CS 4780/6780: Fundamentals of Data Science **Spring 2019**

Topic 7: Nearest Neighbors

Instructor: Daniel L. Pimentel-Alarcón

GO GREEN. AVOID PRINTING, OR PRINT DOUBLE-SIDED.

7.1 Introduction

One of the most fundamental problems in data science is classification, which can be summarized as assigning a label (class) $\mathrm{y} \in \{1, 2, \ldots, \mathrm{C}\} =: [\mathrm{C}]$ to a data point $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$. For example:

- Given a data vector $\mathbf{x}$ containing an individual's information (e.g., genome or demographics), determine whether such individual has Alzheimer's ($\mathrm{y} = 1$) or not ($\mathrm{y} = 0$).
- Given a vectorized image $\mathbf{x}$ of a human face, determine which person (y) amongst a database of C individuals is depicted in $\mathbf{x}$.
- Given a vectorized image $\mathbf{x}$ of a character, determine which symbol (y) amongst an alphabet of C corresponds to $\mathbf{x}$.

Nearest neighbors is arguably the simplest classification method. The main idea is to assign each data point to the class of its most *similar* pre-classified points. Since there are several ways to measure similarity, there are also several flavors of nearest neighbors. In all of them, we assume we already have a collection of training data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{N} \in \mathbb{R}^{\mathrm{D}}$ whose corresponding classes $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N} \in [\mathrm{C}]$ are already known. Given a new data point $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$ whose class is unknown, the goal is to determine its corresponding class $\mathrm{y} \in [\mathrm{C}]$.

7.2 Nearest Neighbor

The simplest form of nearest neighbors simply assigns the new point to the class of its *closest* point in our data, that is, $\mathrm{y} = \mathrm{y}_{\hat{\mathrm{i}}}$, where

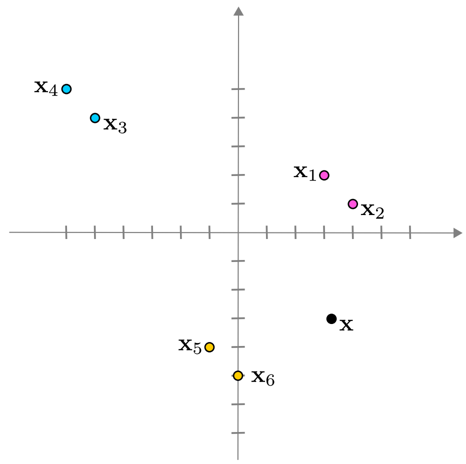
$$\hat{\mathrm{i}} = \underset{\mathrm{i} \in [\mathrm{N}]}{\arg\min} \|\mathbf{x} - \mathbf{x}_\mathrm{i}\|. \tag{7.1}$$

Notice that there are several ways to measure distance, so the norm in (7.1) could be chosen according to the notion of similarity that best suits the application; the most widely used is the euclidean distance, $\|\cdot\|_2$.

Example 7.1. Consider the following training data:

$$\mathbf{x}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 4 \\ 1 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} -5 \\ 4 \end{bmatrix} \quad \mathbf{x}_4 = \begin{bmatrix} -6 \\ 5 \end{bmatrix} \quad \mathbf{x}_5 = \begin{bmatrix} -1 \\ -4 \end{bmatrix} \quad \mathbf{x}_6 = \begin{bmatrix} 0 \\ -5 \end{bmatrix},$$

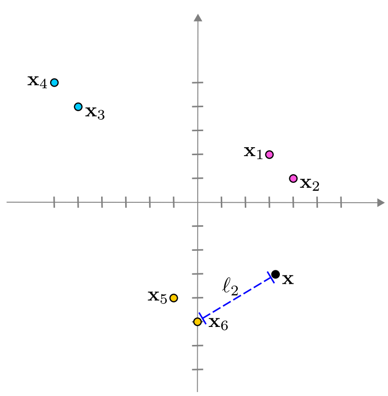
with classes $y_1 = 1$, $y_2 = 1$, $y_3 = 2$, $y_4 = 2$, $y_5 = 3$, $y_6 = 3$ and suppose you observe an additional sample $\mathbf{x} = [3.25 \quad -3]^\mathsf{T}$:

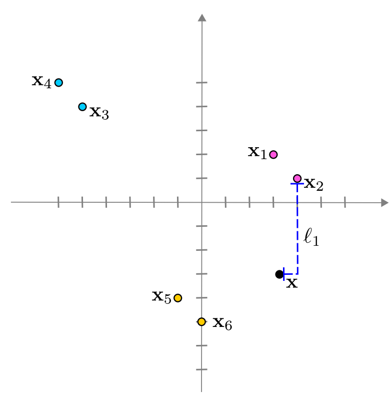


Then $\|\mathbf{x} - \mathbf{x}_1\| = \sqrt{(3.25-3)^2 + (-3-2)^2} \; = \; 5.0062$. Similarly

$$\|\mathbf{x} - \mathbf{x}_2\| = 4.07, \quad \|\mathbf{x} - \mathbf{x}_3\| = 10.82, \quad \|\mathbf{x} - \mathbf{x}_4\| = 12.23, \quad \|\mathbf{x} - \mathbf{x}_5\| = 4.37, \quad \|\mathbf{x} - \mathbf{x}_6\| = 3.82.$$

Since $\mathbf{x}$ is closest to $\mathbf{x}_6$, we assign $\mathbf{x}$ to the same class as $\mathbf{x}_6$, i.e., $y = 3$. Notice, however, that if we use the ℓ_1-norm, then $\|\mathbf{x}-\mathbf{x}_1\| = 5.25$, $\|\mathbf{x}-\mathbf{x}_2\| = 4.75$, $\|\mathbf{x}-\mathbf{x}_3\| = 15.25$, $\|\mathbf{x}-\mathbf{x}_4\| = 17.25$, $\|\mathbf{x}-\mathbf{x}_5\| = 5.25$, and $\|\mathbf{x} - \mathbf{x}_6\| = 5.25$, whence $\mathbf{x}$ is closest to $\mathbf{x}_2$, in which case we conclude $y = 1$:





7.3 K-Nearest Neighbors

Using a single neighbor as reference can be quite sensitive to noise. Hence, to make our nearest neighbor algorithm more robust, we can simply use a *consensus* approach: rather than using a single neighbor, we an use K, and assign $\mathbf{x}$ to the class with the most votes. In this case, the K-*nearest neighbors* algorithm can be summarized as follows:

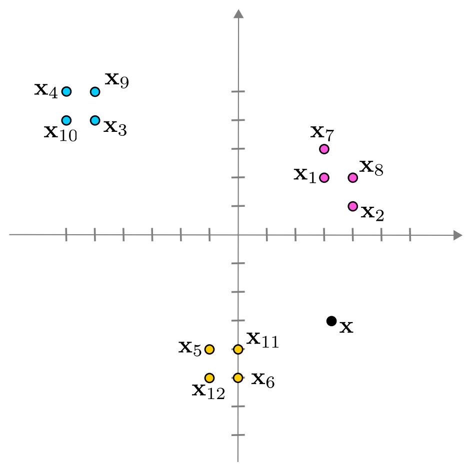
Algorithm 1: K-Nearest Neighbors

Input: Training pairs $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$
Compute distances: For each i, $\delta_i = \|\mathbf{x} - \mathbf{x}_i\|$.
Find K **nearest neighbors:** $\hat{i}_1, \ldots, \hat{i}_K$ = indices of data points that produce K smallest distances δ_i.
Output: Class of new point y = class of the majority amongst $y_{\hat{i}_1}, y_{\hat{i}_2}, \ldots, y_{\hat{i}_K}$.

Example 7.2. In addition to the data in , suppose we also have:

$$\mathbf{x}_7 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \mathbf{x}_8 = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \mathbf{x}_9 = \begin{bmatrix} -5 \\ 5 \end{bmatrix} \quad \mathbf{x}_{10} = \begin{bmatrix} -6 \\ 4 \end{bmatrix} \quad \mathbf{x}_{11} = \begin{bmatrix} 0 \\ -4 \end{bmatrix} \quad \mathbf{x}_{12} = \begin{bmatrix} -1 \\ -5 \end{bmatrix},$$

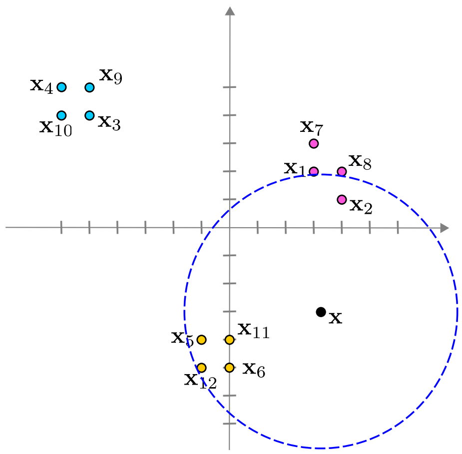
with classes $\{1, 1, 2, 2, 3, 3\}$:

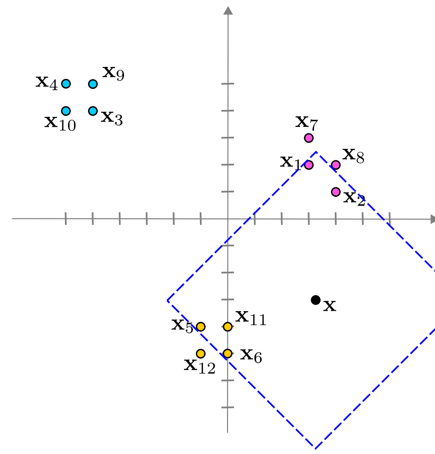


The distances to $\mathbf{x}$ are:

$$\begin{aligned} &\|\mathbf{x} - \mathbf{x}_7\| = 6.01, && \|\mathbf{x} - \mathbf{x}_8\| = 5.06, && \|\mathbf{x} - \mathbf{x}_9\| = 11.49, \\ &\|\mathbf{x} - \mathbf{x}_{10}\| = 11.60, && \|\mathbf{x} - \mathbf{x}_{11}\| = 3.40, && \|\mathbf{x} - \mathbf{x}_{12}\| = 4.6971. \end{aligned}$$

We thus conclude that the closest K = 5 points (with respect to the ℓ_2-norm) are $\{11, 6, 2, 5, 12\}$, which belong to classes $\{3, 3, 1, 3, 3\}$. By consensus, we conclude that $\mathbf{x}$ belongs to class 3. You can verify that if we use the ℓ_1 norm, the closest K = 5 points are $\{11, 2, 1, 5, 6\}$, which belong to classes $\{3, 1, 1, 3, 3\}$, and hence this time we also conclude by consensus that $\mathbf{x}$ belongs to class 3:





Notice that the shape of the ℓ_2-ball (boundary with all equidistant points) is a circle, while the shape of the ℓ_1-ball is a diamond! 😱

Also notice that I chose to use $\mathrm{K} = 5$ neighbors. Why 5? Why not 3, or 6, or any other number?