Assignment 4

CMSC 478 – Machine Learning

February 29, 2024

Item	**Summary**
Assigned	Feb 29, Thursday
Due	Mar 7, 11:59 PM
Topic	KMeans, KNN
Points	40

You are to complete this assignment on your own: that is, the code and writeup you submit must be entirely your own. However, you may discuss the assignment at a high level with other students or on the discussion board. Note at the top of your assignment who you discussed this with or what resources you used (beyond course staff, any course materials, or public Campuswire discussions).
Language and External Resources Your code must be compiled, and you will use Python. Failure to do so, or the use of any external resources, including generative AI, and other people's work, without prior written permission from the instructor, will be considered an academic integrity violation and result, at a minimum, in a 0 on this assignment.

1. KNN (20 points):
Install scikit-learn for Python.
Installation help: If you are a Windows user, download the latest WinPython distribution from here. Go through the installation process, and find "IPython QT Console" and run it. That will give you a Python command prompt with all of the packages you need. If you are a Linux user, this Stack Overflow question and answer shows you how to install it.
Once you've got Python and scikit-learn running, execute the following code and submit the output.

```
from sklearn import datasets
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import cross_val_score
import numpy as np


digits = datasets.load_digits()

for n in range(1, 501, 50):
    clf = KNeighborsClassifier(n_neighbors = n)
    scores = cross_val_score(clf, digits.data, digits.target, cv = 10)
    print('k = %d, acc = %.2f' % (n, np.mean(scores)))
```

You'll notice that the cross-validated accuracy scores are decreasing. **Explain why that is happening in a sentence or two.**

2. K-means (20 points): Suppose you have a dataset in which the instances are 1-dimensional. The instances are 1, 2, 4, 5, 10, 11, 12, 25. Run k-means clustering on this dataset for k = 3 with initial centroids on the first 3 instances in the dataset (i.e., 1, 2, 4).

Draw the points on a number line and show which points belong to which clusters for each iteration of k-means. Either point out the clusters or use the $C^i = j$ notation.
Run the algorithm until two consecutive iterations yield the same assignment of points to clusters.

3. Bonus question (20 points): Draw the elbow curve with inertia as the evaluation criteria. Mention the optimal number of clusters for the above problem.