CS 189: Introduction to Machine Learning - Discussion 8

1. Distance Metric on a set X is defined as a function $d : X \times X \to \Re$ which satisfies the following conditions:

   - $d(x, y) \geq 0 \quad \forall x, y \in X$
   - $d(x, x) = 0$
   - $d(x, y) = d(y, x) \quad \forall x, y \in X$
   - $d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in X$

   Prove the following distances satisfy the conditions.

   a) Euclidean distance $d(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_2 = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ where $\mathbf{p}$ and $\mathbf{q}$ are two n-dimensional real vectors.

   b) Manhattan distance $d(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n}|p_i - q_i|$ where $\mathbf{p}$ and $\mathbf{q}$ are two n-dimensional real vectors.

   c) Jaccard distance $d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$

2. Curse of Dimensionality

   We use 1-NN algorithm to solve a classification problem. The training set contains $(x_1, y_1), \ldots, (x_n, y_n)$. Each x_i is a vector in the d-dimensional space. Each $y_i \in \{-1, 1\}$ is a binary label. Using 1-NN, we classify an unknown point x by

   $$\text{class}(x) = y_{i^*} \quad \text{where } x_{i^*} \text{ is the nearest neighbor of } x.$$

   We know as a prior knowledge that the query point x belongs to the Euclidean ball of radius 1, i.e. $\|x\|_2 \leq 1$. To ensure confident prediction, we also want the distance between x and its nearest neighbour to be small. That is

   $$\|x - x_{i^*}\|_2 \leq \epsilon \quad \text{for all } \|x\|_2 \leq 1. \tag{1}$$

   To make inequality (1) holds, at least how many samples should be in the training set? How does the required sample size depends on the dimension d?