Machine Learning, Fall 2019: Project 1
out on 9/5/19 - due on 9/19/19 by noon on Blackboard

Your Name Here

Header: List the major resources you used to complete this project and the programming language you used.

You may use any programming language you like (Matlab, C++, C, Java...). All programming must be done individually from first principles. You are only permitted to use existing tools for simple linear algebra such as matrix multiplication/inversion. **Do NOT use any toolkit that performs machine learning functions and do NOT collaborate with your classmates. Cite any resources that were used.**

In this project you will practice the basics of Machine Learning Classification by creating a K-NN classifier for two datasets. You will also practice good practices for how to describe, evaluate, and write up a report on the classifier performance. The last problem on decision trees does not require programming.

It is expected that your project report may require 2 pages per dataset if you are good about making interesting figures and making them not too large, or 3-4 pages if your figures are big. The LaTex that generated this page is available here: https://www.overleaf.com/4472941215hqtgjkfvtybq. Please submit a pdf file for project 1.

Datasets: The project will explore two datasets, the famous MNIST dataset of very small pictures of handwritten numbers, and a dataset that explores the prevelance of diabetes in a native american tribe named the Pima. You can access the datasets here:

1. https://www.kaggle.com/uciml/pima-indians-diabetes-database
2. https://www.kaggle.com/c/digit-recognizer/data

Programming Task: For each dataset, you must create a K-NN classifier that uses the training data to build a classifier, and evaluate and report on the classifier performance.

(20 points) Dataset details: Describe the data and some simple visualizations (for images, a few examples from each category; for other data, perhaps some scatter plots or histograms that show a big picture of the data). Describe your training/test split for K-NN and justify your choices.

(10 points) Algorithm Description: K-NN is a very clear algorithm, so here describe any data preprocessing, feature scaling, distance metrics, or otherwise that you did.

(40 points) Algorithm Results: Show the accuracy of your algorithm — in the case of the Pima Dataset, show accuracy with tables showing false positive, false negative, true positive and true negatives. For the Pima Dataset, use three different distance metrics and compare the results.

In the case of the MNIST digits show the complete confusion matrix. Choose a single digit to measure accuracy and show how that number varies as a function of K.

(10 points) Runtime: Describe the run-time of your algorithm and also share the actual "wall-clock" time that it took to compute your results.

Decision Trees: Consider the following set of training examples for the unknown target function $< X_1, X_2 > \rightarrow Y$.

Y	X_1	X_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

1. **(10 points)** What is the sample entropy $H(Y)$ for this training data (with logarithms base 2)?
2. **(10 points)** What are the information gains $IG(X_1) \equiv H(Y) - H(Y|X_1)$ and $IG(X_2) \equiv H(Y) - H(Y|X_2)$ for this sample of training data?
3. **(5 points)** Draw the decision tree that would be learned by ID3 (without postpruning) from this sample of training data.