Notes on Recursive Sampling for the Nyström Method

BY YALIN LIAO

Summary

The classical Nystrom algorithm samples landmarks by a uniform distribution to approximate the Gram matrix, which can lead to large deviations in practice since uniform sampling tends to oversample landmarks from dense parts of the dataset and undersample or miss smaller but still important parts of the dataset. This paper proposes to use the sum of ridge leverage scores to sample the landmarks and then apply the classical Nystrom algorithm. The combined algorithm, called RLS-Nystrom (via recursive sampling scheme), has the advantages in terms of computation complexity and storing space and gives strong theoretucal guarantees for Nyström approximation.

Some Understanding

Nyström approximation

The Nyström method selects a subset of "landmark" points and uses them to construct a low-rank approximation to K. Given a matrix $S \in \mathbb{R}^{n\times s}$ that has a single entry in each column equal to 1 so that KS is a subset of s columns from K, the associated Nyström approximation is:

$$\tilde{K} = KS(S^TKS)^\dagger S^TK$$

Assume that K is a 4×4 matrix ($n=4$). Let $K = [\ K_1\ \ K_2\ \ K_3\ \ K_4\ ]$ and $S = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$. Then

$$KS = [\ K_2\ \ K_4\ ]$$

which is indeed a submatrix of K (consists of 2th and 4th columns from K). Note $S^TS = I_2$ and generally $S^TS = I_s$.

We view Nyström approximation as a low-rank approximation to the dataset in feature space. How to understand?

First, $\tilde{K}$ is a matrix of rank no more than s as

$$\operatorname{rank}(\tilde{K}) = \operatorname{rank}(KS(S^TKS)^\dagger S^TK) \le \operatorname{rank}(S) = s$$

So $\hat{K}$ is a low rank matrix. Why is $\tilde{K}$ an approximator of K? Write $K = \Phi\Phi^T$, where

$$\Phi = \begin{bmatrix} \phi(x_1) \\ \vdots \\ \phi(x_n) \end{bmatrix}$$

is kernelized data matrix. S select s kernelized data matrix $S^T\Phi$, a submatrix of Φ, which has the form

$$S^T\Phi = \begin{bmatrix} \phi(x_{i_1}) \\ \vdots \\ \phi(x_{i_s}) \end{bmatrix}$$

and we approxmiate Φ using its projection (Φ's projection) onto these points $S^T\Phi$ (particularly onto the row space of $S^T\Phi$, $R(S^T\Phi)$). There are infinite matrices which can project data matrix Φ to the subspace $R(S^T\Phi)$, which is the ***optimal projection*** matrix to keep the F norm of data matrix Φ? The answer from linear algebra is the orthogonal projection, $P_S = \Phi^T S(S^T\Phi\Phi^T S)^\dagger S^T\Phi$.[1] Projecting Φ by P_S , we have $\tilde{\Phi} = \Phi P_S$ and

$$\tilde{K} = \tilde{\Phi}\tilde{\Phi}^T = \Phi P_S P_S^T \Phi^T = \Phi P_S \Phi^T = \Phi\Phi^T S(S^T\Phi\Phi^T S)^\dagger S^T\Phi\Phi^T = KS(S^T\Phi\Phi^T S)^\dagger S^T K$$

where we have used the definition of the orthogonal matrix, $P_S = P_S^T = P_S^2$. Now we get the answer why Nyström approximation is a low rank approximation of K.

Recursive sampling

So far, I don't fully understanding the sampling mechanism. Seeing the word "recursive" reminds me that dynamically adjusting model parameters could be applied in my research.

$$D_{\mathrm{KL}}(p_Y \| p_U) = \sup_f \{\mathbb{E}_Y[f(Y)] - \mathbb{E}_U[e^{f(U)-1}]\}$$

In pratice, we have the unstable training issue since samples from p_Y and p_U do not overlap. We first train neural network to solve

$$D_{\mathrm{KL}}(p_Y \| \theta p_U + (1-\theta)p_Y), \theta = \frac{1}{2}$$

and then gradually increase θ to 1 over training process.

$$\lim_{\theta\to 1} D_{\mathrm{KL}}(p_Y \| \theta p_U + (1-\theta)p_Y) = D_{\mathrm{KL}}(p_Y \| p_U)$$

To be specific, we mix the samples from p_Y and p_U with equal data points to estimate the second term of dual form of KL divergence. Over training course, we gradually increase the samples from p_U or decrease samples from p_Y in the second term.

1. The orthogonal projection, which project any vector to the column space of A, $C(A)$, is

$$P_A = A(A^T A)^{-1}A^T$$

Also, P_A is the solution to

$$\min_{P: Px \in C(A)} \|x - Px\|_2^2, \forall x$$

So the optimization problem (each row optimizer is the same)

$$\min_{P: \Phi P \in R(S^T\Phi)} \|\Phi - \Phi P\|_F^2$$

have solution $P_S = (S^T\Phi)^T[(S^T\Phi)(S^T\Phi)^T]^\dagger(S^T\Phi) = \Phi^T S(S^T\Phi\Phi^T S)^\dagger S^T\Phi$. Here probably it is more suitable to use the notation $P_{S^T\Phi}$ for the projection matrix.