k-Nearest Neighbors

1 Hypothesis

Consider the domain $\mathcal{X} = \mathbb{R}^D$ and the codomain $\mathcal{Y} = [1, M]$. In k-Nearest Neighbors classification, our hypothesis $h_{k,d} : \mathcal{X} \to \mathcal{Y}$ is defined by the most common category from the k nearest training points in S determined with the distance function d. As such, k-Nearest Neighbors doesn't require any training (it is non-parametric): it simply classifies new points based on the training dataset S. However, there are hyperparameters that can be tuned, such as picking a distance function d or choosing k itself.

2 Choosing k

Larger values for k smooth out boundaries and increase training error, while smaller values make boundaries more complex and increase training error, possibly inducing overfitting.

3 Distance functions

Below are some popular distance functions that can be used:

- **Euclidean distance:** regular geometric distance. Is given by

$$d_{\text{euc}}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{D} (x^{(i)} - x'^{(i)})^2} = ((\mathbf{x} - \mathbf{x}')^\top (\mathbf{x} - \mathbf{x}'))^{\frac{1}{2}}$$

- **Manhattan distance:** sum of absolute distances of Cartesian coordinates; can be thought of as the distance from navigating city blocks in Manhattan. Is given by

$$d_{\text{man}}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{D} |x^{(i)} - x'^{(i)}|$$

- **Chebyshev distance:** also known as chessboard distance; greatest distance along any coordinate dimension. Is given by

$$d_{\max}(\mathbf{x}, \mathbf{x}') = \max_i(|x^{(i)} - x'^{(i)}|)$$

- **p-norm:** generalization of the previous distances (norms). Is given by

$$d_p(\mathbf{x}, \mathbf{x}') = \left(\sum_{i=1}^{D} |x^{(i)} - x'^{(i)}|^p \right)^{\frac{1}{p}}$$