Sparse Sampling for Approximation of Kernel Functions

Aims

- *Basic problem*: Low-rank approximation of kernel functions by sparse sampling.

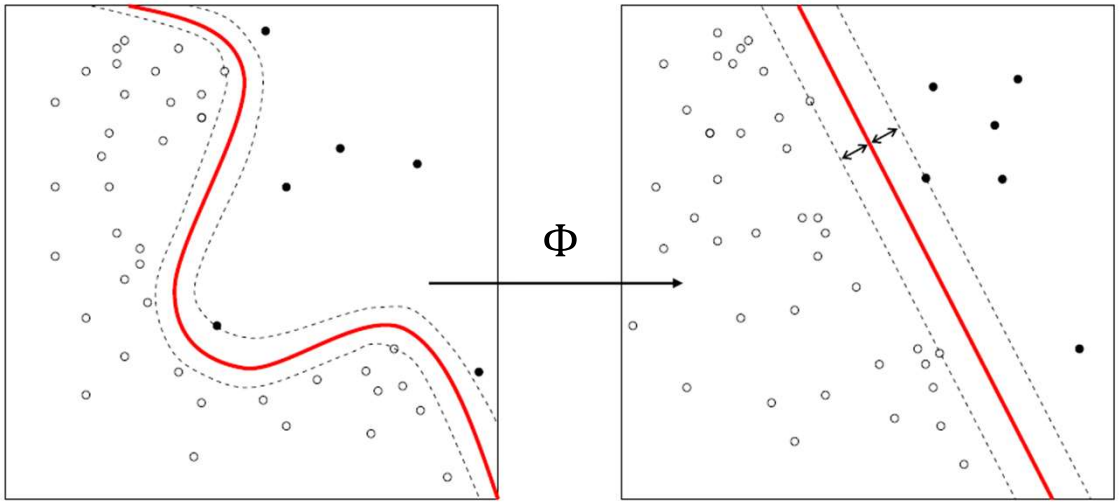
- *Application problem*: Efficient algorithms for neighbor search.

Kernel function

$$k(\mathbf{x}, \mathbf{y}) = k(\|\mathbf{x} - \mathbf{y}\|)$$

e.g., $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$

A kernel function implicitly defines a high-dimensional feature mapping:



$\mathcal{X}$ $\mathcal{H}$

$k(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(x)^\top \mathbf{\Phi}(y)$

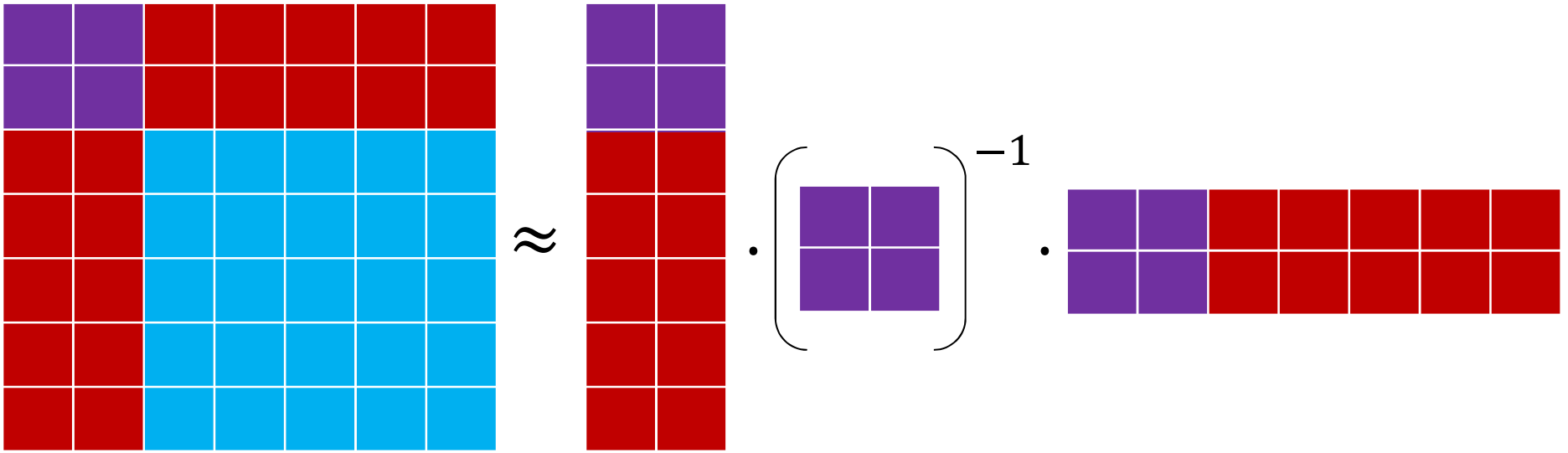
Sparse sampling

- *Analytic description*:

Find a small number of key points $x_1, \ldots, x_n$ from the data set so that

$$k(x, y) \approx [k(x, x_1), \ldots, k(x, x_n)]\mathbf{W}[k(y, x_1), \ldots, k(y, x_n)]^{\mathsf{T}}$$

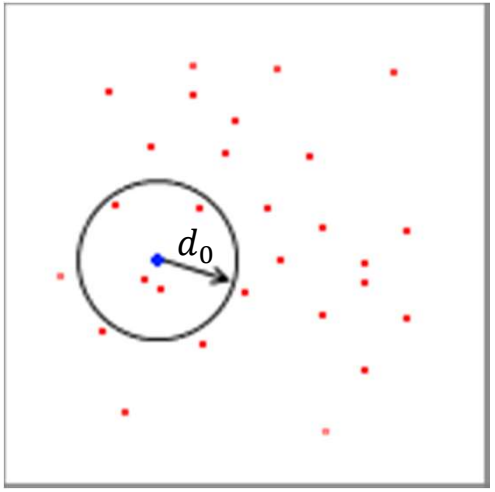
- *Algebraic description*:



Gram matrix $[k(\mathbf{x}_i, \mathbf{y}_j)]$

Neighbor Search

- Search all neighbors of M given points in a data set of size N.



- Based on

Euclidean distance: $d(x, y) = \|x - y\|$

Kernel distance: $d(x, y) = \|\Phi(x) - \Phi(y)\|$

- The complexity of the trivial algorithm is $O(MN)$, which is unacceptable in practice.

- The efficient neighbor search is important for non-parametric modeling and clustering.

Tasks

- Compare the efficiency and accuracy of existing sparse sampling methods.
- Propose (or modify) a sparse sampling algorithm suitable for big data.
- Apply the algorithm to data-driven Langevin modeling or density based clustering.
- Theoretical or experimental analysis of approximation error.

References

- C. Williams and M. Seeger, “Using the Nyström method to speed up kernel machines,” *NIPS*, 2001: 682-688.
- R. Patel, T. A. Goldstein, E. L. Dyer, etc., “oASIS: Adaptive column sampling for kernel matrix approximation,” *Arxiv*: 1505.05208.
- A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, 2014, 344(6191): 1492-1496.
- N. Schaudinnus, B. Bastian, R. Hegger, and G. Stock, “Multidimensional Langevin modeling of nonoverdamped dynamics,” *PRL*, 2015, 115(5): 050602.