Machine Learning Techniques
(機器學習技巧)

Lecture 6: Kernel Models for Regression

Hsuan-Tien Lin (林軒田)
htlin@csie.ntu.edu.tw

Department of Computer Science & Information Engineering
National Taiwan University
(國立台灣大學資訊工程系)
Agenda

Lecture 6: Kernel Models for Regression

- Kernel Ridge Regression
- Support Vector Regression Primal
- Support Vector Regression Dual
- Summary of Kernel Models
Recall: Representer Theorem

for any L2-regularized linear model

\[
\min_w \frac{\lambda}{N} w^T w + \frac{1}{N} \sum_{n=1}^{N} \text{err}(y_n, w^T z_n)
\]

optimal \( w_* = \sum_{n=1}^{N} \beta_n z_n \).

—any L2-regularized linear model can be kernelized!

regression with squared error

\[
\text{err}(y, w^T z) = (y - w^T z)^2
\]

—analytic solution for linear/ridge regression

analytic solution for kernel ridge regression?
Kernel Ridge Regression Problem

Solving ridge regression:

\[
\min_w \frac{\lambda}{N} w^T w + \frac{1}{N} \sum_{n=1}^{N} (y_n - w^T z_n)^2
\]

Yields optimal solution:

\[
w_* = \sum_{n=1}^{N} \beta_n z_n
\]

With out loss of generality, can solve for optimal \( \beta \) instead of \( w \)

\[
\min_{\beta} \frac{\lambda}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_n \beta_m K(x_n, x_m) + \frac{1}{N} \sum_{n=1}^{N} \left( y_n - \sum_{m=1}^{N} \beta_m K(x_n, x_m) \right)^2
\]

= \[
\frac{\lambda}{N} \beta^T K \beta + \frac{1}{N} \left( \beta^T K^T K \beta - 2 \beta^T K^T y + y^T y \right)
\]

Kernel ridge regression:

Use representer theorem for kernel trick on ridge regression.
Kernel Models for Regression

Solving Kernel Ridge Regression

\[
E_{\text{aug}}(\beta) = \frac{\lambda}{N} \beta^T K \beta + \frac{1}{N} \left( \beta^T K^T K \beta - 2 \beta^T K^T y + y^T y \right)
\]

\[
\nabla E_{\text{aug}}(\beta) = \frac{2}{N} \left( \lambda K^T I \beta + K^T K \beta - K^T y \right) = \frac{2}{N} K^T \left( \lambda I + K \right) \beta - y
\]

want \( \nabla E_{\text{aug}}(\beta) = 0 \): one analytic solution

\[
\beta = (\lambda I + K)^{-1} y
\]

- \( \cdot^{-1} \) always exists for \( \lambda > 0 \), because \( K \) positive semi-definite (\textit{Mercer's condition, remember? :-)}\)
- time complexity: \( O(N^3) \) with simple \textit{dense} matrix inversion

\[
\text{can now do non-linear regression 'easily'}
\]
Kernel Models for Regression

Kernel Ridge Regression

Linear versus Kernel Ridge Regression

**Linear Ridge Regression**

\[ w = (\lambda I + X^T X)^{-1} X^T y \]

- more restricted
- \(O(d^3 + d^2 N)\) training;
  \(O(d)\) prediction
- efficient when \(N \gg d\)

**Kernel Ridge Regression**

\[ \beta = (\lambda I + K)^{-1} y \]

- more flexible with \(K\)
- \(O(N^3)\) training;
  \(O(N)\) prediction
- hard for big data

Linear versus kernel: trade-off between efficiency and flexibility
Soft-Margin SVM versus Least-Squares SVM

least-squares SVM (LSSVM) = \textit{kernel ridge regression} for classification

- LSSVM: similar boundary, \textbf{many more SVs} \implies \textit{slower prediction, dense }\beta (BIG \ g)
- dense \beta: LSSVM, kernel LogReg; \textbf{sparse }\alpha: \textit{standard SVM}

want: \textbf{sparse }\beta \textit{ like standard SVM}
will consider **tube regression**

- within a **tube**: no error
- outside a tube: error by distance to tube

error measure:

\[
\text{err}(y, s) = \max(0, |s - y| - \epsilon)
\]

- \( |s - y| \leq \epsilon \): 0
- \( |s - y| > \epsilon \): \(|s - y| - \epsilon\)

—usually called \(\epsilon\)-insensitive error with \(\epsilon > 0\)

todo: L2-regularized tube regression to get **sparse** \(\beta\)
**Tube versus Squared Regression**

**tube:** \( \text{err}(y, s) = \max(0, |s - y| - \epsilon) \)

**squared:** \( \text{err}(y, s) = (s - y)^2 \)

\( \text{tube} \approx \text{squared} \) when \( |s - y| \) small & less affected by outliers
L2-Regularized Tube Regression

\[
\min_w \quad \frac{\lambda}{N} w^T w + \frac{1}{N} \sum_{n=1}^{N} \max \left( 0, |w^T z_n - y| - \epsilon \right)
\]

Regularized Tube Regr.

\[
\min \frac{\lambda}{N} w^T w + \frac{1}{N} \sum \text{tube violation}
\]

- unconstrained, but \text{max} not differentiable
- ‘representer’ to kernelize, but no obvious sparsity

Standard SVM

\[
\min \frac{1}{2} w^T w + C \sum \text{margin vio.}
\]

- not differentiable, but QP
- dual to kernelize, KKT conditions \Rightarrow sparsity

will mimic standard SVM derivation:

\[
\min_{b,w} \quad \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max \left( 0, |w^T z_n + b - y_n| - \epsilon \right)
\]
Standard Support Vector Regression Primal

\[
\begin{align*}
\min_{b, w} \quad & \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max \left(0, |w^T z_n + b - y_n| - \epsilon\right) \\
\text{s.t.} \quad & |w^T z_n + b - y_n| \leq \epsilon + \xi_n \\
& \xi_n \geq 0
\end{align*}
\]

mimicking standard SVM

\[
\begin{align*}
\min_{b, w, \xi} \quad & \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \\
\text{s.t.} \quad & |w^T z_n + b - y_n| \leq \epsilon + \xi_n \\
& \xi_n \geq 0
\end{align*}
\]

making constraints linear

\[
\begin{align*}
\frac{1}{2} w^T w + C \sum_{n=1}^{N} \left(\xi_n^{\vee} + \xi_n^{\wedge}\right) \\
-\epsilon - \xi_n^{\vee} \leq w^T z_n + b - y_n \leq \epsilon + \xi_n^{\wedge} \\
\xi_n^{\vee} \geq 0, \xi_n^{\wedge} \geq 0
\end{align*}
\]

Support Vector Regression (SVR) primal: minimize regularizer + upper tube violations \(\xi_n^{\wedge}\) & lower violations \(\xi_n^{\vee}\)
Fun Time
Quadratic Programming for SVR

\[
\begin{align*}
\min_{b, w, \xi^\vee, \xi^\wedge} & \quad \frac{1}{2} w^T w + C \sum_{n=1}^{N} (\xi_n^\vee + \xi_n^\wedge) \\
\text{s.t.} & \quad -\epsilon - \xi_n^\vee \leq w^T z_n + b - y_n \leq \epsilon + \xi_n^\wedge \\
& \quad \xi_n^\vee \geq 0, \xi_n^\wedge \geq 0
\end{align*}
\]

- Parameter \( C \): trade-off of regularisation & tube violation
- Parameter \( \epsilon \): vertical tube width
  — one more parameter to choose!
- QP of \( \tilde{d} + 1 + 2N \) variables, \( 2N + 2N \) constraints

next: remove dependence on \( \tilde{d} \) by SVR primal \( \Rightarrow \) dual?
Lagrange Multipliers $\alpha^\land$ & $\alpha^\lor$

Objective function

$$\frac{1}{2} w^T w + C \sum_{n=1}^{N} (\xi^\lor_n + \xi^\land_n)$$

Lagrange multiplier $\alpha^\lor_n$ for $-\epsilon - \xi^\lor_n \leq w^T z_n + b - y_n$

Lagrange multiplier $\alpha^\land_n$ for $w^T z_n + b - y_n \leq \epsilon + \xi^\land_n$

Some of the KKT Conditions

- $\frac{\partial L}{\partial w_i} = 0$:
  $$w = \sum_{n=1}^{N} \left( \alpha^\lor_n - \alpha^\land_n \right) z_n$$

- Complementary slackness:
  $$\alpha^\lor_n (-\epsilon - \xi^\lor_n - w^T z_n - b + y_n) = 0$$
  $$\alpha^\land_n (-\epsilon - \xi^\land_n + w^T z_n + b - y_n) = 0$$

Standard dual can be derived using the same steps as Lecture 20
### Support Vector Regression Dual

#### SVM Dual and SVR Dual

**SVM Dual**

\[
\min \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \\
\text{s.t.} \quad y_n(w^T z_n + b) \geq 1 - \xi_n \\
\xi_n \geq 0
\]

**SVR Dual**

\[
\min \frac{1}{2} w^T w + C \sum_{n=1}^{N} \left( \xi_n^\wedge + \xi_n^\vee \right) \\
\text{s.t.} \quad 1(w^T z_n + b - y_n) \leq \epsilon + \xi_n^\wedge \\
1(y_n - w^T z_n + b) \leq \epsilon + \xi_n^\vee \\
\xi_n^\wedge \geq 0, \xi_n^\vee \geq 0
\]

**Kernel SVM Dual**

\[
\min \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m K(x_n, x_m) \\
- \sum_{n=1}^{N} 1 \cdot \alpha_n \\
\text{s.t.} \quad \sum_{n=1}^{N} y_n \alpha_n = 0 \\
0 \leq \alpha_n \leq C
\]

**Kernel SVR Dual**

\[
\min \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (\alpha_n^\vee - \alpha_n^\wedge)(\alpha_m^\vee - \alpha_m^\wedge) k_{n,m} \\
- \sum_{n=1}^{N} \left( (\epsilon + y_n) \cdot \alpha_n^\wedge + (\epsilon - y_n) \cdot \alpha_n^\vee \right) \\
\text{s.t.} \quad \sum_{n=1}^{N} 1 \cdot (\alpha_n^\vee - \alpha_n^\wedge) = 0 \\
0 \leq \alpha_n^\wedge \leq C, 0 \leq \alpha_n^\vee \leq C
\]

*similar QP, solvable by similar solver*
Kernel Models for Regression

Support Vector Regression Dual

Sparsity of SVR Solution

- \( \mathbf{w} = \sum_{n=1}^{N} \left( \alpha_n^\vee - \alpha_n^\wedge \right) \mathbf{z}_n \)

- complementary slackness:
  \[
  \begin{align*}
  \alpha_n^\vee \left( -\epsilon - \xi_n^\vee - \mathbf{w}^T \mathbf{z}_n - b + y_n \right) &= 0 \\
  \alpha_n^\wedge \left( -\epsilon - \xi_n^\wedge + \mathbf{w}^T \mathbf{z}_n + b - y_n \right) &= 0
  \end{align*}
  \]

- strictly within tube \(|\mathbf{w}^T \mathbf{z}_n + b - y_n| < \epsilon\)
  \[
  \implies \alpha_n^\wedge = 0 \text{ and } \alpha_n^\vee = 0
  \]
  \[
  \implies \beta_n = 0
  \]

- SVs \((\beta_n \neq 0)\): on or outside tube

SVR: allows \textbf{sparse} \(\beta\)
Fun Time
PLA/pocket
minimize $\text{err}_0/1$ specially

linear soft-margin SVM
minimize regularized $\hat{\text{err}}_{\text{SVM}}$ by QP

linear SVR
minimize regularized $\text{err}_{\text{TUBE}}$ by QP

linear ridge regression
minimize regularized $\text{err}_{\text{SQR}}$ analytically

regularized logistic regression
minimize regularized $\text{err}_{\text{CE}}$ by GD/SGD

second row: popular in LIBLINEAR
Summary of Kernel Models

Map of Linear/Kernel Models

- PLA/pocket
- linear soft-margin SVM
- SVM
- SVR
- probabilistic SVM

- linear SVR
- linear ridge regression
- kernel ridge regression
- kernel logistic regression
- kernelized linear ridge regression
- kernelized regularized logistic regression

- minimize SVM dual by QP
- minimize SVR dual by QP
- run SVM-transformed logistic regression

fourth row: popular in LIBSVM

Hsuan-Tien Lin (NTU CSIE)
Map of Linear/Kernel Models

first row: less used due to worse performance
third row: less used due to dense $\beta$
Kernel Models

possible kernels:

polynomial, Gaussian, ..., your own from Mercer’s condition,
coupled with

kernel ridge regression
kernel logistic regression
SVM
SVR
probabilistic SVM

powerful extension of linear models
—*with great power comes great responsibility*
in *Spiderman*, remember? :-)

Hsuan-Tien Lin (NTU CSIE)
Fun Time
Summary

Lecture 6: Kernel Models for Regression

- Kernel Ridge Regression
  - representer theorem on RidgeReg
- Support Vector Regression Primal
  - minimize regularized tube errors
- Support Vector Regression Dual
  - a QP similar to SVM
- Summary of Kernel Models
  - with great power comes great responsibility