CS4780/5780 Homework 1 Solution

Problem 1: Train/Test Splits

1. We split the training data by person - for example, we can allocate 80% of the images as the training set, then 10% each for the validation and test sets respectively. It is critical that there is no overlap among three datasets, as the different datasets should be independent from one another.

2. If you solely used the additional dataset on the well-trained model, then it would go wrong that there will be bias towards these 5 categories among your results. As the system will be used to identify more objects, we can use all of the original training dataset along with, for example, 80% of additional images to form the new training set - ideally we want roughly equal numbers of images per category. Then, the remaining 20% of the additional images can be split equally and merged into the original test and validation sets. (Other percentage splits are acceptable if reasonable.)

Problem 2: K-nearest Neighbors

1. See figure 1. Blue = positive, Yellow = negative, Red = boundary line.

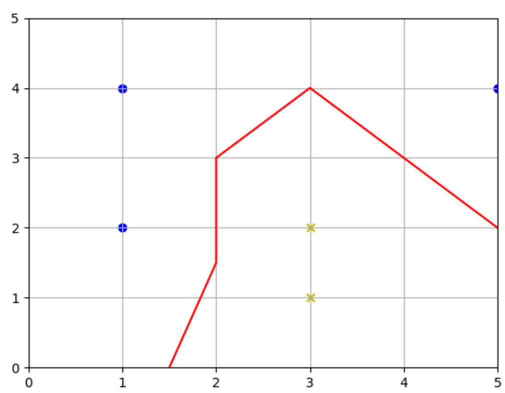

Figure 1: Decision boundary for 1-NN

2. She will classify (500, 1) as +1 with the given 1-NN classifier since the point is closest to the point (500, 4).
   For original test point $(5, 1)$, she will predict -1 because the point is closest to $(3, 1)$

3. Since we are performing 2-NN, the two closest points by Euclidean distance are (0, 0) and (1, 1), each with labels 1.0, 2.5 respectively. Since this is a regression problem, we take the average of these two labels and return 1.75 as our answer.

4. Yes, we can remove those features vectors that have missing values and use K-NN on the new dataset.

5. Unless there is serious preprocessing done on the training data that aids in distance computation, applying a k-NN classifier will take more time: training consists solely of storing the points, while applying must compute the distance between the test point and all stored training points.

6. K-NN still works on images because the underlying latent representation behind images is of low dimension - the curse of dimensionality only takes hold on data with high latent dimensionality.

Problem 3: Curse of Dimensionality

For general d, $\frac{V_d(r_1)}{V_d(r_2)} = \frac{\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r_1^d}{\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r_2^d} = \left(\frac{r_1}{r_2}\right)^d$.

1. $\frac{V_3(0.99r)}{V_3(r)} = (0.99)^3 \approx 0.97$

2. $\frac{V_{10000}(0.99r)}{V_{10000}(r)} = (0.99)^{10000} \approx 2.24 * 10^{-44} \approx 0$

Although the ratio is decreased by only 1%, the relative volume of remaining ball is very tiny when the dimension is high.