CS 189 Introduction to Machine Learning
Fall 2017 DIS14

1 Kernel k-Means

Suppose we have a dataset $\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^n$ that we want to split into K clusters. Furthermore, suppose we know a priori that this data is best clustered in a large feature space $\mathbb{R}^m$, and that we have a feature map $\phi : \mathbb{R}^n \to \mathbb{R}^m$. How should we perform clustering in this space?

(a) Write the objective for K-means clustering in the feature space (using the squared L_2 norm in the feature space). Do so by explicitly constructing cluster centers $\{\mu_k\}_{k=1}^K$ with all $\mu_k \in \mathbb{R}^m$.

Solution:

$$L = \sum_{k=1}^{K} \sum_{x_i \in S_k} \|\phi(x_i) - \mu_k\|^2$$

(b) Write an algorithm that minimizes the objective in (a).

Solution: 1. Compute $\phi(x_i)$ for every point x_i.

2. Do the standard k-means on $\{\phi(x_i)\}$.

(c) Write an algorithm that minimizes the objective in (a) without explicitly constructing the cluster centers $\{\mu_k\}$. Assume you are given a kernel function $\kappa(x,y) = \phi(x) \cdot \phi(y)$.

Solution: We proceed by coordinate descent on the objective in (a). First, given a clustering, the setting of μ_i that minimizes L is

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} \phi(x)$$

Second, given a setting of the μ's, the optimal clustering is given by assigning x_i to the cluster $\arg\min_{1 \le k \le K} f(i,k)$, where

$$f(i,k) = \|\phi(x_i) - \mu_k\|^2$$

To kernelize this, we write

$$f(i,k) = \phi(x_i) \cdot \phi(x_i) - 2\phi(x_i) \cdot \mu_k + \mu_k \cdot \mu_k$$

Substituting the setting of μ_k,

$$= \phi(x_i) \cdot \phi(x_i) - \frac{2}{|S_k|} \sum_{x_j \in S_k} \phi(x_i) \cdot \phi(x_j) + \frac{1}{|S_k|^2} \sum_{x_j, x_l \in S_k \times S_k} \phi(x_j) \cdot \phi(x_l)$$

Now we can replace the inner products with kernel evaluations

$$= \kappa(x_i, x_i) - \frac{2}{|S_k|} \sum_{x_j \in S_k} \kappa(x_i, x_j) + \frac{1}{|S_k|^2} \sum_{x_j, x_l \in S_k} \kappa(x_j, x_l)$$

This yields the following algorithm:

(a) Compute the kernel matrix $G_{ij} = \kappa(x_i, x_j)$.

(b) Start with an initial clustering $\{S_k\}$.

(c) Compute the new cluster index for each x_i as $\arg\min_{1 \le k \le K} f(i,k)$.

(d) Update $\{S_k\}$

(e) Repeat steps (3) and (4) until convergence.