CSCI 416/516 Homework #1

DUE: September 27, 2023, at 11:59 pm

CSCI 416/516: Each Problem begins with an allocation of points, represented as [u pts/g pts]. If you are registered in CSCI 416, you can receive up to u pts on this Problem; if you are registered in CSCI 516, you can receive up to g pts on this Problem. The last Problem is optional for undergraduates (CSCI 416) but required for graduates (CSCI 516). **Write down which session you are in / are you a graduate or undergraduate student**.

Submission: You need to submit both your homework report (answers to the Problems) as a pdf file and your code (Jupyter Notebook exported as a pdf file) through Blackboard. The pdf file exported from Jupyter Notebook must show the same output as the ones indicated in your submitted homework report.

- **Problem 1 [3pts/3pts]: Euclidean Distance.**
  Consider the following 3-dimensional points, $x^{(a)} = [1, -3, 5]$ and $x^{(b)} = [-2, 4, -6]$. Write the formula for the Euclidean distance between two points in a 3-dimensional space. Then, using the formula, calculate the Euclidean distance between $x^{(a)}$ and $x^{(b)}$.

- **Problem 2 [3pts/3pts]: Curse of Dimensionality.**
  Imagine you're working with a dataset of e-commerce product reviews. Each review is represented as a vector, where each dimension corresponds to the frequency of a particular word from a predefined vocabulary. The dataset has 10,000 reviews, and the vocabulary size is 50,000 words. Explain what is meant by the "curse of dimensionality" in the context of this high-dimensional dataset. How many dimensions are there in this case?

- **Problem 3 [3pts/2pts]: Normalization.**
  Explain why it is often recommended to normalize or standardize features when using the K-NN algorithm.

- **Problem 4 [6pts/4pts]: KNN on MNIST.**
  The MNIST dataset is a large collection of handwritten digits that is commonly used for training image processing systems. Each image in the MNIST dataset is a 28×28 grayscale image, represented as a 784-dimensional vector ($28 \times 28 = 784$). You want to use the `KNeighborsClassifier` from sklearn to classify handwritten digits from the MNIST dataset. Complete the following tasks listed below.

  – Load the MNIST dataset. You can refer to the Tutorials covered during lectures.

- Split the dataset into a training set and a test set. Use 60,000 instances for training and the last 10,000 for testing.
- Since K-NN is a distance-based algorithm, normalize the features (pixel intensities) to have a mean of 0 and a standard deviation of 1.
- Train a `KNeighborsClassifier` on the training set with K=3 (i.e., 3 nearest neighbors).
- Predict the digit labels on the test set and calculate the accuracy of the model.

- **Problem 5 [bonus 3pt/3pt]: Grid Search and Hyparameter Tuning.**
Perform a hyperparameter tuning on $K = \{1, 2, 3, 4, 5\}$. Which K yields the best accuracy score during testing? Report the testing accuracy scores for all the 5 possible values for K.