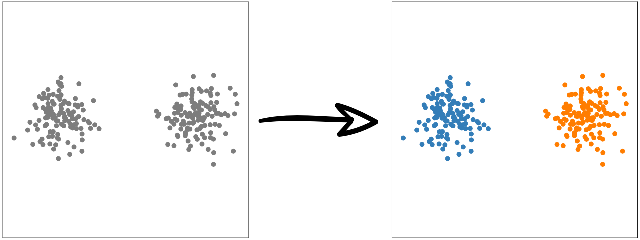
Statistical and Computational Trade-Offs in Kernel K-Means

Daniele Calandriello, Lorenzo Rosasco
LCSL - IIT/MIT and Università di Genova

NeurIPS, December 2018

K-Means



Given n points, partition them into k clusters.

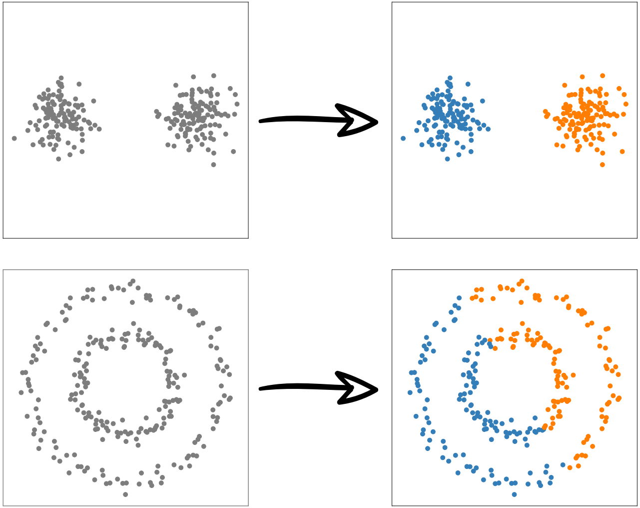
$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|\mathbf{x}_i - \mathbf{c}_j\|^2$$

K-Means

Given n points, partition them into k clusters.

$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\ldots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\| \mathbf{x}_i - \mathbf{c}_j \right\|^2$$

Problem: only linear separation

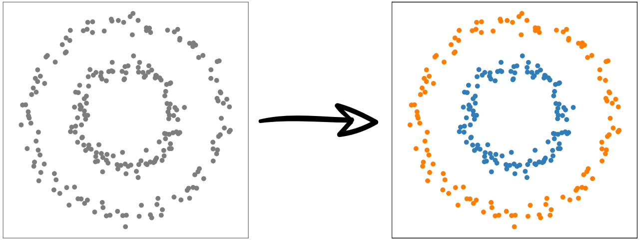


Kernel K-Means

Given n points, partition them into k clusters.

$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\ldots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\| \varphi(\mathbf{x}_i) - \mathbf{c}_j \right\|^2$$

Feature map $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^D$



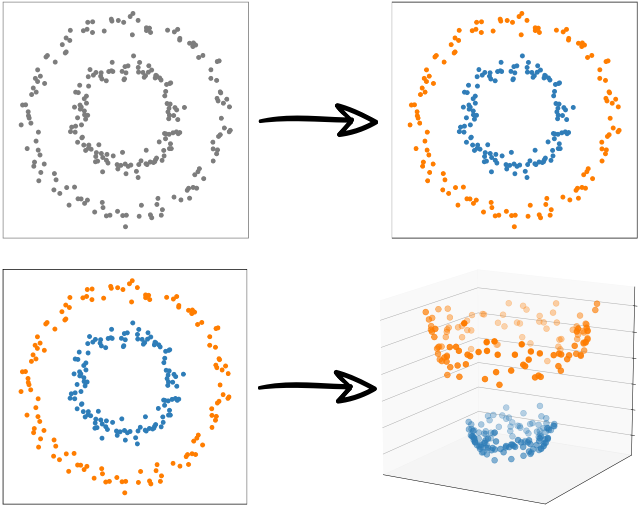
Kernel K-Means

Given n points, partition them into k clusters.

$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \left\| \varphi(\mathbf{x}_i) - \mathbf{c}_j \right\|^2$$

Feature map $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^D$

(e.g., $\varphi([x, y]) = [x, y, x^2 + y^2]$)



Computing Kernel K-Means

$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|\varphi(\mathbf{x}_i) - \mathbf{c}_j\|^2$$

$$\|\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j)\|^2 = \|\varphi(\mathbf{x}_i)\|^2 + \|\varphi(\mathbf{x}_j)\|^2 - 2\,\underbrace{\varphi(\mathbf{x}_i)^\intercal\varphi(\mathbf{x}_j)}_{\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)}$$

kernel

Computing Kernel K-Means

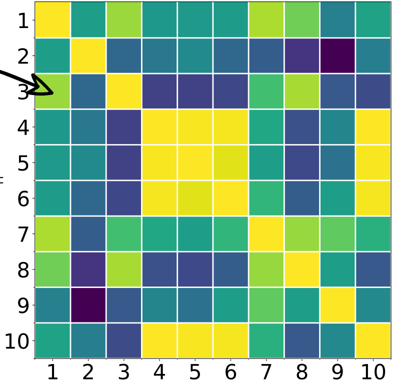
$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \|\varphi(\mathbf{x}_i) - \mathbf{c}_j\|^2$$

$$\|\varphi(\mathbf{x}_i)-\varphi(\mathbf{x}_j)\|^2 = \|\varphi(\mathbf{x}_i)\|^2 + \|\varphi(\mathbf{x}_j)\|^2 - 2\,\underbrace{\varphi(\mathbf{x}_i)^{\mathsf{T}}\varphi(\mathbf{x}_j)}_{\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)}$$

kernel

$\mathcal{K}(\mathbf{x}_3,\mathbf{x}_1)$

$\mathbf{K} =$



Computing Kernel K-Means

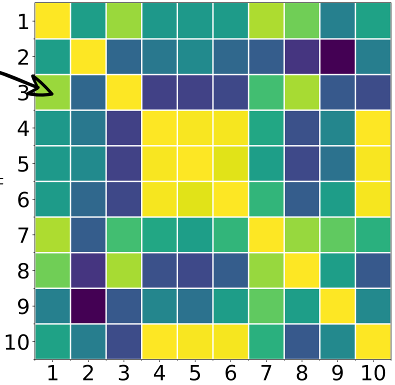
$$\widehat{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \|\varphi(\mathbf{x}_i) - \mathbf{c}_j\|^2$$

$$\|\varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j)\|^2 = \|\varphi(\mathbf{x}_i)\|^2 + \|\varphi(\mathbf{x}_j)\|^2 - 2\,\underbrace{\varphi(\mathbf{x}_i)^\mathsf{T}\varphi(\mathbf{x}_j)}_{\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)}$$

kernel

$\mathcal{K}(\mathbf{x}_3, \mathbf{x}_1)$

$\mathbf{K} =$



Space n^2, Construct $\mathbf{K}$ n^2, Iter. time: n^2

K-Means with Uniform Nyström Embedding

$$\widetilde{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \left\| \varphi_m(\mathbf{x}_i) - \mathbf{c}_j \right\|^2$$

K-Means with Uniform Nyström Embedding

$$\widetilde{\mathbf{C}} = \min_{[\mathbf{c}_1,\ldots,\mathbf{c}_j]} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\| \varphi_m(\mathbf{x}_i) - \mathbf{c}_j \right\|^2$$

$$\|\varphi_m(\mathbf{x}_i) - \varphi_m(\mathbf{x}_j)\|^2 = \|\varphi_m(\mathbf{x}_i)\|^2 + \|\varphi_m(\mathbf{x}_j)\|^2 - 2\underbrace{\varphi_m(\mathbf{x}_i)^{\mathsf{T}}\varphi_m(\mathbf{x}_j)}_{\mathcal{K}_m(\mathbf{x}_i,\mathbf{x}_j)}$$

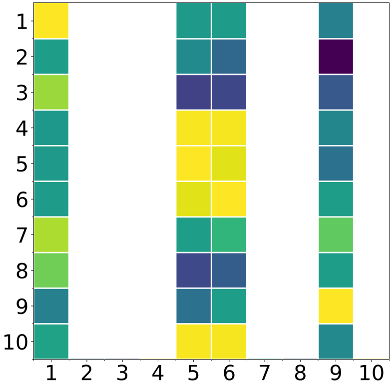
Nyström approximation

K-Means with Uniform Nyström Embedding

$$\widetilde{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \left\| \varphi_m(\mathbf{x}_i) - \mathbf{c}_j \right\|^2$$

$$\|\varphi_m(\mathbf{x}_i) - \varphi_m(\mathbf{x}_j)\|^2 = \|\varphi_m(\mathbf{x}_i)\|^2 + \|\varphi_m(\mathbf{x}_j)\|^2 - 2\underbrace{\varphi_m(\mathbf{x}_i)^\mathsf{T}\varphi_m(\mathbf{x}_j)}_{\mathcal{K}_m(\mathbf{x}_i,\mathbf{x}_j)}$$

Nyström approximation

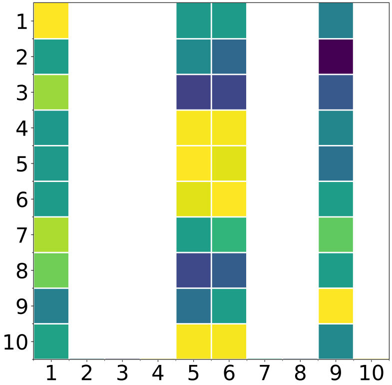


K-Means with Uniform Nyström Embedding

$$\widetilde{\mathbf{C}} = \min_{[\mathbf{c}_1,\dots,\mathbf{c}_j]} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} \left\| \varphi_m(\mathbf{x}_i) - \mathbf{c}_j \right\|^2$$

$$\|\varphi_m(\mathbf{x}_i) - \varphi_m(\mathbf{x}_j)\|^2 = \|\varphi_m(\mathbf{x}_i)\|^2 + \|\varphi_m(\mathbf{x}_j)\|^2 - 2\,\underbrace{\varphi_m(\mathbf{x}_i)^{\mathsf{T}}\varphi_m(\mathbf{x}_j)}_{\mathcal{K}_m(\mathbf{x}_i,\mathbf{x}_j)}$$

Nyström approximation



Space $\cancel{n^2}$ nm, Construct $\widetilde{\mathbf{K}}_m$ $\cancel{n^2}$ nm^2, Iter. time: $\cancel{n^2}$ nmk

K-Means with Uniform Nyström Embedding

$$\widetilde{\mathbf{C}} = \min_{[\mathbf{c}_1,\ldots,\mathbf{c}_j]} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\|\varphi_m(\mathbf{x}_i) - \mathbf{c}_j\right\|^2$$

$\|\varphi_m(\mathbf{x}_i) - \varphi$ …

$\mathcal{K}_m(\mathbf{x}_i, \mathbf{x}_j)$

Nyström approximation

How to choose m for optimal statistical vs computational trade-off?

Space $\cancel{n^2}$ nm, Construct $\widetilde{\mathbf{K}}_m$ $\cancel{n^2}$ nm^2, Iter. time: $\cancel{n^2}$ nmk

Main result

Let $\mathbf{x}_i \sim \mu$ and the test error $\mathcal{E}(\widetilde{\mathbf{C}}) = \mathbb{E}_{\mathbf{x}\sim\mu}[\min_{j=1,\dots,k} \|\varphi(\mathbf{x}) - \widetilde{\mathbf{c}}_j\|^2]$

Main result

Let $\mathbf{x}_i \sim \mu$ and the test error $\mathcal{E}(\widetilde{\mathbf{C}}) = \mathbb{E}_{\mathbf{x}\sim\mu}[\min_{j=1,\dots,k} \|\varphi(\mathbf{x}) - \widetilde{\mathbf{c}}_j\|^2]$

Theorem

$$\mathcal{E}(\widetilde{\mathbf{C}}) \leq \underbrace{\mathcal{O}(k/\sqrt{n})}_{\textit{statistical error}} + \underbrace{\mathcal{O}(k/m)}_{\textit{computational error}}$$

Main result

Let $\mathbf{x}_i \sim \mu$ and the test error $\mathcal{E}(\widetilde{\mathbf{C}}) = \mathbb{E}_{\mathbf{x}\sim\mu}[\min_{j=1,\dots,k} \|\varphi(\mathbf{x}) - \widetilde{\mathbf{c}}_j\|^2]$

Theorem

$$\mathcal{E}(\widetilde{\mathbf{C}}) \leq \underbrace{\mathcal{O}(k/\sqrt{n})}_{\textit{statistical error}} + \underbrace{\mathcal{O}(k/m)}_{\textit{computational error}}$$

$m = \sqrt{n}$ is sufficient for $k/\sqrt{n}$ rate!

Previous results require $m = n$

Main result

Let $\mathbf{x}_i \sim \mu$ and the test error $\mathcal{E}(\widetilde{\mathbf{C}}) = \mathbb{E}_{\mathbf{x}\sim\mu}[\min_{j=1,\dots,k} \|\varphi(\mathbf{x}) - \widetilde{\mathbf{c}}_j\|^2]$

Theorem

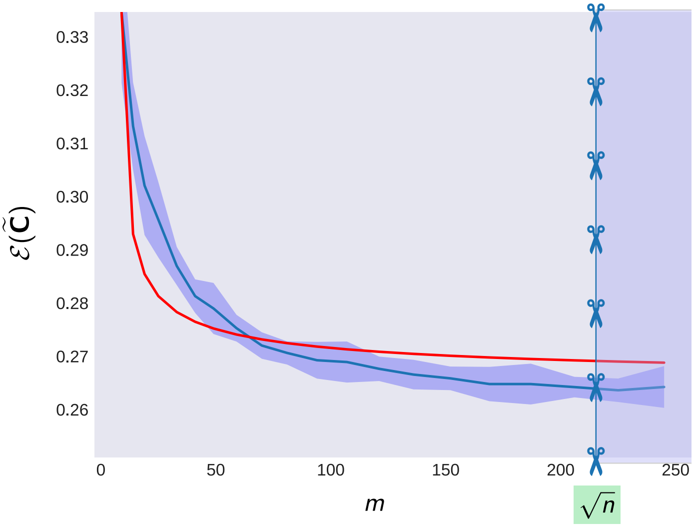
$$\mathcal{E}(\widetilde{\mathbf{C}}) \leq \underbrace{\mathcal{O}(k/\sqrt{n})}_{\textit{statistical error}} + \underbrace{\mathcal{O}(k/m)}_{\textit{computational error}}$$

$m = \sqrt{n}$ is sufficient for $k/\sqrt{n}$ rate!

Previous results require $m = n$

	Space	Construct $\mathbf{K}/\widetilde{\mathbf{K}}_m$	Iter. time
Kernel k-means	n^2	n^2	n^2
Nyström k-means	$n\sqrt{n}$	n^2	$n\sqrt{n}k$

MNIST-60k: test cost vs embedding size m



Recap

Improved statistical vs computational trade-off for k-means

First computation saving with no loss of statistical accuracy

Similar results for k-means++ (efficient)

Open question: fast $\mathcal{O}(k/n)$ rate?

Recap

Improved statistical vs computational trade-off for k-means

First computation saving with no loss of statistical accuracy

Similar results for k-means++ (efficient)

Open question: fast $\mathcal{O}(k/n)$ rate?

Taking suggestions at poster #129