & Data Science for Petco. You are trying to determine the pricing algorithm for cat food. Specifically, you need to estimate how the prices of certain items should change over time at specific stores. You have access to a database containing years of customer, store, and product data (store location, Petco product price over time, competitor product price over time, number of customers/day, etc.).

5. You work for Discover bank and need to create an algorithm to detect fraudulent transactions. You have historic transaction data including transaction amount, time of transaction, location, etc. Based on this data, you need to build an understanding of the underlying distribution of standard transactions. You can then monitor new transactions to see how well they fit into the expected distribution, flagging unusual transactions as potentially fraudulent.

6. You are working on the election campaign for your favorite senator. You must determine for each potential voter whether they are “strong supporters,” “undecided,” “swing voters,” or “unlikely to vote.” With this information, your team can focus its campaign towards winning over the undecided and swing voter categories rather than wasting time and money on strong supporters or unlikely voters. You have demographic data on potential voters including age, gender, location, party affiliation, etc.

7. You work for CarGurus and need to design a generalized search algorithm. Given the users’ search words you are tasked with finding the most relevant cars and presenting them in order of relevance. You are provided with a dataset of queries and their corresponding search results. Each result is labeled with a relevance score describing how well it matches the search query.

8. You are the owner of a restaurant that is famous for your vegetable soup. You are trying to determine how many pounds of vegetables to buy for next week. If you buy too much, the leftover vegetables go to waste. If you buy too little, you will run out of vegetables prematurely and disappoint your customers. You have data about all past weeks (how many customers you had, whether there was a holiday, number of rainy days, etc.).

9. You are a marketing analyst at Express. You are trying to determine the public opinion on an experimental line of green suits. You develop a natural language processing algorithm to read Threads and Twitter/X posts with the appropriate hashtags and determine whether each post is “positive” or “negative.”

10. You work for Spotify, improving their music recommendation system. You have access to millions of users’ listening data. Your goal is to find common connections between the genres and songs/pieces that these users are listening to. Using what is learned, the system can recommend music that is likely to interest a user based on their listening/search history.

2 Model Selection - Theory (Yiyang)

Consider yourself as a data scientist working for a healthcare company. Your team has been tasked with developing a predictive model to identify the risk of patients getting strokes (-1 for healthy patients and 1 for stroke patients) based on various factors such as age, lifestyle, genetic markers, and medical history.

To solve this task, you have developed two models, **M1** and **M2**, both of which have similar mean predictive accuracy on the training set. However, **M1** is a high-degree polynomial classifier with 10 parameters, while **M2** is a linear classifier with 5 parameters.

(a) Based on the current result, which model would you expect to generalize better to the test set, and why? Explain your reason in one sentence.

(b) Suppose you have developed a third model, **M3**, which is a polynomial regression model with 100 parameters. This model has significantly better predictive accuracy than both **M1** and **M2** on the training dataset. However, when you test **M3** on a test dataset, its performance significantly drops. What might be the reason for this drop in performance?

(c) What can you do to avoid the problem of performance dropping in Problem (b)? How does that change **M3**’s model complexity?

Now, you look closer into your collected dataset and find that the proportion of stroke patients and non-stroke patients in both your training and validation dataset is 1:8, and the confusion matrix for **M1** and **M2** for the validation dataset is shown below:

M1	**True Stroke**	**True Healthy**
Predicted Stroke	85	10
Predicted Healthy	15	890

M2	**True Stroke**	**True Healthy**
Predicted Stroke	70	8
Predicted Healthy	30	892

(d) Calculate the sensitivity, specificity, and accuracy performance for **M1** and **M2**. Would your answer to the first question change? Why or why not? If you want to train a good classifier using the same dataset, what kind of techniques can be employed at training time to overcome challenges working with this imbalanced dataset?

3 Regularization (Jon)

Let's try to determine whether adding an additional regularization term to a model's objective function reduces the model's complexity.

Consider 0-1 loss, and models (parameterized by θ) that use $m(\theta)$ variables, where we regularize the number of variables. So the objective is:

$$\min_{\theta} \mathcal{L}(\theta) + \lambda m(\theta),$$

where $\mathcal{L}(\theta)$ is $1/n$ times the number of misclassifications. n is the number of data points.

1. What is the largest value of λ that cannot affect the accuracy of the optimal solution θ_0? In other words, if $\lambda < \mathcal{N}$ for some number $\mathcal{N}$, then for any $\theta_0 \in \arg\min_\theta \mathcal{L}(\theta)$ and any $\theta_\lambda \in \arg\min_\theta \mathcal{L}(\theta) + \lambda m(\theta)$, $\mathcal{L}(\theta_0) = \mathcal{L}(\theta_\lambda)$. What is the largest possible $\mathcal{N}$? Hint: this answer relies on using the 0-1 loss function. $\mathcal{N}$ will depend on the number of variables in θ_0. There are two parts for this proof: showing that if $\lambda < \mathcal{N}$, the condition holds, and if $\lambda > \mathcal{N}$, it is possible for the condition **not** to hold.

2. Consider θ_1 and θ_2, which both have the same optimal objective value, $\mathcal{L}(\theta_1)+\lambda m(\theta_1) = \mathcal{L}(\theta_2)+\lambda m(\theta_2)$. We have that model θ_2 has 2 fewer variables than θ_1. Express θ_1's training error $\mathcal{L}(\theta_1)$ in terms of θ_2's training error $\mathcal{L}(\theta_2)$ and λ.

3. We again have two models, θ_3 and θ_4 which were optimized using objectives that had different regularization parameters λ_3 and λ_4 where $\lambda_3 > \lambda_4$. We know that the second one is more accurate on the training set, $\mathcal{L}(\theta_4) < \mathcal{L}(\theta_3) + \epsilon$. We also know that the objective of θ_3 equals that of θ_4, $\mathcal{L}(\theta_3) + \lambda_3 m(\theta_3) = \mathcal{L}(\theta_4) + \lambda_4 m(\theta_4)$. We also know that θ_3 is not too much smaller than θ_4 in that it uses at most z fewer variables. Then, it is true that θ_4 is at most a certain size, specifically: $m(\theta_4) <$ function$(\lambda_3, \lambda_4, \epsilon, z)$. What is this function?

4 Classifiers and Metrics - Coding (Stark)

Age	like Rowing	Experience	Income	Y
20	1	0	20	0
18	1	1	33	0
11	0	1	21	1
31	0	0	9	1
22	1	1	7	1
21	1	0	10	0
13	1	0	23	1
15	1	1	16	0
16	0	1	15	1
17	1	0	6	0

You are given the dataset above with feature vector $\mathbf{x}$ including Age, likeRowing, Experience, and Income, and the binary label Y, whether the student is accepted to the Stanford rowing team. You are also given a linear classifier $g(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$ and a non-linear classifier $f(\mathbf{x}) = \tanh(\boldsymbol{\theta}^\top \mathbf{x} + \theta_0)$, where $\boldsymbol{\theta} = (0.05, -3, 2.1, 0.008)$, $\theta_0 = 0.3$, and "tanh" function $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. (In this question, you are expected to write functions from scratch, but packages including Matplotlib and NumPy are allowed.)

(a) First calculate the value of $g(\mathbf{x})$ for each data point. What is the largest threshold value that would minimize (mis)classification error?

(b) Calculate the value $f(\mathbf{x})$ for each data point. What is the largest threshold value that would minimize (mis)classification error? Compute the confusion matrix, precision, recall, and F1 score for one such threshold.

(c) For classifiers $f(\mathbf{x})$ and $g(\mathbf{x})$, plot the ROC curves. Please plot each ROC curve as a continuous, connected set of lines. Plot all the points on the ROC curve that represent decision points with the minimum classification error.

(d) For the ROC curves in (c), calculate the AUC from scratch using only the numpy package (do not use sklearn or similar packages).

5 K-Nearest Neighbors with Parameter Tuning - Coding (Harry & Eric)

In this problem, you will implement from scratch the k-NN algorithm on the breast cancer dataset. You will also implement your own cross-validation algorithm in order to tune your model. You should not use a pre-existing k-NN algorithm or cross-validation algorithm such as from Sklearn. If you are unsure whether a package is allowed, feel free to ask in EdDiscussion.

The breast cancer dataset contains 30 feature variables and a target variable which you are trying to predict. The data has already been split into a training and test set for you.

(a) Is accuracy or F1 score a more appropriate performance metric to use for this task? Why?

(b) Implement a k-NN algorithm from scratch to classify the dataset. Use $k = 31$ and the Euclidean distance, and make sure to normalize the data. Report your model's F1 score on the test set.

Note: See the sklearn article on the MinMaxScaler for more information on how to perform the normalization without leaking information between the train and test sets. Make sure to still implement it from scratch.

(c) Use cross-validation with 5 folds and the F1 score to tune the value of k and the distance function used (possible distance functions to use could be Euclidean distance, Manhattan distance, or cosine similarity). Make sure to use at least five values of k between 1 and 63, and try at least two distance functions.

For each distance function, show a plot where the x-axis depicts the value of k and the y-axis depicts the average F1 score for that value of k during cross-validation. Which pair of parameters performed the best?

Note: You may find the array_split method from NumPy helpful when implementing cross-validation.

(d) Using the best parameters determined in part **(c)**, report the performance of a k-NN classifier on the test set. Compare this to your model in part **(b)**. Is it as you expected?