Using the Nyström Method to Speed Up Kernel Machines

Student Seminar (2018/6/27)

Liyuan Xu

Motivation

- Kernel Method

$$f(x) = \sum_{i=1}^{n} w_i k(x, x_i),$$

where $k(x, x')$ is **reproducing kernel** on RKHS H.

$$\forall f \in H, \langle f, k(\,\cdot\,, x) \rangle_H = f(x)\,.$$

Many applications, Expressive, Theoretically sound

Extremely Slow ! (typically takes $O(n^3)$)

Motivation

- Large-Scale Kernel Methods
  - Random Feature Mapping [Rahimi+, 2007]
    - Approximate $k(x, x') \simeq z(x)^\top z(x')$
    - $z(x)$: Low dimensional vector
  - Nyström Methods [Williams+, 2001]
    - Approximate Gram Matrix $K = \left(k(x_i, x_j)\right)_{i,j}$ with low rank decomposition

Reason of choosing the paper

"Using the Nyström Method to Speed Up Kernel Machines"
[Williams+, 2001]

- Straight-forward idea
- Many application beyond the kernel
- Still hot topic in research (2 papers @AISTATS2018)

The true reason of choosing the paper...

At AISTATS2018 poster session

Researcher

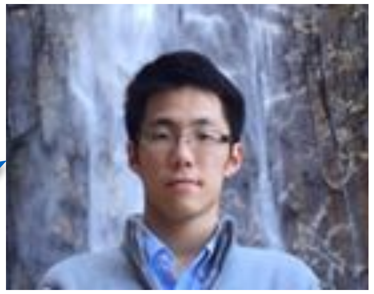
Our research is about Nyström method.

Oh, Nesterov method!

$$x_{k+1} = x_k - \eta \nabla f(x_k + \gamma_k \Delta x_k) + \gamma_k \Delta x_k$$

Kernel method is .....

Me



???

Do not repeat my failure again !

Nyström Method for Approximating Eigenfunction

Decomposition for Kernel (Mercer's theorem)

$$k(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N} \lambda_k \phi_k(\mathbf{x}) \phi_k(\mathbf{y})$$

for $N \leq \infty, \lambda_1 \geq \lambda_2 \geq \ldots \lambda_N \geq 0$, ϕ_k is p-orthogonal

$$\int \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} = \delta_{i,j}$$

Nyström Method for Approximating Eigenfunction

Due to orthogonality,

$$\int k(\mathbf{x}, \mathbf{y})\phi_i(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} = \lambda_i \phi_i(\mathbf{y})$$

Approximate integral by samples $\{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_q\} \sim p(\mathbf{x})$

$$\frac{1}{q}\sum_{k=1}^{q} k(\mathbf{x}'_k, \mathbf{y})\phi_i(\mathbf{x}'_k) \simeq \lambda_i \phi_i(\mathbf{y})$$

Thus, setting $y = x'_1, \ldots, x'_q$ yields following eigenproblem

$$K^{(q)}U^{(q)} = U^{(q)}\Lambda^{(q)}$$

$$K^{(q)} = \left(k(\mathbf{x}'_i, \mathbf{x}'_j)\right)_{i,j \in 1,\ldots,q}, U^{(q)}U^{(q)\top} = I, \Lambda^{(q)} = \mathrm{diag}(\lambda_1^{(q)}, \ldots, \lambda_q^{(q)})$$

Nyström Method for Approximating Eigenfunction

Thus, the eigenvalue and eigenfunction for $\{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_q\}$is

$$\phi_i(\mathbf{x}'_j) \simeq \sqrt{q}\, U^{(q)}_{j,i} \quad \lambda_i \simeq \lambda^{(q)}_i / q$$

Therefore, the eigenfunction is

$$\phi_i(\mathbf{y}) = \frac{1}{\lambda_i} \int k(\mathbf{y}, \mathbf{x}) \phi_i(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} \simeq \frac{1}{\lambda_i} \sum_{k=1}^{q} k(\mathbf{y}, \mathbf{x}'_k) \phi_i(\mathbf{x}'_k)$$

$$\simeq \frac{\sqrt{q}}{\lambda^{(q)}_i} \sum_{k=1}^{q} k(\mathbf{y}, \mathbf{x}'_k) U^{(q)}_{k,i}$$

which is called Nyström Approximation of Eigenfunction

Nyström approximation of Gram matrix

Using the Nyström approximation for eigenfunction

$$
\begin{aligned}
k(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{k=1}^{N} \lambda_k \phi_i(\mathbf{x}_i) \phi_i(\mathbf{x}_j) \\
&\simeq \sum_{k=1}^{q} \frac{\lambda_k^{(q)}}{q} \left(\frac{\sqrt{q}}{\lambda_k^{(q)}} \sum_{l=1}^{q} k(\mathbf{x}_i, \mathbf{x}'_l) U_{l,k}^{(q)} \right) \left(\frac{\sqrt{q}}{\lambda_k^{(q)}} \sum_{l=1}^{q} k(\mathbf{x}_j, \mathbf{x}'_l) U_{l,k}^{(q)} \right)
\end{aligned}
$$

Low-rank approximation of Gram matrix $K = \tilde{U}^{(q)} \tilde{\Lambda}^{(q)} \tilde{U}^{(q)\top}$

$$
\tilde{\Lambda}^{(q)} = \operatorname{diag}\left(\frac{1}{\lambda_1^{(q)}}, \frac{1}{\lambda_2^{(q)}}, \ldots, \frac{1}{\lambda_q^{(q)}} \right), \quad \tilde{U}^{(q)} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}'_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}'_q) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}'_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}'_q) \end{bmatrix} U^{(q)}
$$

Nyström approximation for Gram matrix

Setting $\mathbf{x}'_1 = \mathbf{x}_1, \ldots, \mathbf{x}'_q = \mathbf{x}_q$ and use $\phi_i(\mathbf{x}'_j) \simeq \sqrt{q}\, U^{(q)}_{j,i}$ yields

$$K \simeq \tilde{K} = C K^{\dagger}_{q,q} C^{\top}$$

$A^{\dagger}$: pseudo-inverse

$$K = \begin{bmatrix} K_{q,q} & B^{\top} \\ B & K_{n-q,n-q} \end{bmatrix} \qquad C = \begin{bmatrix} K_{q,q} \\ B \end{bmatrix}$$

Nyström approximation for Gram matrix

Setting $\mathbf{x}'_1 = \mathbf{x}_1, \ldots, \mathbf{x}'_q = \mathbf{x}_q$ and use $\phi_i(\mathbf{x}'_j) \simeq \sqrt{q}\,U^{(q)}_{j,i}$ yields

$$K \simeq \tilde{K} = C K^{\dagger}_{q,q} C^{\top}$$

$$K = \begin{bmatrix} K_{q,q} & B^{\top} \\ B & K_{n-q,n-q} \end{bmatrix} \qquad \tilde{K} = \begin{bmatrix} K_{q,q} & B^{\top} \\ B & \tilde{K}_{n-q,n-q} \end{bmatrix}$$

$$K_{n-q,n-q} - \tilde{K}_{n-q,n-q} = \text{Schur Complement}$$

Application for kernel method

In Gaussian process regression, we have to calculate

$$\mathbf{a} = (K + \sigma I)^{-1}\mathbf{t}$$

Using Nyström approximation $K = \tilde{U}^{(q)}\tilde{\Lambda}^{(q)}\tilde{U}^{(q)\top}$, we have

$$\mathbf{a} = \frac{1}{\sigma}\left(\mathbf{t} - \tilde{U}^{(q)}\left(\sigma I + \tilde{\Lambda}^{(q)}\tilde{U}^{(q)\top}\tilde{U}^{(q)}\right)^{-1}\tilde{\Lambda}^{(q)}{\tilde{U}^{(q)}}^{\top}\mathbf{t}\right)$$

by Woodbury formula, which can be computed in $O(q^2 n)$

Practically, setting $n \gg q$ does not harm the performance

Conclusion

"Using the Nyström Method to Speed Up Kernel Machines" [Williams+, 2001]

- Introduced Nyström approximation
- Apply it to speed up kernel method
- Showed low-rank does not harm the performance empirically