CS4780/5780 Homework 1

Due: Thursday 09/13/2018 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. (Please answer all questions within each paragraph.)

Problem 1: Train/Test Splits

1. Suppose your boss Robin asks you to develop a machine learning system that can classify images into object categories. The dataset consists of 100,000 real images with the following properties:

   - Each image consists of only one object and is associated with exactly one category.
   - There are 20 categories.
   - There are 5,000 images per category.
   - In real life (after deployment) all categories are equally likely to appear.

   Using your knowledge from CS4780, please frame this task as a supervised learning problem. Describe your setup formally and explain how you would split the data into train/validation/test.

2. Once deployed, a dissatisfied customer (Kilian) points out that five of the categories have a disproportionally high classification error rate. As he is a very nice guy he offers you additional 10,000 images for each one of these five categories, to be added to the original data set and for you to re-train your system. What would go wrong if you simply add these additional images to the original data set and proceed as described in the previous question? What changes would you suggest should be made to the new setup?

Problem 2: K-nearest Neighbors

1. Consider you have the following 2D dataset (with binary class labels) as shown in Figure 1:

   - Class $+1$: $\{(1,2),(1,4),(5,4)\}$
   - Class -1: $\{(3,1),(3,2)\}$

   Suppose the data is strictly confined within the $[0,5]\times[0,5]$ grid. Draw the decision boundary for a 1-NN classifier with Euclidean distance. How would the point $(5,1)$ be classified?

2. Your Finnish friend Aleksandra works with the same data set, however she measured the first coordinate in centimeters instead of meters (the second dimension is unchanged). The data thus becomes:

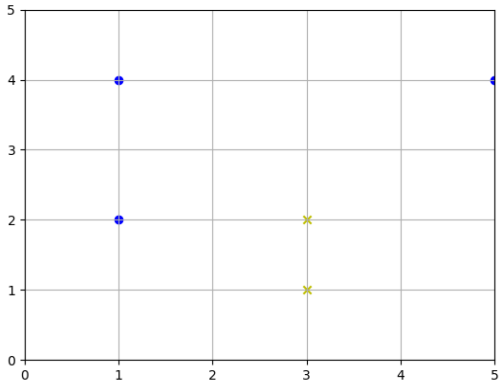

Figure 1: Data points in Problem 2.1. Blue = +1, Yellow = -1

- Class +1: $\{(100, 2), (100, 4), (500, 4)\}$
- Class -1: $\{(300, 1), (300, 2)\}$

Does her 1-NN classifier classify points differently? What will she predict for the original test point $(5, 1)$, which she represents as $(500, 1)$?

3. k-NN can also be used for regression (*i.e.* your labels are real values now). Here, instead of predicting the most common label amongst the neighbors, we predict the label average. Suppose you have the following dataset:

$\mathcal{X}$	$\mathcal{Y}$
(0,0)	1.0
(1,1)	2.5
(2,3)	3.0
(3,1)	1.0
(2,1)	2.5

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

4. In real life it can happen that a test point has some features missing (e.g. because a sensor dropped, or a measurement couldn't be made). Can we still use K-NN in these cases? 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.

Problem 3: Curse of Dimensionality

1. Remember that the volume of a d-dimensional with radius r can be computed as:

$$V_d(r) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d, \tag{1}$$

where $\Gamma()$ is a function independent of r and can be treated as a constant.

   (a) Consider a sphere in $d = 3$ with radius r. Compute what fraction of the volume remains if we decreases the radius by just 1%, i.e. compute $\frac{V_3(0.99r)}{V_3(r)}$.

   (b) Compute the same ratio in $d = 10,000$ (a typical dimension for a machine learning data set).

   What do you conclude about the volume around the surface and inside the interior of high dimensional spheres? (Optional: Close your eyes and try to visualize a sphere in $d = 10,000$. Discuss as a group what it could look like.)