CS-233 Theoretical Exercise 1

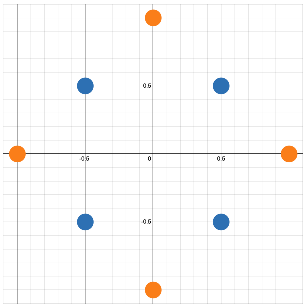
Feburary 2025

1 K-Nearest Neighbors

As seen in the lecture, the notion of nearest neighbors depends upon the distance measure, with popular choices being the L1 and L2 norm. However, why does the choice of norm matter? One way of approaching this question is by understanding the difference between the L1 and L2 norm. Specifically, when do the L1 and L2 norm differ for two points $\mathbf{x}_i$ and $\mathbf{x}_j$?

1. Formally, show that $\|\mathbf{x}_i, \mathbf{x}_j\|_1 \geq \|\mathbf{x}_i, \mathbf{x}_j\|_2$, where $\|\mathbf{x}_i, \mathbf{x}_j\|_p$ is the L-p norm between D-dimensional vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. When is the equality met?

2. Let us assume that we have 8 samples corresponding to 2 classes. The class A samples are located at $(\pm 0.5, \pm 0.5)$, whereas the class B samples are located at $(\pm 1, 0)$ and $(0, \pm 1)$. What would the nearest neighbors be for a sample at any of the locations $(\pm 1, \pm 1)$ according to the L1 and L2 norm? How would the classification of this sample change as we increase the number of nearest neighbors for both the L1 and L2 norm?

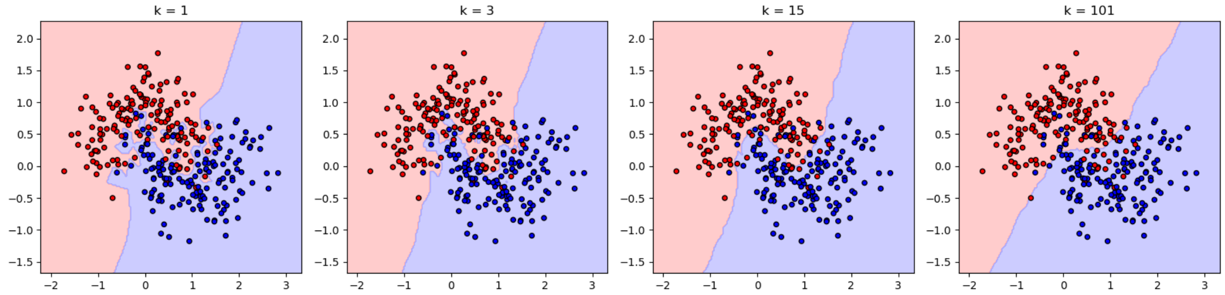


2 The Impact of K in KNN

You are given a dataset with two classes: red and blue. The data is distributed in a 2D space such that there is a region where the two classes are closely mixed. You apply k-nearest neighbors (KNN) with different values of K and observe the following results:
1. Why is K typically chosen to be an odd number?

2. What happens when K is too small?

3. Why does a very large K result in underclassification?



4. How might you choose a good value of K in practice?

3 Preprocessing in KNN

You are applying KNN to a dataset with the following features:

- Age: Ranges from 0 to 100
- Income: Ranges from $0 to $1,000,000
- Binary Gender: Encoded as 0 or 1

1. What preprocessing step is crucial before applying KNN to this dataset and why?

2. What could happen if you skip this step?

4 Data Imbalance in KNN

You are working with a dataset where the minority class represents only 10% of the total samples. When using KNN, you notice that most predictions favor the majority class.

1. What potential challenge might KNN face with this imbalanced dataset?

2. How can you address class imbalance when using KNN?