CS 189 Introduction to Machine Learning
Spring 2018 DIS14

1 Regularized and Kernel k-Means

Recall that in k-means clustering we are attempting to minimize an objective defined as follows:

$$\min_{C_1, C_2, \ldots, C_k} \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2, \text{ where}$$

$$\mu_i = \text{argmin}_{\mu_i \in \mathbb{R}^d} \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2 = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \quad i = 1, 2, \ldots, k.$$

The samples are $\{x_1, \ldots, x_n\}$, where $x_j \in \mathbb{R}^d$, and C_i is the set of samples assigned to cluster i. Each sample is assigned to exactly one cluster, and clusters are non-empty.

(a) **What is the minimum value of the objective when $k = n$ (the number of clusters equals the number of samples)?**

Solution: The value is 0 since every point can get its own cluster.

(b) (Regularized k-means) Suppose we add a regularization term to the above objective. That is, the objective now becomes

$$\sum_{i=1}^{k} \left(\lambda \|\mu_i\|_2^2 + \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2 \right).$$

Show that the optimum of

$$\min_{\mu_i \in \mathbb{R}^d} \lambda \|\mu_i\|_2^2 + \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2$$

is obtained at $\mu_i = \frac{1}{|C_i| + \lambda} \sum_{x_j \in C_i} x_j$.

Solution:

Taking gradient with respect to μ_i for the function

$$f(\mu_j) = (\sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2) + \lambda \|\mu_i\|_2^2,$$

we get

$$\nabla_{\mu_j} f(\mu_j) = (2 \sum_{x_j \in C_i} (\mu_i - x_j)) + 2\lambda \mu_i$$

$$= 2((|C_i| + \lambda)\mu_i - \sum_{x_j \in C_i} x_j).$$

Setting it to zero, we get $\mu_i = \frac{1}{|C_i|+\lambda} \sum_{x_j \in C_i} x_j$. As the function f is convex, the minimum is obtained at $\mu_i = \frac{1}{|C_i|+\lambda} \sum_{x_j \in C_i} x_j$.

(c) Here is an example where we would want to regularize clusters. Suppose there are n students who live in a $\mathbb{R}^2$ Euclidean world and who wish to share rides efficiently to Berkeley for their final exam in CS189. The university permits k vehicles which may be used for shuttling students to the exam location. The students need to figure out k good locations to meet up. The students will then walk to the closest meet up point and then the shuttles will deliver them to the exam location. Define x_j to be the location of student j, and let the exam location be at $(0,0)$. Assume that we can drive as the crow flies, i.e. by taking the shortest paths between two points. **Write down an appropriate objective function to minimize the total distance that the students and vehicles need to travel.** Your result should be similar to the regularized k-means.

Solution:

The objective function that minimize total distance that the students and vehicles need to travel is

$$\min_{\mu_i \in \mathbb{R}^d} \sum_{i=1}^{k} \left(\|\mu_i\|_2 + \sum_{x_j \in C_i} \|x_j - \mu_j\|_2 \right). \tag{1}$$

Cathy Wu, Ece Kamar, Eric Horvitz. *Clustering for Set Partitioning with a Case Study in Ridesharing*. IEEE Intelligent Transportation Systems Conference (ITSC), 2016.

(d) (Kernel k-means) Suppose we have a dataset $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^\ell$ that we want to split into k clusters, i.e., finding the best k-means clustering *without the regularization*. Furthermore, suppose we know *a priori* that this data is best clustered in an impractically high-dimensional feature space $\mathbb{R}^m$ with an appropriate metric. Fortunately, instead of having to deal with the (implicit) feature map $\phi : \mathbb{R}^\ell \to \mathbb{R}^m$ and (implicit) distance metric[1], we have a kernel function $\kappa(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle$ that we can compute easily on the raw samples. How should we perform the kernelized counterpart of k-means clustering?

[1] Just as how the interpretation of kernels in kernelized ridge regression involves an implicit prior/regularizer as well as an implicit feature space, we can think of kernels as generally inducing an implicit distance metric as well. Think of how you would represent the squared distance between two points in terms of pairwise inner products and operations on them.

Derive the underlined portion of this algorithm.

```
Algorithm 1: Kernel K-means
  Require: Data matrix $X \in \mathbb{R}^{n\times d}$; Number of clusters $K$; kernel function $\kappa(\mathbf{x}_1, \mathbf{x}_2)$
  Ensure: Cluster class $\mathrm{class}(j)$ for each sample $x_j$.
    function KERNEL-K-MEANS($X, c$)
        Randomly initialize $\mathrm{class}(j)$ to be an integer in $1, 2, \ldots, K$ for each $x_j$.
        while not converged do
            for $i \leftarrow 1$ to $K$ do
                Set $S_i = \{j \in \{1, 2, \ldots, n\} : \mathrm{class}(j) = i\}$.
            for $i \leftarrow 1$ to $n$ do
                Set $\mathrm{class}(i) = \mathrm{argmin}_k$ ______________________________
        Return $S_i$ for $i = 1, 2, \ldots, c$.
    end function
```

(Hint: there will be no explicit representation of the "means" $\boldsymbol{\mu}_\mathbf{i}$, instead each cluster's membership itself will implicitly define the relevant quantity, in keeping with the general spirit of kernelization that we've seen elsewhere as well.)

Solution: First, given a clustering S_i, we will put

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

to minimize $\sum_{x \in S_i} \|\phi(x) - \mu_i\|_2^2$.

Second, given a setting of the μ's, the optimal clustering is given by assigning x_i to the cluster $\arg\min_k f(i,k)$, where

$$\begin{aligned} f(i,k) &= \|\phi(x_i) - \mu_k\|^2 \\ &= \langle \phi(x_i), \phi(x_i) \rangle - 2\langle \phi(x_i), \mu_k \rangle + \langle \mu_k, \mu_k \rangle \\ &= \langle \phi(x_i), \phi(x_i) \rangle - \frac{2}{|S_k|}\sum_{x_j \in S_k} \langle \phi(x_i), \phi(x_j) \rangle + \frac{1}{|S_k|^2}\sum_{x_j, x_l \in S_k \times S_k} \langle \phi(x_j), \phi(x_l) \rangle \\ &= \kappa(x_i, x_i) - \frac{2}{|S_k|}\sum_{x_j \in S_k} \kappa(x_i, x_j) + \frac{1}{|S_k|^2}\sum_{x_j, x_l \in S_k} \kappa(x_j, x_l). \end{aligned}$$

Therefore, we should write

$$\mathrm{class}(i) = \arg\min_k \frac{1}{|S_k|^2}\sum_{x_j, x_l \in S_k} \kappa(x_j, x_l) - \frac{2}{|S_k|}\sum_{x_j \in S_k} \kappa(x_i, x_j). \tag{2}$$