Nyström Based Kernel Classification for Big Data

September 18, 2015

Background

Classical Setting in Machine Learning

- n training examples: $\{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$ where $\mathbf{x_i} \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$.

Background

feature representation

Classical Setting in Machine Learning

- n training examples: $\{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$ where $\mathbf{x_i} \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$.

Background

target variable

Classical Setting in Machine Learning

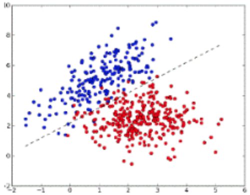
- n training examples: $\{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$ where $\mathbf{x_i} \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_i \in \mathcal{Y}$.

Background

Classical Setting in Machine Learning

- n training examples: $\{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$ where $\mathbf{x_i} \in \mathcal{X} \subseteq \mathbb{R}^d$,$y_i \in \mathcal{Y}$.
- The goal of machine learning is to learn a predictive function $f(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Y}$

  - Classification: $\mathcal{Y} = \{-1, +1\}$
  - Regression: $\mathcal{Y} \subseteq \mathbf{R}$



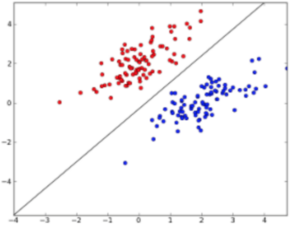
Background

Linear Model

- Predictive function $f(x) = \mathbf{w}^T\mathbf{x}$
- Optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^T\mathbf{x}_i, y_i) + \lambda R(\mathbf{w})$$

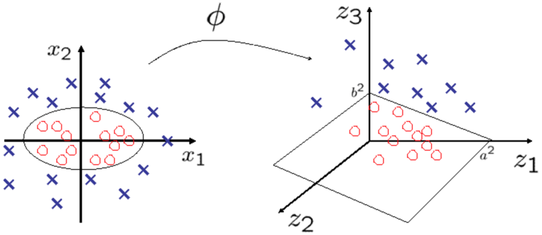
- Example: SVM, Logistic Regression



- Pros: efficient to solve
- Cons: suffer lower performance when data points are not linearly separable

Background

Non-Linear Model



Background

Non-Linear Model

- Predictive function $f(x) = \mathbf{w}^T\phi(\mathbf{x})$, where $\phi(\cdot)$ is mapping function
- Optimization problem (Primal form):

$$\min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \frac{\lambda}{2}||f||^2_{\mathcal{H}_K} \tag{1}$$

- Optimization problem (Dual form):

$$\max_{\alpha\in\mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T K\alpha \tag{2}$$

- Where ℓ^* is the conjugate function of loss function ℓ and $K \in \mathbb{R}^{n\times n}$ is kernel matrix

Background

Non-Linear Model

- Kernel trick $\kappa(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$

$$K = \begin{bmatrix} \phi(\mathbf{x}_1)^T\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_1)^T\phi(\mathbf{x}_n) \\ \phi(\mathbf{x}_2)^T\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_2)^T\phi(\mathbf{x}_n) \\ \vdots & \vdots & \vdots \\ \phi(\mathbf{x}_n)^T\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_n)^T\phi(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

- Kernel function: $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2)$, $\kappa(\mathbf{x}, \mathbf{y}) = (a\mathbf{x}^T\mathbf{y} + c)^d$
- Pros: enjoys high performance
- Cons: hard to solve

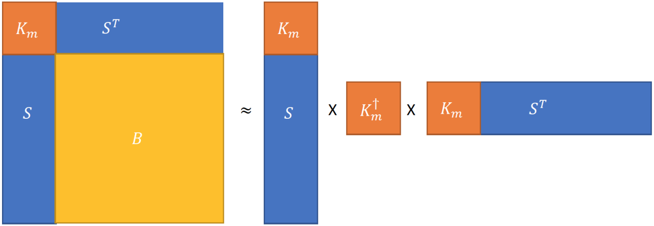
Challenge in Big Data

- X is $n \times d$ matrix, $n \longrightarrow$ millions, billions, $\cdots$
- Need to compute kernel matrix K

$$K = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

- Very very large kernel matrix K
- Computation and memory cost

Nyström Method



$$K \approx K_b K_m^\dagger K_b^T$$

Theorems about Nyström Method

Theorem[Drineas and Mahoney, 2005]

For any m uniformly sampled columns, with a high probability,

$$||K - K_b K_m^{\dagger} K_b^T||_2 = O(\frac{n}{\sqrt{m}})$$

Theorem[Jin et al., 2011]

For any m uniformly sampled columns, assume there exists $\rho \in (0, 1/2)$ such that $\lambda_m = \Omega(n/m^{\rho})$ and $\lambda_{m+1} = O(n/m^{1-\rho})$, with a high probability,

$$||K - K_b K_m^{\dagger} K_b^T||_2 = O(\frac{n}{m^{1-\rho}})$$

Nyström based Kernel Classification

- Nyström approximation dual optimization

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T (K_b K_m^{\dagger} K_b^T)\alpha \tag{3}$$

- Short feature representation

$$\begin{aligned}\hat{K} &= K_b K_m^{\dagger} K_b^T \\ &= K_b V D^{-1} V^T K_b^T \\ &= (D^{-1/2} V^T K_b^T)^T (D^{-1/2} V^T K_b^T) \\ &= \hat{X}^T \hat{X}\end{aligned}$$

- Recall $K = \Phi(X)^T \Phi(X)$ while $\hat{K} = \hat{X}^T \hat{X}$

Nyström based Kernel Classification

- Nyström approximation dual optimization with short feature representation

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T \hat{X}^T \hat{X}\alpha \tag{4}$$

$$\hat{X} = D^{-1/2} V^T K_b^T$$

- $X \in \mathbb{R}^{n \times d}, \Phi(X) \in \mathbb{R}^{n \times hd}, \hat{X} \in \mathbb{R}^{n \times m}$

Nyström based Kernel Classification

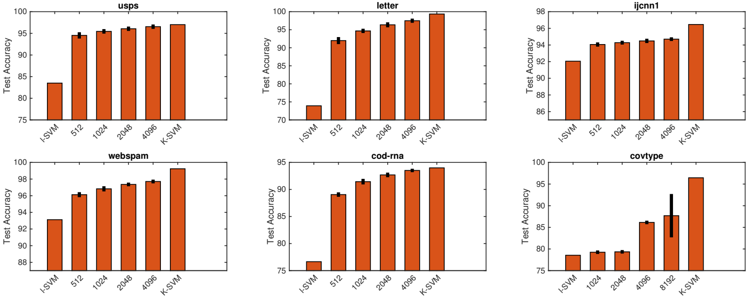
- Draw m samples $\{\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_m}\}$ from n examples
- Compute sub-kernel matrix $K_m \in \mathbb{R}^{m \times m}$ among m samples and $K_b \in \mathbb{R}^{n \times m}$ between all examples and m samples
- Singular Value Decompostion on $K_m = VDV^T$
- $\hat{X} = D^{-1/2} V^T K_b^T$
- Reduced the kernel problem to linear model

Statistic of Datasets

Name	usps	letter	ijcnn1	webspam	cod-rna	covtype
#Training	7,291	12,000	91,701	280,000	271,617	464,810
#Testing	2,007	6000	49,990	70,000	59,535	116,202
#Features	256	16	22	254	8	54

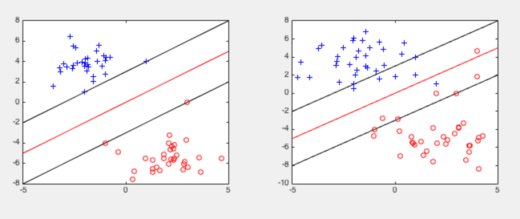
Test Accuracy

- Compare linear-SVM, Kernel SVM, and Nyström Method using $m = 512, 1024, 2048$ and 4096



Adding ℓ_1 regularization

- Approximation error



$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T \hat{X}^T \hat{X}\alpha - \frac{\tau}{n}||\alpha||_1 \tag{5}$$

Analysis

- Approximation error

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T \hat{X}^T \hat{X}\alpha - \frac{\tau}{n}\|\alpha\|_1 \quad (6)$$

Lemma

Let $\mathcal{S}$ be the support set of α_* and $\mathcal{S}^c$ denote its complement. By setting $\tau \geq \frac{2}{\lambda n}\sum_{i=1}^{n} |[\alpha_*]_i| \|\widehat{K}_{*i} - K_{*i}\|_\infty$, we have $\|[\widetilde{\alpha}_* - \alpha_*]_{\mathcal{S}^c}\|_1 \leq 3\|[\widetilde{\alpha}_* - \alpha_*]_{\mathcal{S}}\|_1$.

Analysis

- Proof:

Analysis

Lemma

Let $q = \frac{1}{n}X^T(A^TA - I)e$*.With a probability at least* $1 - \delta$, *we have*

$$||q||_\infty \leq \frac{c\eta R}{n}\sqrt{\frac{\log(d/\delta)}{m}} \tag{7}$$

where c *is the universal constant in the JL lemma,* $||e||_2 \leq \eta$ *and* $max_{1<j<d}||x_j||_2 \leq R$

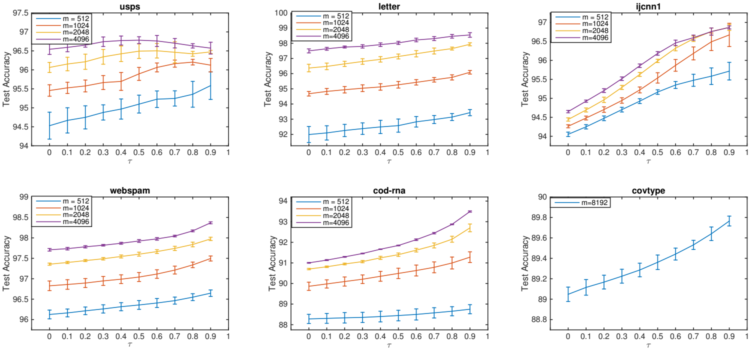
- In our case, we need to similar $\Delta = \frac{1}{\lambda n}(\hat{X}^T\hat{X} - X^TX)\alpha*$, Can we use the same strategy?

Analysis

-

Test Accuracy

- Adding ℓ_1 regularization



Two Strategies to Refine Nystrom

- Probability sampling data points

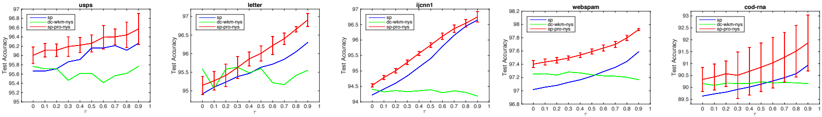
$$Pr(X_i \text{ is selected}) = \frac{|\tilde{\alpha}_i|}{\sum_{i=1}^{n} |\tilde{\alpha}_i|} \quad (8)$$

- Weighted kmean to constructed data points

$$\min \sum_{i=1}^{n} [\alpha_i]^2 ||x_i - c_{\pi_i}||^2 \quad (9)$$

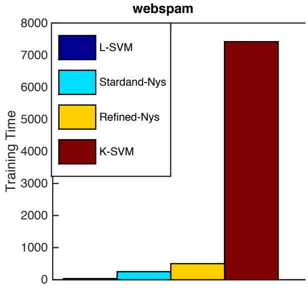
Test Accuracy

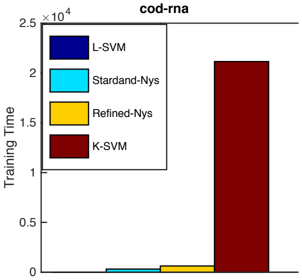
- adding ℓ_1 Nyström, ℓ_1-pro-Nyström and weight-kmean-Nyström



Time complexity

- Linear SVM, Standard-Nyström, Refined-Nyström and Kernel SVM





Conclusion and Future Work

- Nyström method is a powerful method for matrix approximation.
- Nyström based kernel classification can acheive high performance with less compuation and memory.
- Adding ℓ_1 norm on Nyström based on kernel classification can improve test accuracy.
- Theoretical analysis for Adding ℓ_1 norm on Nyström based on kernel classification is the future work.

Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

Rong Jin, Tianbao Yang, and Mehrdad Mahdavi. Improved bound for the nystrom's method and its application to kernel classification. *CoRR*, abs/1111.2262, 2011. URL `http://arxiv.org/abs/1111.2262`.