CS4780/5780 Homework 1

Due: Tuesday 02/13/2018 11:55pm on Gradescope

Note: For homework, you can work in a team of 5. Please include your teammates' NetIDs and names on the front page and form a group on Gradescope. Also, you are given two late days for this homework.

Problem 1: Train/Test Splits

1. Suppose your boss Chris asks you to develop a Machine Learning application that could identify the phonemes (the basic sound units) despite different speakers. The dataset is a series of recordings on how different people pronounce the 44 phonemes with the following properties:

   - Each recording is associated with only one phoneme and person.
   - The recordings are from 100 people.
   - The data collection is done in a week. In that week, all the participants come in for 5 days (Monday to Friday) to create the recordings.
   - Each person records around 200 phonemes per day.

   Armed with the knowledge you learned in class, you know that you can frame this problem as a supervised learning problem. Devise a valid scheme to split the dataset into training, validation and test sets.

2. After you deploy the application, Kilian says the application is not able to identify phonemes from his voice. In order to develop a phoneme identification system that will work exclusively on him, he is willing to invest an enormous amount of money and sends in 10000 recordings of him pronouncing different phonemes. In this case, how would you split the dataset into training, validation and test set? Please assume you have to include the dataset from problem 1 part 1 when you train your model.

Problem 2: K-nearest Neighbors

1. Consider you have the following 2D dataset:

   - Positive: $\{(1,2),(1,4),(5,4)\}$
   - Negative: $\{(3,1),(3,2)\}$

   Suppose the data comes from the grid $[0,5] \times [0,5]$. Draw the decision boundary for 1-NN classifier with the Euclidean distance.

2. Consider you have the following 2D dataset:

   - Positive: $\{(100, 2), (100, 4), (500, 4)\}$
   - Negative: $\{(300, 1), (300, 2)\}$

   Suppose the data lies in the grid $[0, 500] \times [0, 5]$. If you use a 1-NN classifier (assume Euclidean distance), will the classifier be able to classify $(500, 1)$ as negative? One of the problems with using Euclidean distance is that when we have features of different scales, the Euclidean distance will be dominated by the features that have larger scales. One way to fix this is to scale all the features linearly to $[0, 1]$. Will a 1-NN classifier be able to classify $(500, 1)$ as negative after we scale the features linearly to $[0, 1]$?

3. K-NN can also be used for regression (meaning, your labels are real values now). Suppose you have the following dataset:

$\mathcal{X}$	$\mathcal{Y}$
(0,0)	1
(1,1)	2
(2,3)	3
(3,1)	1
(2,1)	2

   where $\mathcal{X}$ is the feature and $\mathcal{Y}$ is the label. What would be the label for $(0, 1)$ if we use 2-NN with Euclidean distance?

4. Real world datasets often have missing values for certain features. Can we still use K-NN on these datasets? If yes, explain how.

5. Does it take more time to train a K-NN classifier or to apply a K-NN classifier? Explain your reasoning. (Please assume that the data is on the magnitude of millions of points).

6. K-NN classifiers are known to suffer from the curse of dimensionality. However, in class we showed that K-NN actually works on images which are often high dimensional. Explain why.