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

Solution: The first three conditions are trivial for all distance metrics. For the triangle inequality of (a), we want to prove $\|\mathbf{x} + \mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$, i.e

$$\sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \leq \sqrt{(\sum_{i=1}^{n} x_i^2)} + \sqrt{(\sum_{i=1}^{n} y_i^2)}$$

Square both sides, you have

$$\sum_{i=1}^{n}(x_i + y_i)^2 \leq \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 + 2\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)}$$

It is equivalent to

$$\sum_{i=1}^{n}(x_i y_i) \leq \sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)}$$

This holds based on Cauchy inequality.

For triangle inequality of (b), we want to prove

$$|\sum_{i=1}^{n}(x_i + y_i)| \leq |\sum_{i=1}^{n} x_i| + |\sum_{i=1}^{n} y_i|$$

This is true since we have $|x_i + y_i| \le |x_i| + |y_i| \quad \forall i$.

For triangle inequality of (c), we want to prove $\forall A, B, C$

$$d(A,C) \le d(A,B) + d(B,C)$$

Assume there exists a counterexample, $\exists A, B, C s.t.$ $d(A,C) > d(A,B) + d(B,C)$. Note that A, C and $A \cap C$ have to be non empty. The left hand keeps unchanging if we change B. If we remove all the elements in B which are not in A or C and get a $B' \subset A \cup C$, then $|A \cap B'| = |A \cap B|$ and $|A \cup B'| \le |A \cup B|$. So we have $d(A,B') \le d(A,B)$ and $d(B',C) \le d(B,C)$ similarly. Then,

$$d(A,C) > d(A,B) + d(B,C) \ge d(A,B') + d(B',C)$$

We now consider remove all the elements in B' that are only in A or C and get $B'' \subset A \cap C$. Similarly, we only decrease the right side $d(A,B') + d(B',C) \ge d(A,B'') + d(B'',C)$.

So we have

$$d(A,B'') + d(B'',C) = 1 - \frac{|B''|}{|A|} + 1 - \frac{|B''|}{|C|} \ge \frac{|A| - |A \cap C|}{|A|} + \frac{|C| - |A \cap C|}{|C|}$$
$$\ge \frac{|A| - |A \cap C|}{|A \cup C|} + \frac{|C| - |A \cap C|}{|A \cup C|} = \frac{|A \cup C| - |A \cap C|}{|A \cup C|} = d(A,C)$$

This is a contradiction!

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?

   Solution: Let B_0 be the ball center at the original, having radius 1. Let $B_i(\epsilon)$ be the ball center at x_i, having radius ϵ. If inequality (1) always holds, then for any point $x \in B_0$, there is at least one index i such that $x \in B_i(\epsilon)$. It means that the union of $B_1(\epsilon), \ldots, B_n(\epsilon)$ covers the ball B_0. Let vol(B) indicates the volume of object B, then we have

   $$n \times \text{vol}(B_1(\epsilon)) = \sum_{i=1}^{n} \text{vol}(B_i(\epsilon)) \geq \text{vol}(\cup_{i=1}^{n} B_i(\epsilon)) \geq \text{vol}(B_0).$$

   It implies

   $$n \geq \frac{\text{vol}(B_0)}{\text{vol}(B_1(\epsilon))} = (1/\epsilon)^d.$$

   This lower bound suggests that to make an accurate prediction on high-dimensional input, we need exponentially many samples in the training set. This exponential dependence is sometimes called the *curse of dimensionality*. It highlights the difficulty of using non-parametric methods for solving high-dimensional problems.