CMSC 25025 / STAT 37601: Machine Learning and Large Scale Data Analysis

Assignment 5 Sample Solutions

1. *k-Means Business* (50 points)

   (a) The optimal population quantization C^* minimizes $R(C)$. Show that $R(\widehat{C})$ being close to $R(C^*)$ does not imply that $\widehat{C}$ is close to C^*.

   Solution:

   Consider this distribution: $\mathbb{P}(x = -1) = \mathbb{P}(x = 0) = \mathbb{P}(x = 1) = \frac{1}{3}$. Then, when $K = 2$, $C^* = \{-\frac{1}{2}, 1\}$ is an optimal population quantization. Meanwhile, $\widehat{C} = \{-1, \frac{1}{2}\}$ gives the same risk. But $\widehat{C}$ is not close to C^*.

   (b) Let $R^{(k)}$ denote the minimal risk among all possible clusterings with k clusters. Show that $R^{(k)}$ is nonincreasing in k.

   Solution:

   Let $\mathcal{C}_k$ denote all codebooks of length k. Suppose C_k^* is the optimal codebook that gives the minimum risk $R^{(k)}$. Let c_{k+1} be an arbitrary point in the sample space. Then,

$$
\begin{aligned}
R^{(k+1)} &= \min_{C \in \mathcal{C}_{k+1}} R(C) \le R(C_k^* \cup \{c_{k+1}\}) = \mathbb{E} \min_{c_j \in C_k^* \cup \{c_{k+1}\}} \|X - c_j\|^2 \\
&\le \mathbb{E} \min_{c_j \in C_k^*} \|X - c_j\|^2 = R^{(k)}
\end{aligned}
$$

   Therefore, $R^{(k)}$ is nonincreasing in k.

   (c) Show that, under appropriate conditions on the distribution P from which the random variables X_i are drawn, the optimal k-means risk satisfies $R^{(k)} \to 0$ as $k \to \infty$.

   Proof:

   Denote the pdf of distribution P as $f(x)$. Assume P has a finite variance, σ^2. i.e. $\int_\Omega \|x - \mu\|^2 f(x)\, dx = \sigma^2 < \infty$, where $\mu = \mathbb{E}X$ and Ω is the sample space. Consider such a sequence of spherical balls centered at μ: $B_n = B(\mu, n)$. Then, we have $\lim_{n\to\infty} \int_{B_n} \|x - \mu\|^2 f(x)\, dx \to \sigma^2$. In another word, for any given $\epsilon > 0$, $\exists n_\epsilon$, s.t. $\int_{\Omega \setminus B_{n_\epsilon}} \|x - \mu\|^2 f(x)\, dx < \epsilon/2$.

   Let C_k be a set of k centers including μ.

$$
\begin{aligned}
R(C_k) &= \int_\Omega \min_{c \in C_k} \|x - c\|^2 f(x)\, dx \\
&= \int_{B_{n_\epsilon}} \min_{c \in C_k} \|x - c\|^2 f(x)\, dx + \int_{\Omega \setminus B_{n_\epsilon}} \min_{c \in C_k} \|x - c\|^2 f(x)\, dx \\
&\le \int_{B_{n_\epsilon}} \min_{c \in C_k} \|x - c\|^2 f(x)\, dx + \int_{\Omega \setminus B_{n_\epsilon}} \|x - \mu\|^2 f(x)\, dx \\
&= I + II
\end{aligned}
$$

We already know $II < \epsilon/2$. By noticing that I integrates over B_{n_ϵ} which is a bounded area, we can add finite number of cluster centers to C_k to form a grid inside B_{n_ϵ} with grid size $\epsilon/4$. This could ensure that, for any $x \in B_{n_\epsilon}$, there exists a cluster center that is within a distance of $\epsilon/2$ to x. Denote this enhanced set of cluster centers as C_ϵ and suppose it has size k_ϵ. Then, $I = \int_{B_{n_\epsilon}} \min_{c \in C_\epsilon} \|x - c\|^2 f(x)\, dx < \epsilon/2$ and, further, $R(C_\epsilon) < \epsilon$.

Since for any ϵ, we may create such a finite set C_ϵ and $R^{(k_\epsilon)} \leq R(C_\epsilon) < \epsilon$, it must be true that $R^{(k)} \to 0$ as $k \to \infty$.

Please notice that, if the distribution does not have a finite variance, the result may not be true. For example, consider a Cauchy distribution with parameter x_0 and γ. It has pdf $f(x) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$, $x \in R$. In this case, any finite set of cluster centers $\bar{C}_k = \{c_1, c_2, ..., c_k\}$ would have population clustering risk:

$$\begin{aligned}
R(\bar{C}_k) &= \int_R \min_{c \in \bar{C}_k} \|x - c\|^2 f(x)\, dx \\
&> \int_{c_{(k)}}^{\infty} (x - c_{(k)})^2 f(x)\, dx \\
&= \int_{c_{(k)}}^{\infty} \frac{(x - c_{(k)})^2}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}\, dx \\
&= \infty
\end{aligned}$$

where $c_{(k)} = \max\{c_1, c_2, ..., c_k\}$. Since this holds for any finite codebook C, $R^{(k)}$ is unbounded regardless of the choice of k.