Homework 1: k-Nearest Neighbors

CS412

Released: January 23rd

Due: January 30th, 11:30pm on Gradescope

1 Getting started

Import your data into the language of your choice. In this zip file, there are two folders, R and Python, which contain starter code which should help with this assignment. Also included is the .csv of the data. We will only be using 1's and 5's for this assignment, but I have provided the entire dataset for those of you who are interested in experimenting with categorical classification beyond binary. The first column in the `data.csv` is the value of the digit it represents. In the starter code, I have selected just the 1's and 5's and then separated them into a training and testing set. If you are not using R or Python, or if you wish to use different packages, make sure these steps are finished.

Notice that each row is 256 pixels saved from left to right, top to bottom forming a 16x16 square. Each pixel has an intensity which ranges from -1 to 1, where 1 is the darkest.

2 Drawing your first graph

Once you have the data imported, you will want to select a way to graphically represent the data in two-dimensions. For this, you will need to choose two features that seem to reasonably separate the data. *Don't worry about there being a linear separation, but there should be some inherent clustering.* Good examples for this could be:

- Mean intensity

- Intensity variation

- Vertical symmetry

- Horizontal symmetry

Once you have done this, create a graph of your data points normalized to $[-1, 1]$, so that all of your points fall between -1 and 1 on the x and y axes. You do **not** need to standardize to mean and standard deviation. Mark the 1's as red and the 5's as blue. **Label this Figure 1.1 in your report**. Make sure that there are appropriate axis labels.

3   1-Nearest Neighbor

In this section, color the regions of the graph by which is the closest in your two-dimensional space. Use Euclidean distance. Use the same normalization and axes that you had in Figure 1.1. **Label this Figure 1.2 in your report**.

Answer the following questions in your report:

a) Do you believe that this model suffers from underfitting or overfitting?
   Why or why not?

b) Is the error for this model equivalent to or different from the 1-Nearest Neighbor model in the 256-dimensioned space?
   Explain your answer.

Use 10-fold cross validation to determine the cross validation error for the set under the following conditions and present their error for the following 1-Nearest-Neighbor problems. The 2-dimension problem should be only based on your two dimensions from part 2 and the 256 dimension problem should be on the points in the 256-dimensional image space. **(1c) Comment on any differences you see in the results and what may have resulted in them**. You may need to find or modify your packages to get these values.

a) Euclidean distance - 2 dimensions

b) (EC) Manhattan distance - 2 dimensions

c) (EC) Chebyshev distance - 2 dimensions

d) Euclidean distance - 256 dimensions

e) (EC) Manhattan distance - 256 dimensions

f) (EC) Chebyshev distance - 256 dimensions

4   k-Nearest Neighbor

Consider all of the odd k-Neighbor models between 1-49. Produce a graph of the 10-fold cross validation results for each of the 25 candidates and show their result. The x axis should be k and the y axis should be the E_{cv} Do this for the 256-dimension space. **Label this Figure 1.3** and report the value of k which you think yields the best result. (2a) Explain your answer.

(2b) Give a graph of the 2-dimensional region for your optimal k-Nearest neighbor model and **label this Figure 1.4**. Does this model suffer from overfitting or underfitting? Explain your answer.

(2c) Graduate student question: (EC for undergrad) Provide the estimated error of the 25 models at the 95% confidence level by utilizing the variance of the region. Select your model as that with the lowest 95% upper bound. Does this make a model more likely to overfit or underfit the data? Explain

5   Extra Credit

Find two datasets, one which has a low optimum value for k and one which has a high optimum value for k. Explain what in the data may have led to this result and support your answer with figures as necessary. Additionally, give the E_{cv} for these models. fTo search for potential datasets, search the UCI datasets. Most are available as `.csv`

Making your report

I have included this LaTeX code here to help with compiling your report. You can include images by placing them in the source file and adding:

```
\includegraphics[width=\textwidth]{sample.png}
```

Also, **HERE** is a resource for coding in R. **HERE** is a resource for starting code in Python. Starter code will be posted on Thursday, Jan 24.