ST510: Foundations of Machine Learning
Exercise Sheet (with Solutions)

Exercise 1. Let $\{\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top\}_{i=1}^n$ be i.i.d. and $\mathbb{P}(\|\mathbf{X}_i\| \leq B) = 1$ for some $B < \infty$. For k-means clustering, see definitions of $R(\widehat{C})$ and $R(C^*)$ in lecture slides. Prove that

$$\mathbb{E}(R(\widehat{C})) - R(C^*) \lesssim \sqrt{\frac{k(p+1)\log(2n)}{2n}}.$$

Proof. Note that

$$\begin{aligned}
R(\widehat{C}) - R(C^*) &= R(\widehat{C}) - R_n(\widehat{C}) + R_n(\widehat{C}) - R(C^*) \\
&\leq R(\widehat{C}) - R_n(\widehat{C}) + R_n(C^*) - R(C^*) \\
&\leq 2 \sup_{C \in \mathcal{C}_h} |R(C) - R_n(C)|.
\end{aligned}$$

For each C, define a function $f_C(x) = \|x - \Pi_C(x)\|^2$. Then $\sup_x |f_C(x)| \leq 4B$ for all C. By the fact that $\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y \geq t)dt$ whenever $Y \geq 0$, we have

$$\begin{aligned}
2 \sup_{C \in \mathcal{C}_h} |R(C) - R_n(C)| &\leq 2 \sup_{C \in \mathcal{C}_h} \left| \frac{1}{n} \sum_{i=1}^n f_C(\mathbf{X}_i) - \mathbb{E}(f_C(\mathbf{X}_i)) \right| \\
&= 2 \sup_{C \in \mathcal{C}_k} \left| \int_0^\infty \left(\frac{1}{n} \sum_{i=1}^n I(f_C(\mathbf{X}_i) > u) - \mathbb{P}(f_C(\mathbf{X}_i) > u) \right) du \right| \\
&\leq 8B \sup_{C \in \mathcal{C}_k, u \in [0, 4B]} \left| \frac{1}{n} \sum_{i=1}^n I(f_C(\mathbf{X}_i) > u) - \mathbb{P}(f_C(\mathbf{X}_i) > u) \right| \\
&= 8B \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i; C, u) - \mathbb{E}[h(\mathbf{X}_i; C, u)] \right|
\end{aligned}$$

where $\mathcal{H} := \{h(x; C, u) = I(f_C(x) > u) | \forall C, u\}$.

The growth function of $\mathcal{H}$ satisfies $\tau_{\mathcal{H}}(n) \lesssim n^{k(p+1)}$ because

- $h(x; C, u) = I(f_C(x) > u) = \prod_{j=1}^{k} I(f_{\{c_j\}}(x) > u)$
- Production of hypothesis classes have sub-multiplicative growth: for hypothesis classes $\mathcal{H}_1, \mathcal{H}_2$ and their intersection $\mathcal{H}_1 \otimes \mathcal{H}_2 := \{h_1 \times h_2 \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$ we have that

$$\tau_{\mathcal{H}_1 \otimes \mathcal{H}_2}(n) \leq \tau_{\mathcal{H}_1}(n) \times \tau_{\mathcal{H}_2}(n)$$

- The hypothesis class $\mathcal{H}_j$ of a ball in $\mathbb{R}^p$ has growth function $\tau_{\mathcal{H}}(n) \leq n^{p+1}$ because:
  - No set of $p+2$ points can be shattered (which is an easy consequence of Radon's theorem and the hyperplane separation theorem for convex sets), and thus the VC dimension is at most $p+1$.
  - By Sauers's lemma, $\tau_{\mathcal{H}_j}(n) = O(n^{p+1})$.

According to the result of generalization gap (Equation (6.4) in [1]), we have:

$$\mathbb{E} \sum_{C \in \mathcal{C}_h} |R(C) - R_n(C)| \leq \frac{4 + \sqrt{\tau_{\mathcal{H}}(2n)}}{\sqrt{2n}} \lesssim \sqrt{\frac{k(p+1)\log(n)}{2n}},$$

we can derive the result.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.