MATH524 – Spring 2025

Problem Set: Week 11

1. (**k-NN**) In this question we will explore some of the key steps in proving consistency of the k-NN algorithm (see Theorem 4 from Chapter 7). Recall that the empirical risk for the clustering scheme is defined by

$$R_n(C) = \frac{1}{n}\sum_{i=1}^n \|x_i - C(x_i)\|^2 = \frac{1}{n}\sum_{i=1}^n \min_j \|x_i - c_j\|^2.$$

The nearest neighbour algorithm is chosen such that

$$R_n(C) = \min_{|C|\le k_n} R_n(C).$$

For this question, assume the data x_1^n is fixed. Let $\mathcal{C}_n = \{c_1, \ldots, c_{k_n}\}$ and $\bar{\mathcal{C}}_n = \{\bar{c}_1, \ldots, \bar{c}_{k_n}\}$ denote two sets of cluster centers corresponding to two clustering schemes C and $\bar{C}$.

(a) Show that

$$|R_n(C) - R_n(\bar{C})| \le \max_{i,j}(\|x_i - c_j\| + \|x_i - \bar{c}_j\|)\max_j \|c_j - \bar{c}_j\|.$$

Solution: Plugging in for R_n and simplifying,

$$\begin{aligned}
|R_n(C) - R_n(\bar{C})| &= \left|\frac{1}{n}\sum_{i=1}^n (\min_{j=1,\ldots,k_n}\|x_i - c_j\| - \min_{j=1,\ldots,k_n}\|x_i - \bar{c}_j\|)\right| \\
&\le \frac{1}{n}\sum_{i=1}^n \max_{j=1,\ldots,n}(\|x_i - c_j\| + \|x_i - \bar{c}_j\|)\|c_j - \bar{c}_j\| \\
&\le \max_{i,j}(\|x_i - c_j\| + \|x_i - \bar{c}_j\|)\max_j \|c_j - \bar{c}_j\|.
\end{aligned}$$

This result shows that the empirical risk on a bounded set of cluster centers is a continuous function of the cluster centers.

(b) Show that for the clustering scheme C_n that minimizes R_n over all nearest neighbour clustering schemes with k_n clusters, $R_n(C_n) \to 0$ as $n \to \infty$ for $\mathbb{E}[\|x\|^2] < \infty$.

Hint: *You may wish to use a truncation argument to split the risk function.*

Solution: Let $\{u_1, u_2, \ldots\}$ be a countable dense subset of $\mathbb{R}^d$ with $u_1 = 0$. Choose $L > 0$. Then,

$$\begin{aligned}
R_n(C_n) &\le \frac{1}{n}\sum_{i=1}^n \min_{j=1,\ldots,k_n}\|x_i - u_j\|^2 \\
&\le \frac{1}{n}\sum_{i:x_i\in[-L,L]^d}\min_{j=1,\ldots,k_n}\|x_i - u_j\|^2 + \frac{1}{n}\sum_{i:x_i\in\mathbb{R}^d\setminus[-L,L]^d}\min_{j=1,\ldots,k_n}\|x_i - u_j\|^2 \\
&\le \max_{x\in[-L,L]^d}\min_{j=1,\ldots,k_n}\|x - u_j\|^2 + \frac{1}{n}\sum_{i:x_i\in\mathbb{R}^d\setminus[-L,L]^d}\|x_i\|^2 \\
&\to 0 + \mathbb{E}[\|x\|^2\mathbf{1}(x\in\mathbb{R}^d\setminus[-L,L]^d)].
\end{aligned}$$

Since we assume $\mathbb{E}[\|x\|^2] < \infty$, letting $L \to \infty$, $\mathbb{E}[\|x\|^2\mathbf{1}(x\in\mathbb{R}^d\setminus[-L,L]^d)] \to 0$ and the result follows.

(c) Show that if Π_n is the collection of all partitions induced by the k_n nearest enighbour clustering scheme, $M(\Pi_n) = k_n$.

Solution: Every k-nearest neighbour clustering scheme generates exactly k cluster centers that partition all the data. Thus, we directly see that $M(\Pi_n) = k$.

(d) Show that for the k_n-NN partitioning scheme, $\Delta(\Pi_n) \leq (n+1)^{(d+1)k_n^2}$

Solution: Putting together the results of Theorem 8 and Theorem 9 from Chapter 5, $\Delta(x_1^n, \Pi_n) \leq (n+1)^{(d+1)}$. Thus, by definition and the fact that each k_n-NN partition is the intersections of at most k_n^2 hyperplanes perpendicular to one of the k_n^2 pairs of cluster centers, $\Delta(\Pi_n) \leq \left((n+1)^{(d+1)}\right)^{k_n^2}$.

2. (**Packing and covering numbers (Chapter 5, Lemma 2)**) $\mathcal{F} = \{f \in \mathbb{R}^d\}$ and ν is a probability measure with $p \geq 1$, $\varepsilon > 0$. Then,

$$\mathcal{M}(2\varepsilon, \mathcal{F}, L_p(\nu)) \leq N(\varepsilon, \mathcal{F}, L_p(\nu)) \leq \mathcal{M}(\varepsilon, \mathcal{F}, L_p(\nu))$$

where $\mathcal{M}$ is the packing number and N is the covering number.

Solution: Let $f_1, \dots, f_l$ be a 2ε-packings of $\mathcal{F}$ w.r.t $L_p(\nu)$. Any set

$$U_\varepsilon(g) = \{h : \mathbb{R}^d \to \mathbb{R} : \|h - g\|_{L_p(\nu)} < \varepsilon\}$$

constains at most one f_i from the packing. This directly implies the first inequality. For the second inequality, we assume $\mathcal{M}(\varepsilon, \mathcal{F}, L_p(\nu)) < \infty$ since otherwise the proof is trivial. Now, let $g_1, \dots, g_l$ be an ε-packing of $\mathcal{F}$ of size $l = \mathcal{M}(\varepsilon, \mathcal{F}, L_p(\nu))$. Letting $h \in \mathcal{F}$ be an arbitrary function, $\{h, g_1, \dots, g_l\}$ is a subset of $\mathcal{F}$ of size $l+1$ and so it cannot by an ε-packing of $\mathcal{F}$. Thus, there exists a $j \in \{1, \dots, l\}$ such that

$$\|h - g_j\|_{L_p(\nu)} < \varepsilon.$$

This means that $\{g_1, \dots, g_l\}$ is an ε-cover of $\mathcal{F}$ and so the second inequality follows.