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?

Solution:

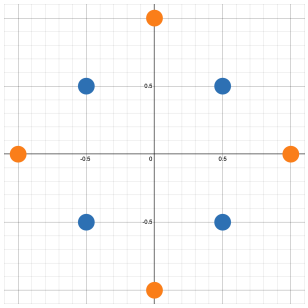
The L1 distance is defined as: $\|\mathbf{x}_i, \mathbf{x}_j\|_1 = \sum_{k=1}^{D} |\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}|$.

The L2 distance is defined as: $\|\mathbf{x}_i, \mathbf{x}_j\|_2 = \sqrt{\sum_{k=1}^{D} (\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)})^2}$.

Since the distance metric is positive, we can compare the squared L1 and L2 distances instead. The squared L1 distance is $(\sum_{k=1}^{D} |\mathbf{x_i}^{(k)} - \mathbf{x}_j^{(k)}|)^2$. Let $a^{(k)}$ represent the k^{th} term $|\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}|$. Then, we can write $(\sum_k a^{(k)})^2 = \sum_k (a^{(k)})^2 + 2\sum_{k=1}^{D} \sum_{\ell<k} a^{(k)} a^{(\ell)}$. Since each term $a^{(k)} \geq 0$, the comparison with the squared L2 distance, which only contains the first summation, yields that $\|\mathbf{x}_i, \mathbf{x}_j\|_1 \geq \|\mathbf{x}_i, \mathbf{x}_j\|_2$.

When is an equality achieved? This occurs only when $\sum_{k=1}^{D} \sum_{\ell<k} a^{(k)} a^{(\ell)} = 0$. The trivial solution to this equation is when $\mathbf{x}_i = \mathbf{x}_j$. However, the equation also admits a solution when at most one $a^{(k)}$ is not zero. This implies that the L2 distance is the same as the L1 distance if the two points differ in at most one coordinate.

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?

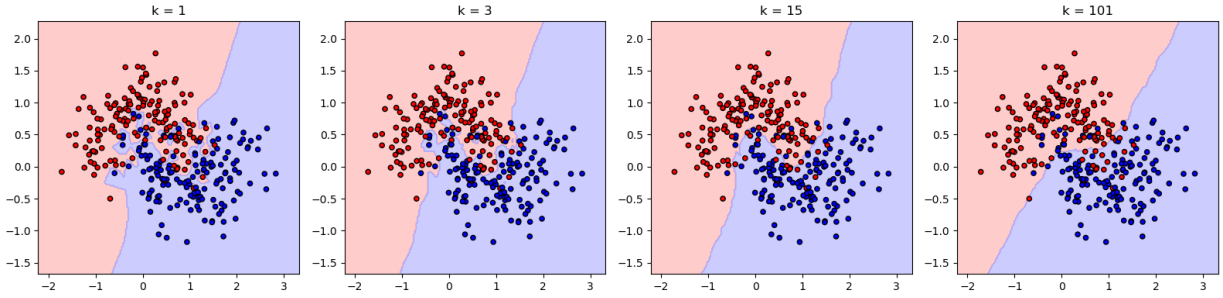


Solution: We first discuss using the L1 norm for K-Nearest Neighbors. For any sample located at $(\pm 1, \pm 1)$, we have three equidistant nearest neighbors. Using majority voting, we assign the sample to class B. On increasing the number of neighbors to 5, we have 3 class A neighbors and 2 class B neighbors. This means the sample belongs to class A. Finally, the number of neighbors increases to 8 with both classes being equally represented. Since the average distance of neighbors belonging to both the classes is the same, the vote is split and there is ambiguity on the class.

We now discuss using the L2 norm. For any sample located at $(\pm 1, \pm 1)$, we have one nearest neighbor belonging to class A. Increasing the nearest neighbors to 3 results in 2 class B and 1 class A neighbor. Increasing the number to 5 results in 3 class A and 2 class B neighbors. However, the classification differs from the L1 case when we increase the number of nearest neighbors to 8. While both classes are equally represented, class A samples have a lower average distance when compared to class B. Therefore, we allocate our sample to class A.

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?

Solution: K is usually chosen as an odd number to avoid ties in classification. If K is even, there might be an equal number of neighbors from different classes, making it ambiguous which class to assign to the query point.

2. What happens when K is too small?

Solution: When K is too small (e.g., K=1), the model is very sensitive to noise and outliers. The model captures the training data well but does not generalize to new data.

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

Solution: A very large K means the model considers a large number of neighbors, potentially including distant points. This smooths the decision boundary too much, ignoring finer details of the dataset. If K is too large, the majority class dominates, and minority class points may get misclassified.

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

Solution: The best K is usually found using cross-validation. A common approach is to test different values of K (e.g., 1, 3, 5, 7, ..., 101) and select the one that minimizes validation error.

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?

Solution: Feature scaling is crucial. Since KNN relies on distance calculations (usually Euclidean), features with large numerical ranges will dominate the distance metric.

2. What could happen if you skip this step?

Solution: If feature scaling is skipped, features like income (which has a much larger range) will have a disproportionate impact on the distance calculation, making features like age and gender almost irrelevant in determining neighbors.

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?

Solution: Since KNN assigns a class based on the majority of its nearest neighbors, a minority class point might be surrounded by majority class points, leading to misclassification. The imbalance makes it harder for KNN to correctly identify the minority class.

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

Solution:

- Weighing neighbors by the inverse of their class size converts neighbor counts into the fraction of each class that falls in your K nearest neighbors.
- Under-sampling the dominant class.
- Augmenting the other class by generating synthetic examples.