Binning in Gaussian Kernel Regularization

Tao Shi
The Ohio State University

Bin Yu
University of California, Berkeley
OUTLINE

• Gaussian Kernel Regularization

• Binning in Regression

• Binning in Support Vector Machine
Regression and Classification

Regression Model:

\[ Y_i = f(X_i) + \epsilon_i \]

Classification:

\[
P(Y_i = 1|X_i) = 1 - P(Y_i = -1|X_i) = f(X_i)
\]
Regression / Function Estimation
Regularization Methods

Regularization methods find function \( f \) that minimizes

\[
L(f, \text{data}) + \lambda J(f)
\]

--- \( L \) is an empirical loss
--- \( J(f) \) is a penalty functional, which is usually a norm or semi-norm of a Reproducing Kernel Hilbert Space (RKHS)

For Example: cubic spline corresponds to

\[
J(f) = \int [f''(x)]^2 dx
\]
Periodic Gaussian Kernel Regression

To find a function $f$ minimizing

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda J_{pg}(f)$$

while $f$ belonging to the RKHS of the periodic Gaussian kernel, and $J_{pg}(f)$ is the norm of $f$ in the RKHS

-- Gaussian kernel:

$$G(s, t) = (2\pi)^{-1/2} \omega^{-1} \exp\left(-\frac{(s - t)^2}{2\omega^2}\right)$$

-- Periodic Gaussian kernel on $(0,1]$:

$$K(s, t) = 2 \sum_{l=0}^{\infty} \exp(-l^2\omega^2/2) \cos(2\pi l(s - t))$$
Periodic Gaussian Kernel Regression (cont)

It is equivalent to find \( \mathbf{c} \) to minimize:

\[
[(y - G^{(n)}c)^T (y - G^{(n)}c) + \lambda c^T G^{(n)}c]
\]

with \( G^{(n)} \) as a \( n \) by \( n \) Gram matrix and \( G^{(n)}_{i,j} = K(x_i, x_j) \). The solution is found as

\[
\hat{c} = (G^{(n)} + \lambda I)^{-1}y.
\]

The fitted value is

\[
\hat{y} = G^{(n)}\hat{c} = G^{(n)}(G^{(n)} + \lambda I)^{-1}y.
\]

NOTICE: The computation complexity is \( O(n^3) \) for the matrix inversion, so it is not computationally attractive when the sample size is large.
Asymptotic Properties of PGKR

Periodic Gaussian Kernel Regularization is rate optimal for estimating functions any finite order Sobolev space $H^k(Q)$

$$H^k(Q) = \{ f \in L^2(0, 1) : f \text{ is periodic, } \int_0^1 [f(t)]^2 + [f^{(k)}(t)]^2 dt \leq Q \}.$$  

Alternative Definition:

$$H^k(Q) = \{ f : f(t) = \sum_{l=0}^{\infty} \theta_l \phi_l(t), \sum_{l=0}^{\infty} \gamma_l \theta_l^2 \leq Q, \gamma_0 = 1, \gamma_{2l-1} = \gamma_{2l} = l^{2k} + 1 \}$$

with $\phi_0(t) = 1$, $\phi_{2l-1}(t) = 2^{1/2} \sin(2\pi lt)$, and $\phi_{2l}(t) = 2^{1/2} \cos(2\pi lt)$

Minimax Rate of PGKR for $H^k(Q)$: (Lin and Brown 2004)

$$(2k + 1) k^{-2k/(2k+1)} Q^{1/(2k+1)} n^{-2k/(2k+1)}$$
Gaussian Kernel Support Vector Machines

With labels \( y_i \in \{-1, 1\} \), SVMs find \( f \in \text{RKHS}_K \) to minimize

\[
\sum_{i=1}^{n} (1 - y_i f(x_i))^+ + \lambda \|f\|_K^2
\]

The solution is in the form of

\[
\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)
\]

with most of \( \alpha \)'s being 0, the non-zero \( \alpha \)'s are called support vectors. The most common used kernel in machine learning literature is the Gaussian kernel

The computation of training SVM is \( O(n^2) \) empirically
OUTLINE

• Gaussian Kernel Regularization

• Binning in Regression

• Binning in Support Vector Machine
Binning in Regression
Binning for PGKR in Regression

Let \( x_1, \ldots, x_n \) be equally spaced in \((0, 1]\) and assume \( n = mp \). Binning the data as:

\[
\bar{x}_j = \frac{x_{(j-1)\times p+1} + \cdots + x_{(j-1)\times p+p}}{p}
\]

\[
\bar{y}_j = \frac{y_{(j-1)\times p+1} + \cdots + y_{(j-1)\times p+p}}{p}
\]

Use PGKR on the binned data, we get:

\[
\hat{y} = G^{(n,m)}(G^{(m)} + \lambda_B I)^{-1} B^{(m,n)} y
\]

with

\[
G_{i,j}^{(n,m)} = K(x_i, \bar{x}_j) \quad G_{i,j}^{(m)} = K(\bar{x}_i, \bar{x}_j)
\]

and

\[
B_{i,j}^{(m,n)} = I\{|j/p| = i\}/p
\]
Asymptotic Properties of Binned PGKR

THEOREM: (Shi and Yu 2006, *Statistica Sinica* 16, 541-567) The binned PGKR reaches the same minimax rate as the unbinned PGKR does in any finite order Sobolev space $H^k(Q)$, when number of bins and regularization parameter $\lambda_B$ are properly chosen.

NOTICE: the binned estimator only requires invert the matrix $G^{(m)} + \lambda_B I$.

The computation complexity is $O(m^3)$. For estimator $k$-th order Sobolev space $H^k(Q)$, we need $m = O(kn^{k/(2k+1)})$ to achieve the optimal rate.

Overall computation: $O(n)$ for $k = 1$, $O(n^{3/5})$ for $k = 2$, and …
Asymptotic Properties of Binned PGKR (cont)

Key ideals in the proof:

• The eigen-vectors $V^{(n)}$ of $G^{(n)}$ are the trigonometric basis evaluated at $x_1, \ldots, x_n$

• $G^{(n,m)} V^{(m)} = \lambda V^{(n)}$

• The variance and bias of the binned estimator can be explicitly written out in the finite sample case

• Trading off between variance and bias finishes the proof
Simulation Study

\[ Y_i = f(X_i) + \epsilon_i \]

sample size: 120
repeat 100 times

\[ f_1(x) = \sin^2(2\pi x)1_{(x \leq 1/2)} \]

\[ f_2(x) = -x + 2(x - 1/4)1_{(x \geq 1/4)} + 2(-x + 3/4)1_{(x \geq 3/4)} \]

\[ f_3(x) = 1/(2 - \sin(2\pi x)) \]

\[ f_4(x) = 2 + \sin(2\pi x) + 2\cos(2\pi x) + 3\sin^2(2\pi x) + 4\cos^3(2\pi x) + 5\sin^3(2\pi x) \]
Simulation Study (cont)

MSE of the Binned PGKR v.s. the number of points in each bin.

MSE of the binned Gaussian KR v.s. the number of points in each bin.

Note: the MSE with 1 point in each bin corresponds to the unbinned estimator.
OUTLINE

• Gaussian Kernel Regularization

• Binning in Regression

• Binning in Support Vector Machine
Gaussian Kernel SVMs

The solution is in the form of

\[ \hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \]

Most of \( \alpha \)'s being 0

The non-zero \( \alpha \)'s are called support vectors

SV’s are the misclassified points and the points within the “margin”
Binning Gaussian Kernel SVMs

- Split each predictor by its marginal distribution
- Find the data points in each bin
- For non-empty bins, average the predictors and take majority vote of labels
Binning Gaussian Kernel SVMs (cont)

- Reduce the variance of labels (see illustration below)
- Keep the marginal distribution of predictors
- Reduce the training sample size and the number of SV’s
Earth Observing System (EOS) is designed by NASA for studying the Earth from space using a multiple-instrument, multiple-satellite approach. Improving scientific understanding of global climate changes and providing scientific basis for environmental policies.
Cloud Detection

- Uncertainties about cloud radiation feedback on the global climate are among the greatest obstacles in understanding and predicting the Earth's future climate.
Polar Cloud Detection
Challengers in Polar Cloud Detection

1. Surface covered by snow and ice
2. Huge data size (1.9GB/minute)
3. High dimensionality (9 angles and 4 spectral bands)
4. Fast computation is needed for online data process
MISR Features

Three features based on 275 m, terrain projected red radiances. Classification results are reported on 1.1 km MISR grid. (Shi. et al 2004, 2006)

1. correlation between angles

\[ \text{CORR} = \frac{(r_{AF-AN} + r_{BF-AN})}{2} \]

2. surface smoothness

\[ \text{SD}_{AN} \]

3. the angular signature of the radiances from different angles: Normalized Difference Angular Index (Nolin, Fetterer, and Scambos 2002)

\[ \text{NDAI} = \frac{(R_{DF} - R_{AN})}{(R_{DF} + R_{AN})} \]
One Example

MISR data collected over Greenland in summer 2002 (54879 labeled pixels in the pictures, white - “cloudy”, gray - “clear”, and black - “unsure”)

Expert labels are given by Prof. Eugene Clothiaux (Meteorologist in collaboration)

Three features (CORR, SD, NDAI) are used for classification
SVMs for Polar Cloud Detection

We train Gaussian Kernel SVMs on three features in 4 setups:

1. SVMs on random sample ~ 1000 data points

2. Bagged SVMs (repeat training SVM on ~ 1000 random sample 21 times; then average the prediction of 21 runs)

3. SVMs on bin centers (split each predictor to 10 bins and use the resulted 996 bin centers and majority vote of labels)

4. SVMs on half of the sample (27179 training points)

All parameters of SVM are tuned by cross-validation
SVMs for Polar Cloud Detection

<table>
<thead>
<tr>
<th></th>
<th>random sample size 966</th>
<th>SVM on 966 bin centers</th>
<th>SVM size 27179</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>SVM</td>
<td>Bagged SVM</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>*85.09%</td>
<td>86.07%</td>
<td>86.08%</td>
</tr>
<tr>
<td>Comp Time (seconds)</td>
<td>$81 \times 1.85$</td>
<td>$21 \times 81 \times 1.85$</td>
<td>$3.87 + 81 \times 1.85$</td>
</tr>
<tr>
<td></td>
<td>= 2.5 minutes</td>
<td>= 52.11 minutes</td>
<td>= 2.56 minutes</td>
</tr>
<tr>
<td># Support Vectors</td>
<td>350</td>
<td>$\sim 7350$</td>
<td>210</td>
</tr>
</tbody>
</table>

$\lambda$ and $\sigma$ are chosen by CV over a 9 x 9 grid.

1. SVM on bin centers give the closest rate to the “full” SVM

2. Binning step itself (3.87 sec) is fast, compared to training SVMs

3. SVM on bin centers have the fewest SV’s, which leads to fast speed in the prediction step
Binning or Clustering

Feng and Mangasarian (2001) suggest using k-mean clustering to reduce the training size

The computation time and memory requirement of K-mean / K-median clustering increase dramatically when the number of centroids increases

The classification rates of clustering+SVM and binning+SVM are similar as shown in the polar cloud detection problem

<table>
<thead>
<tr>
<th></th>
<th>512 bins</th>
<th>966 bins</th>
<th>Rate of SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binning(+SVM)</td>
<td>3.45 sec</td>
<td>3.87 sec</td>
<td>85.64%</td>
</tr>
<tr>
<td>Clustering(+SVM)</td>
<td>21.65 min</td>
<td>Out of Memory</td>
<td>85.72%</td>
</tr>
</tbody>
</table>
Concluding Remarks

• Binning on Gaussian Kernel Regularization keeps the accuracy and reduce the computation significantly

• Binning on Gaussian kernel SVMs speeds up both the training and testing speed

• The computation of binning is faster than other training sample size reduction methods, such as clustering and bagged SVMs
Acknowledgements

• Collaborators:
  Dr. Amy Braverman (JPL, NASA)
  Prof. Eugene Clothiaux (PSU)

• Financial support from NSF, NASA and ARO

• MISR data were obtained at the courtesy of the NASA Langley Research Atmospheric Sciences Data Center