Lecture 6: Support Vector 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)
Roadmap

1. Embedding Numerous Features: Kernel Models

   Lecture 5: Kernel Logistic Regression
   - two-level learning for SVM-like sparse model for soft classification, or using representer theorem with regularized logistic error for dense model

   Lecture 6: Support Vector Regression
   - Kernel Ridge Regression
   - Support Vector Regression Primal
   - Support Vector Regression Dual
   - Summary of Kernel Models

2. Combining Predictive Features: Aggregation Models

3. Distilling Implicit Features: Extraction Models
Recall: Representer Theorem

for any \textbf{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 \textbf{L2-regularized} linear model can be \textit{kernelized}!

regression with squared error

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

—analytic solution for linear/ridge regression

\textbf{analytic solution} for \textit{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 \( w \)

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

Regularization of \( \beta \) on K-based regularizer

Linear regression of \( \beta \) on K-based features

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

Kernel ridge regression:
Use representer theorem for kernel trick on ridge regression
Solving Kernel Ridge Regression

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

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

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

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

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

can now do non-linear regression ‘easily’
**Linear versus Kernel Ridge Regression**

**Linear Ridge Regression**

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

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

**Kernel Ridge Regression**

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

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

**linear versus kernel:**
trade-off between **efficiency** and **flexibility**
After getting the optimal $\beta$ from kernel ridge regression based on some kernel function $K$, what is the resulting $g(x)$?

1. \[ \sum_{n=1}^{N} \beta_n K(x_n, x) \]
2. \[ \sum_{n=1}^{N} y_n \beta_n K(x_n, x) \]
3. \[ \sum_{n=1}^{N} \beta_n K(x_n, x) + \lambda \]
4. \[ \sum_{n=1}^{N} y_n \beta_n K(x_n, x) + \lambda \]
After getting the optimal $\beta$ from kernel ridge regression based on some kernel function $K$, what is the resulting $g(x)$?

1. $\sum_{n=1}^{N} \beta_n K(x_n, x)$
2. $\sum_{n=1}^{N} y_n \beta_n K(x_n, x)$
3. $\sum_{n=1}^{N} \beta_n K(x_n, x) + \lambda$
4. $\sum_{n=1}^{N} y_n \beta_n K(x_n, x) + \lambda$

Reference Answer: ①

Recall that the optimal $w = \sum_{n=1}^{N} \beta_n z_n$ by representer theorem and $g(x) = w^T z$. The answer comes from combining the two equations with the kernel trick.
Soft-Margin SVM versus Least-Squares SVM

least-squares SVM (LSSVM)  
= kernel ridge regression for classification

- LSSVM: similar boundary, **many more SVs**  
  \( \implies \) slower prediction, **dense \( \beta \) (BIG \( g \))

- dense \( \beta \): LSSVM, kernel LogReg;  
  **sparse \( \alpha \): standard SVM**

want: **sparse \( \beta \)** 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**

\[ \text{err}(y, s) = (s - y)^2 \]

- Tube ≈ Squared when \( |s - y| \) small
- Less affected by outliers
L2-Regularized Tube Regression

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

Regularized Tube Regr.

\[
\min \frac{\lambda}{N} w^T w + \frac{1}{N} \sum \text{tube violation}
\]
- unconstrained,
- but \textbf{max not differentiable}
- ‘representer’ to kernelize,
- but \textbf{no obvious sparsity}

Standard SVM

\[
\min \frac{1}{2} w^T w + C \sum \text{margin vio.}
\]
- not differentiable,
- but \textbf{QP}
- dual to kernelize,
- KKT conditions \Rightarrow \textbf{sparsity}

will mimic \textbf{standard SVM} derivation:

\[
\min_{b,w} \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max \left( 0, |w^T z_n + b - y_n| - \epsilon \right)
\]
Support Vector Regression (SVR) primal:
minimize regularizer + (upper tube violations $\xi_n^\vee$ & lower violations $\xi_n^\wedge$)

**Standard Support Vector Regression Primal**

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

**Support Vector Regression (SVR) primal:**

 Mimicking standard SVM

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

Making constraints linear

$$\begin{align*}
\frac{1}{2} w^T w + C \sum_{n=1}^{N} (\xi_n^\vee + \xi_n^\wedge) \\
-\epsilon - \xi_n^\vee \leq y_n - w^T z_n - b \leq \epsilon + \xi_n^\wedge \\
\xi_n^\vee \geq 0, \xi_n^\wedge \geq 0
\end{align*}$$
Quadratic Programming for SVR

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

- parameter \( C \): trade-off of regularization & 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 \( \Rightarrow \text{dual?} \)
Consider solving support vector regression with $\epsilon = 0.05$. At the optimal solution, assume that $w^T z_1 + b = 1.234$ and $y_1 = 1.126$. What is $\xi^\vee_1$ and $\xi^\wedge_1$?

1. $\xi^\vee_1 = 0.108$, $\xi^\wedge_1 = 0.000$
2. $\xi^\vee_1 = 0.000$, $\xi^\wedge_1 = 0.108$
3. $\xi^\vee_1 = 0.058$, $\xi^\wedge_1 = 0.000$
4. $\xi^\vee_1 = 0.000$, $\xi^\wedge_1 = 0.058$
Consider solving support vector regression with $\epsilon = 0.05$. At the optimal solution, assume that $w^T z_1 + b = 1.234$ and $y_1 = 1.126$. What is $\xi^\vee_1$ and $\xi^\wedge_1$?

1. $\xi^\vee_1 = 0.108, \xi^\wedge_1 = 0.000$

2. $\xi^\vee_1 = 0.000, \xi^\wedge_1 = 0.108$

3. $\xi^\vee_1 = 0.058, \xi^\wedge_1 = 0.000$

4. $\xi^\vee_1 = 0.000, \xi^\wedge_1 = 0.058$

Reference Answer: 3

$y_1 - w^T z_1 - b = -0.108 < -0.05$, which means that there is a lower tube violation of amount 0.058. When there is a lower tube violation on some example, trivially there is no upper tube violation.
Support Vector Regression Dual

Lagrange Multipliers $\alpha^\wedge$ & $\alpha^\vee$

Objective function

$$\frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^{N} (\xi^\vee_n + \xi^\wedge_n)$$

Lagrange multiplier $\alpha^\wedge_n$ for $y_n - \mathbf{w}^T \mathbf{z}_n - b \leq \epsilon + \xi^\wedge_n$

Lagrange multiplier $\alpha^\vee_n$ for $-\epsilon - \xi^\vee_n \leq y_n - \mathbf{w}^T \mathbf{z}_n - b$

Some of the KKT Conditions

- $\frac{\partial L}{\partial \mathbf{w}_i} = 0$: $\mathbf{w} = \sum_{n=1}^{N} (\alpha^\wedge_n - \alpha^\vee_n) \mathbf{z}_n$; $\frac{\partial L}{\partial b} = 0$: $\sum_{n=1}^{N} (\alpha^\wedge_n - \alpha^\vee_n) = 0$

- Complementary slackness:
  - $\alpha^\wedge_n (\epsilon + \xi^\wedge_n - y_n + \mathbf{w}^T \mathbf{z}_n + b) = 0$
  - $\alpha^\vee_n (\epsilon + \xi^\vee_n + y_n - \mathbf{w}^T \mathbf{z}_n - b) = 0$

Standard dual can be derived using the same steps as Lecture 4
Support Vector Regression (SVR)

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

Support Vector Regression Dual (SVRD)

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

Hsuan-Tien Lin (NTU CSIE)
Support Vector Regression

Support Vector Regression Dual

Sparsity of SVR Solution

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

- Complementary slackness:
  \[
  \alpha_n^\wedge (\epsilon + \xi_n^\wedge - y_n + \mathbf{w}^T \mathbf{z}_n + b) = 0 \\
  \alpha_n^\vee (\epsilon + \xi_n^\vee + y_n - \mathbf{w}^T \mathbf{z}_n - b) = 0
  \]

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

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

SVR: allows sparse \( \beta \)
What is the number of variables within the QP problem of SVR dual?

1. \( \tilde{d} + 1 \)
2. \( \tilde{d} + 1 + 2N \)
3. \( N \)
4. \( 2N \)
What is the number of variables within the QP problem of SVR dual?

- 1. $\tilde{d} + 1$
- 2. $\tilde{d} + 1 + 2N$
- 3. $N$
- 4. $2N$

Reference Answer: 4

There are $N$ variables within $\alpha^\vee$, and another $N$ in $\alpha^\wedge$. 
Map of Linear Models

**PLA/pocket**
minimize $\text{err}_0/1$ specially

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

**linear soft-margin SVM**
minimize regularized $\hat{\text{err}}_{\text{SVM}}$ 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: minimize SVM dual by QP
- SVR: minimize SVR dual by QP
- linear SVR
- linear ridge regression
- kernel ridge regression
- kernel logistic regression
- regularized logistic regression
- kernelized linear ridge regression
- kernelized regularized logistic regression
- probabilistic SVM: run SVM-transformed logistic regression

fourth row: popular in LIBSVM
Support Vector Regression

Summary of Kernel Models

Map of Linear/Kernel Models

- PLA/pocket
- linear soft-margin SVM
- SVM
- linear SVR
- linear ridge regression
- kernel ridge regression
- SVM
- SVR
- regularized logistic regression
- kernel logistic regression
- probabilistic SVM

First row: less used due to worse performance

Third row: less used due to dense $\beta$
possible kernels:

polynomial, Gaussian, . . ., your design (with 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? :-)
Which of the following model is less used in practice?

1. pocket
2. ridge regression
3. (linear or kernel) soft-margin SVM
4. regularized logistic regression

Reference Answer:

The pocket algorithm generally does not perform better than linear soft-margin SVM, and hence is less used in practice.
Which of the following model is less used in practice?

1. pocket
2. ridge regression
3. (linear or kernel) soft-margin SVM
4. regularized logistic regression

Reference Answer: 1

The pocket algorithm generally does not perform better than linear soft-margin SVM, and hence is less used in practice.
Summary

1. Embedding Numerous Features: Kernel Models
   - Lecture 6: Support Vector Regression
     - Kernel Ridge Regression
       * representer theorem on ridge regression
     - Support Vector Regression Primal
       * minimize regularized tube errors
     - Support Vector Regression Dual
       * a QP similar to SVM dual
     - Summary of Kernel Models
       * with great power comes great responsibility

2. Combining Predictive Features: Aggregation Models
   - next: making cocktail from learning models

3. Distilling Implicit Features: Extraction Models