Written Assignment 3

Yushan Liu Student ID: 2024214103

November 16, 2024

All the tasks in this WA is done entirely by MYSELF.

3.1 VC Dimension

The concept class $C = \{c_{a,b} \mid a, b \in \mathbb{R}, a < b\}$ is defined as

$$c_{a,b}(x) = \begin{cases} 1, & x \in [a,b], \\ 0, & \text{otherwise.} \end{cases}$$

To proof VC$dim(C) = 2$, means that the concept class can distinguish all labelings for up to 2 points but cannot distinguish all labelings for 3 points.

First let's proof VC$dim(C) \geq 2$: Let $S = \{x_1, x_2\}$, where $x_1 < x_2$. We verify that the concept class C can distinguish all possible labelings of S. For the 4 possible labelings:

- **Labeling (0, 0)**: Neither point is covered. Choose $c_{a,b}$ such that $b < x_1$.
- **Labeling (1, 0)**: Only x_1 is covered. Choose $c_{a,b}$ such that $a \leq x_1 < b \leq x_2$.
- **Labeling (0, 1)**: Only x_2 is covered. Choose $c_{a,b}$ such that $a > x_1$ and $b \geq x_2$.
- **Labeling (1, 1)**: Both x_1 and x_2 are covered. Choose $c_{a,b}$ such that $a \leq x_1$ and $b \geq x_2$.

For each labeling, we can find a function $c_{a,b} \in C$ that assigns the labels correctly. Hence, C can distinguish all possible labelings of 2 points, so $\text{VCdim}(C) \geq 2$.

Then let's proof VC$dim(C) \leq 2$: Consider $S = \{x_1, x_2, x_3\}$, where $x_1 < x_2 < x_3$. We verify that C cannot distinguish all possible labelings of S. For example, consider the labeling $(1, 0, 1)$, where x_1 and x_3 are covered, but x_2 is not. Since every function $c_{a,b} \in C$ corresponds to a continuous interval $[a, b]$, it is impossible to construct an interval that includes x_1 and x_3 but excludes x_2. Thus, C cannot distinguish all possible labelings of 3 points, so $\text{VCdim}(C) \leq 2$.

Therefore, we have

$$\text{VCdim}(C) = 2$$

.

3.2 Rademacher Complexity

The Rademacher complexity of a function class F over a sample $S = \{x_1, x_2, \ldots, x_m\}$ is defined as:

$$\mathcal{R}_m(F) = \mathbb{E}_\sigma \left[\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right],$$

where σ_i are independent Rademacher random variables ($\sigma_i \in \{-1, +1\}$).

(a)

For $g(x) = af(x) + b \in aF + b$, we compute:

$$\frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i) = \frac{1}{m} \sum_{i=1}^{m} \sigma_i \big(af(x_i) + b\big) = a \cdot \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) + b \cdot \frac{1}{m} \sum_{i=1}^{m} \sigma_i.$$

Since σ_i are symmetric and independent, the expectation of the term involving b vanishes:

$$\mathbb{E}_\sigma\left[b \cdot \frac{1}{m}\sum_{i=1}^{m}\sigma_i\right] = 0.$$

Thus, the Rademacher complexity becomes:

$$\mathcal{R}_m(aF+b) = \mathbb{E}_\sigma\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \cdot af(x_i)\right] = |a| \cdot \mathbb{E}_\sigma\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right].$$

Therefore, we have:

$$\mathcal{R}_m(aF+b) = |a|\mathcal{R}_m(F).$$

(b)

For $l_h(x,y)$, we can rewrite as:

$$l_h(x,y) = \frac{1-h(x)y}{2} = -\frac{1}{2}\cdot h(x)y + \frac{1}{2}.$$

By the result from part (a), the Rademacher complexity of a linearly transformed function class satisfies:

$$\mathcal{R}_m(\mathcal{L}(H)) = \left|-\frac{1}{2}\right|\mathcal{R}_m(H) = \frac{1}{2}\mathcal{R}_m(H).$$

Threrfore, we have:

$$2\mathcal{R}_m(H) = \mathcal{R}_m(\mathcal{L}(H)).$$

3.3 K-means

The objective of the k-means clustering problem is to partition the data into k clusters $C_1, \ldots, C_k$ such that the within-cluster sum of squares is minimized:

$$\min_C \sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2,$$

where μ_j is the mean (center) of the j-th cluster:

$$\mu_j = \frac{1}{|C_j|}\sum_{x\in C_j} x.$$

(a)

We can expand the squared deviation term $\|x-\mu_j\|^2$ for each cluster C_j as follows:

$$\sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2 = \sum_{j=1}^{k}\left(\sum_{x\in C_j}\|x\|^2 - 2\sum_{x\in C_j}x^\top\mu_j + |C_j|\|\mu_j\|^2\right)$$

Consider the second term in Eq:

$$-2\sum_{x\in C_j} x^\top \mu_j = -2\sum_{x\in C_j} x^\top \left(\frac{1}{|C_j|}\sum_{x'\in C_j} x'\right) = \frac{-2}{|C_j|}\sum_{x\in C_j}\sum_{x'\in C_j} x^\top x'$$

Consider the third term in Eq:

$$|C_j|\|\mu_j\|^2 = |C_j|\left\|\frac{1}{|C_j|}\sum_{x'\in C_j} x'\right\|^2 = \frac{1}{|C_j|}\sum_{x'\in C_j}\|x'\|^2$$

We can rewrite the objective function as:

$$\begin{aligned}
&\sum_{j=1}^{k}\left(\sum_{x\in C_j}\|x\|^2 - 2\sum_{x\in C_j} x^\top \mu_j + |C_j|\|\mu_j\|^2\right)\\
=&\sum_{j=1}^{k}\left(\frac{1}{C_j}\sum_{x\in C_j}\sum_{x'\in C_j}\|x\|^2 - 2\frac{1}{|C_j|}\sum_{x,x'\in C_j} x^\top x' + \frac{1}{|C_j|}\sum_{x'\in C_j}\|x'\|^2\right)\\
=&\sum_{j=1}^{k}\frac{1}{C_j}\sum_{x\in C_j}\sum_{x'\in C_j}\left(\|x\|^2 - 2x^\top x' + \|x'\|^2\right) = \sum_{j=1}^{k}\frac{1}{C_j}\sum_{x\in C_j}\sum_{x'\in C_j}\|x-x'\|^2
\end{aligned}$$

Because the calculation $\sum_{x\in C_j}\sum_{x'\in C_j}\|x-x'\|^2$ actually calculates the squared distance for every pair of points in cluster C_j. Since this is a double summation, each pair of points (x, x') is counted twice, so we can write the objective function as:

$$\sum_{j=1}^{k}\frac{1}{|C_j|}\sum_{x\in C_j}\sum_{x'\in C_j}\|x-x'\|^2 = \sum_{j=1}^{k}\frac{1}{2|C_j|}\sum_{x,x'\in C_j}\|x-x'\|^2$$

Therefore, we can know that:

$$\sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2 = \sum_{j=1}^{k}\frac{1}{2|C_j|}\sum_{x,x'\in C_j}\|x-x'\|^2$$

which shows that the k-means clustering problem is equivalent to minimizing the pairwise squared deviation between points in the same cluster.

(b)

We begin with a simplification as follows:

$$S \triangleq \sum_{i=1}^{k}\sum_{j=1}^{k}|C_i||C_j|\|\mu_i-\mu_j\|^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}|C_i||C_j|\left(\|\mu_i\|^2 - 2\mu_i^\top\mu_j + \|\mu_j\|^2\right)$$

Before the proof, we define some notations as follows:

- $m = \sum_{j=1}^{k}|C_j|$ is the total number of data points.
- $\bar{x} = \frac{1}{m}\sum_{x\in X} x$ is the overall mean of the data.

For the first and the third term, we have:

$$\sum_{i=1}^{k}\sum_{j=1}^{k}|C_i||C_j|\|\mu_i\|^2 = \left(\sum_{i=1}^{k}|C_i|\|\mu_i\|^2\right)\left(\sum_{j=1}^{k}|C_j|\right)$$
$$= m\sum_{i=1}^{k}|C_i|\|\mu_i\|^2 = m\sum_{j=1}^{k}|C_j|\|\mu_j\|^2$$

For the second term, we have:

$$-2\sum_{i=1}^{k}\sum_{j=1}^{k}|C_i||C_j|\mu_i^\top\mu_j = -2\sum_{i=1}^{k}|C_i|\mu_i^\top\left(\sum_{j=1}^{k}|C_j|\mu_j\right)$$

Since we have:

$$\sum_{j=1}^{k}|C_j|\mu_j = \sum_{j=1}^{k}\sum_{x\in C_j}x = \sum_{x\in X}x = m\bar{x}.$$

Therefore, we can rewrite the second term as:

$$-2\left(\sum_{i=1}^{k}|C_i|\mu_i^\top\right)m\bar{x} = -2m\bar{x}^\top m\bar{x} = -2m^2\|\bar{x}\|^2.$$

Combine the results above, we have:

$$S = m\sum_{i=1}^{k}|C_i|\|\mu_i\|^2 - 2m^2\|\bar{x}\|^2 + m\sum_{j=1}^{k}|C_j|\|\mu_j\|^2 = 2m\sum_{j=1}^{k}|C_j|\|\mu_j\|^2 - 2m^2\|\bar{x}\|^2$$

Now we define the **Total Sum of Squares, TSS** as:

$$\text{TSS} = \sum_{x\in X}\|x-\bar{x}\|^2$$

The TSS of the data can be devided into two parts:

- **Within-cluster Sum of Squares, WSS**: $\sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2$
- **Between-cluster Sum of Squares, BSS**: $\sum_{j=1}^{k}|C_j|\|\mu_j-\bar{x}\|^2$

We can easily know that:

$$\text{TSS} = \text{WSS} + \text{BSS}$$

Now we find the relationship between S and BSS:
First, we simplify the BSS as follows:

$$\text{BSS} = \sum_{j=1}^{k}|C_j|\left(\|\mu_j\|^2 - 2\mu_j^\top\bar{x} + \|\bar{x}\|^2\right)$$

$$= \sum_{j=1}^{k} |C_j| \|\mu_j\|^2 - 2 \left(\sum_{j=1}^{k} |C_j| \mu_j^\top \right) \bar{x} + \left(\sum_{j=1}^{k} |C_j| \right) \|\bar{x}\|^2$$

as we have proved that $\sum_{j=1}^{k} |C_j| \mu_j = m\bar{x}$, we can know that:

$$\text{BSS} = \sum_{j=1}^{k} |C_j| \|\mu_j\|^2 - 2m\|\bar{x}\|^2 + m\|\bar{x}\|^2 = \sum_{j=1}^{k} |C_j| \|\mu_j\|^2 - m\|\bar{x}\|^2$$

We can rewriting as:

$$\sum_{j=1}^{k} |C_j| \|\mu_j\|^2 = \text{BSS} + m\|\bar{x}\|^2.$$

So, we can know that:

$$S = 2m \left(\text{BSS} + m\|\bar{x}\|^2 \right) - 2m^2\|\bar{x}\|^2 = 2m \cdot \text{BSS} + 2m^2\|\bar{x}\|^2 - 2m^2\|\bar{x}\|^2 = 2m \cdot \text{BSS}$$

This means that maximizing S is equivalent to maximizing the between-cluster sum of squares (BSS), since m is a constant (the total number of data points).

Since the TSS only depends on the data, means that TSS is contant. So minimizing the within-cluster sum of squares (WSS) is equivalent to maximizing the between-cluster sum of squares (BSS).

$$\arg\min_{C} \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2 \quad \Longleftrightarrow \quad \arg\max_{C} \sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j| \|\mu_i - \mu_j\|^2.$$

3.4 Spectral Clustering