Kernel K-means

Saturday, March 11, 2017 10:28 AM

Mercer's Theorem. Let $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be an inner product on the Hilbert space $\mathcal{H}$. There exists a map $\varphi: \mathcal{X} \to \mathcal{H}$ such that

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$$

if and only if for any square integrable function $g(x)$, i.e. $\int |g(x)|^2 dx < \infty$, the following holds:

$$\iint dx\, dy\, g(x)\, k(x, y)\, g(y) \geq 0.$$

$$\|\varphi(x) - \varphi(y)\|_{\mathcal{H}}^2 = (\varphi(x) - \varphi(y))^T (\varphi(x) - \varphi(y))$$
$$= \varphi(x)^T \varphi(x) - 2\varphi(x)^T \varphi(y) + \varphi(y)^T \varphi(y)$$
$$= k(x, x) - 2k(x, y) + k(y, y)$$

$\uparrow$ $D_{\mathcal{H}}(x, y)$

K-means $E = \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ki} \|x_i - m_k\|^2$

where $z_{ki} = \begin{cases} 1 & \text{if } x_i \in C_k \\ 0 & \text{otherwise} \end{cases}$, $m_k = \frac{\sum_{i=1}^{n} z_{ki} x_i}{\sum_{i=1}^{n} z_{ki}} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$

Kernel k-means do the same thing but on $\mathcal{H}$.

$$E = \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ki} \| \varphi(x_i) - m_k \|_{\mathcal{H}}^2$$

$$m_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} \varphi(x_i) = \frac{\sum_{i=1}^{n} z_{ki} \varphi(x_i)}{\sum_{i=1}^{n} z_{ki}}.$$

$$\| \varphi(x_i) - m_k \|_{\mathcal{H}}^2 = \left\| \varphi(x_i) - \frac{1}{n_k} \sum_{\ell=1}^{n} z_{k\ell} \varphi(x_\ell) \right\|_{\mathcal{H}}^2$$

$$= \varphi(x_i)^T \varphi(x_i) - 2 \varphi(x_i)^T \frac{1}{n_k} \sum_{\ell=1}^{n} z_{k\ell} \varphi(x_\ell) + \frac{1}{n_k^2} \sum_{\ell=1}^{n} \sum_{p=1}^{n} z_{k\ell} z_{kp} \varphi(x_\ell)^T \varphi(x_p)$$

Mercer's theo.

$$= k(x_i, x_i) - \frac{2}{n_k} \sum_{\ell=1}^{n} k(x_i, x_\ell) + \frac{1}{n_k^2} \sum_{\ell=1}^{n} \sum_{p=1}^{n} z_{k\ell} z_{kp} k(x_\ell, x_p)$$

Thus the kernel k-mean objective functio is

$$E = \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ki} \left(k(x_i, x_i) - \frac{2}{n_k} \sum_{\ell=1}^{n} z_{k\ell} k(x_i, x_\ell) + \frac{1}{n_k^2} \sum_{\ell=1}^{n} \sum_{p=1}^{n} z_{k\ell} z_{kp} k(x_\ell, x_p) \right)$$

where $n_k = \sum_{i=1}^{n} z_{ki}$. Only depends on the Gram matrix $K_{ij} = k(x_i, x_j)$

We will see that this can be written as

$$Tr(GKG^T) \quad \text{kernel k-means.}$$

Spectral clustering is a weighted kernel k-means

$$T(G D^{-1/2} K D^{-1/2} G^T)$$

Problems:

1. k-means: $O(n \times d)$
   kernel k-means: $O(n^2)$
2. In a sufficiently high dimensional space, every data can be separable. However, a choice of kernel is data/application dependent, and usually introduces new parameters.
3. The wrong choice of kernel can lead to poor results, even worse than k-means.

Options for scalability issues:

1. Reordered Kernel k-means (distributed, does not improve complexity)
2. Sampling based approximation (kernel matrix approx, random projections)

Approximating the kernel:

Sample $m \ll n$ points $\rightarrow \{y_1, \ldots, y_m\}$

Compute $\begin{cases} K_A \in \mathbb{R}^{m \times m}, (K_A)_{ij} = \varphi(y_i)^T \varphi(y_j) \\ K_B \in \mathbb{R}^{n \times m}, (K_B)_{ij} = \varphi(x_i)^T \varphi(y_j) \end{cases}$

$$m_k \approx \sum_{j=1}^{m} \alpha_{kj} \varphi(y_j)$$

$$\min_{z} \max_{\alpha} \frac{1}{2} \sum_{k=1}^{k} \sum_{i=1}^{n} z_{ij} \left\| \varphi(x_i) - \sum_{j=1}^{m} \alpha_{kj} \varphi(y_j) \right\|^2$$

- This is equivalent to run k-means on $K_B K_A^{-1} K_B^T$ (Nystrom approx.)
- $O(n \cdot m)$ almost linear on n.
- $O(\frac{1}{m})$ approx. error.

Methods:
- Nystrom. Randomly select columns $K = K_B K_A^{-1} K_B^T$
- CUR. Select k columns C and rows R. Find U s.t $\min \| K - CUR \|_F$

- Find N such that
  $\hat{K} = K + N$
  is sparse, and $\mathbb{E}[N_{ij}] = 0$ and $\mathrm{var}[N_{ij}]$ is small

Kernel PCA Project data into the first C eigenvectors of K, cluster on the eigenspace.

Nonlinear Random Projections

Pick a random matrix $R \in \mathbb{R}^{d \times r}$.
Transform data from the Hilbert space:

$$x' = \frac{1}{\sqrt{r}} R^T x, \quad y' = \frac{1}{\sqrt{r}} R^T y$$

Johnson-Lindenstrauss theorem

$$(1-\varepsilon)\|x-y\|^2 \leq \|x'-y'\|^2 \leq (1+\varepsilon)\|x-y\|^2$$

Provided R is orthogonal on r is large enough. This means that distances are "almost" preserved.

original space (n) → (data not separable) → Hilbert space ($m > n$) → (data is separable) → Random projection space ($r \ll m$) → (preserve separati) → Do k-means in the projected space

Unsupervised Nonparametric Kernel Learning Algorithm. Liu et al (2013)

$$\min_{K,U} \operatorname{Tr}(KL) + \frac{1}{2}U\left(K + \frac{\gamma}{c}KL + \frac{I}{c}\right)U^T$$

s.t. $K \succeq 0$ SDP

$\operatorname{Tr}(K^p) \leq B$ } avoid overfitting

$-l \leq n_p - n_q \leq l$ } and all points in the same cluster

$L = I - D^{-1/2} W D^{-1/2}$ (Laplacian)

This matrix is the spectrum of the data after regularization.

$O(n^2)$ to $O(n^3)$ complexity.

Refs

Kernel Theory

[1] Statistical Learning Theory, V.N. Vapnik

[2] Learning with Kernels, B. Scholkopf and A. Smola

Matrix Approximation

[3] On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel-Based Learning, Drineas and Mahoney

[4] CUR matrix decompositions for improved data analysis, Drineas and Mahoney

[5] Improving CUR Matrix Decomposition and the Nystrom Approximation via Adaptive Sampling, Wang and Zhang

[6] Fast Computation of Low Rank Matrix Approximations, Achlioptas and McSherry

Random Projections

[7] Kernels as Features: On kernels,margins, and low-dimensional mappings, Balcan, Blum and Vempala

[8] Random Features for Large-Scale Kernel Machines, Rahimi and Recht

Kernel Learning

[9] Generalized Maximum Margin Clustering and Unsupervised Kernel Learning, Valizadegan and Jin