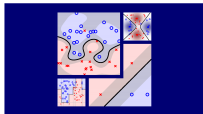


Machine Learning Techniques (機器學習技法)



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 L2-regularized linear model

$$\min_{\mathbf{w}} \quad \frac{\lambda}{N} \mathbf{w}^T \mathbf{w} + \frac{1}{N} \sum_{n=1}^N \text{err}(y_n, \mathbf{w}^T \mathbf{z}_n)$$

optimal $\mathbf{w}_* = \sum_{n=1}^N \beta_n \mathbf{z}_n$.

—any L2-regularized linear model can be **kernelized!**

regression with squared error

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

—analytic solution for linear/ridge regression

analytic solution for **kernel ridge regression?**

Kernel Ridge Regression Problem

solving ridge regression $\min_{\mathbf{w}} \frac{\lambda}{N} \mathbf{w}^T \mathbf{w} + \frac{1}{N} \sum_{n=1}^N (y_n - \mathbf{w}^T \mathbf{z}_n)^2$

yields optimal solution $\mathbf{w}_* = \sum_{n=1}^N \beta_n \mathbf{z}_n$

with out loss of generality, can solve for optimal β instead of \mathbf{w}

$$\min_{\beta} \underbrace{\frac{\lambda}{N} \sum_{n=1}^N \sum_{m=1}^N \beta_n \beta_m K(\mathbf{x}_n, \mathbf{x}_m)}_{\text{regularization of } \beta \text{ on } K\text{-based regularizer}} + \underbrace{\frac{1}{N} \sum_{n=1}^N \left(y_n - \sum_{m=1}^N \beta_m K(\mathbf{x}_n, \mathbf{x}_m) \right)^2}_{\text{linear regression of } \beta \text{ on } K\text{-based features}}$$

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

kernel ridge regression:

use **representer theorem** for kernel trick on **ridge regression**

Solving Kernel Ridge Regression

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

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

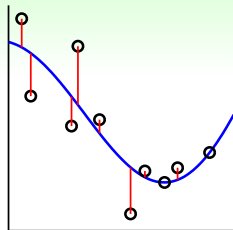
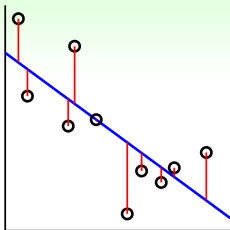
want $\nabla E_{\text{aug}}(\boldsymbol{\beta}) = \mathbf{0}$: one analytic solution

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

- $(\cdot)^{-1}$ always exists for $\lambda > 0$, because \mathbf{K} 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
- $O(d^3 + d^2 N)$ 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**

Fun Time

After getting the optimal β from kernel ridge regression based on some kernel function K , what is the resulting $g(\mathbf{x})$?

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

Fun Time

After getting the optimal β from kernel ridge regression based on some kernel function K , what is the resulting $g(\mathbf{x})$?

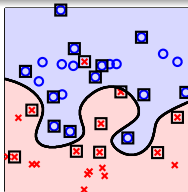
- 1 $\sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x})$
- 2 $\sum_{n=1}^N y_n \beta_n K(\mathbf{x}_n, \mathbf{x})$
- 3 $\sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x}) + \lambda$
- 4 $\sum_{n=1}^N y_n \beta_n K(\mathbf{x}_n, \mathbf{x}) + \lambda$

Reference Answer: 1

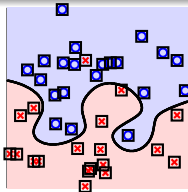
Recall that the optimal $\mathbf{w} = \sum_{n=1}^N \beta_n \mathbf{z}_n$ by representer theorem and $g(\mathbf{x}) = \mathbf{w}^T \mathbf{z}$. The answer comes from combining the two equations with the kernel trick.

Soft-Margin SVM versus Least-Squares SVM

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



soft-margin Gaussian SVM



Gaussian LSSVM

- LSSVM: similar boundary, **many more SVs**
 \implies slower prediction, **dense β (BIG g)**
- dense β : LSSVM, kernel LogReg;
sparse α : standard SVM

want: **sparse β** like standard SVM

Tube Regression

will consider **tube regression**

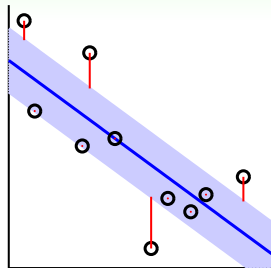
- within a **tube**: **no error**
- outside a tube: **error** by distance to tube

error measure:

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

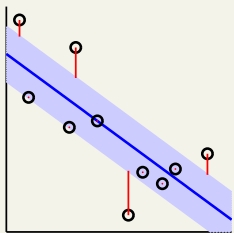
- $|s - y| \leq \epsilon$: 0
- $|s - y| > \epsilon$: $|s - y| - \epsilon$

—usually called ϵ -insensitive error with $\epsilon > 0$

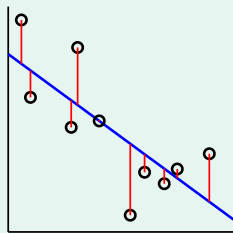


todo: **L2-regularized tube regression**
to get **sparse β**

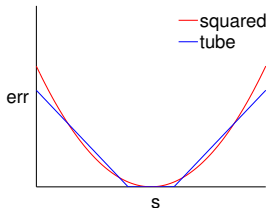
Tube versus Squared Regression



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



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



tube \approx **squared** when $|s - y|$ small
& **less affected by outliers**

L2-Regularized Tube Regression

$$\min_{\mathbf{w}} \quad \frac{\lambda}{N} \mathbf{w}^T \mathbf{w} + \frac{1}{N} \sum_{n=1}^N \max \left(0, |\mathbf{w}^T \mathbf{z}_n - y| - \epsilon \right)$$

Regularized Tube Regr.

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

- unconstrained, but **max not differentiable**
- 'representer' to kernelize, but **no obvious sparsity**

standard SVM

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

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

will mimic **standard SVM** derivation:

$$\min_{b, \mathbf{w}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \max \left(0, |\mathbf{w}^T \mathbf{z}_n + b - y_n| - \epsilon \right)$$

Standard Support Vector Regression Primal

$$\min_{b, \mathbf{w}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \max(0, |\mathbf{w}^T \mathbf{z}_n + b - y_n| - \epsilon)$$

mimicking standard SVM

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

making constraints linear

$$\begin{aligned} \min_{b, \mathbf{w}, \xi} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\xi_n^V + \xi_n^A) \\ \text{s.t.} \quad & -\epsilon - \xi_n^V \leq y_n - \mathbf{w}^T \mathbf{z}_n - b \leq \epsilon + \xi_n^A \\ & \xi_n^V \geq 0, \xi_n^A \geq 0 \end{aligned}$$

Support Vector Regression (SVR) primal:

minimize regularizer + (upper tube violations ξ_n^A & lower violations ξ_n^V)

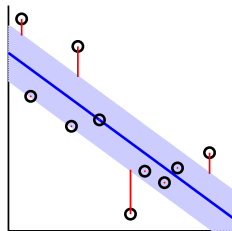
Quadratic Programming for SVR

$$\min_{b, \mathbf{w}, \xi_n^V, \xi_n^A} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\xi_n^V + \xi_n^A)$$

$$\text{s.t.} \quad -\epsilon - \xi_n^V \leq y_n - \mathbf{w}^T \mathbf{z}_n - b \leq \epsilon + \xi_n^A$$

$$\xi_n^V \geq 0, \xi_n^A \geq 0$$

- parameter C : trade-off of regularization & tube violation
- parameter ϵ : 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?

Fun Time

Consider solving support vector regression with $\epsilon = 0.05$. At the optimal solution, assume that $\mathbf{w}^T \mathbf{z}_1 + b = 1.234$ and $y_1 = 1.126$. What is ξ_1^V and ξ_1^\wedge ?

- 1 $\xi_1^V = 0.108, \xi_1^\wedge = 0.000$
- 2 $\xi_1^V = 0.000, \xi_1^\wedge = 0.108$
- 3 $\xi_1^V = 0.058, \xi_1^\wedge = 0.000$
- 4 $\xi_1^V = 0.000, \xi_1^\wedge = 0.058$

Fun Time

Consider solving support vector regression with $\epsilon = 0.05$. At the optimal solution, assume that $\mathbf{w}^T \mathbf{z}_1 + b = 1.234$ and $y_1 = 1.126$. What is ξ_1^V and ξ_1^A ?

- 1 $\xi_1^V = 0.108, \xi_1^A = 0.000$
- 2 $\xi_1^V = 0.000, \xi_1^A = 0.108$
- 3 $\xi_1^V = 0.058, \xi_1^A = 0.000$
- 4 $\xi_1^V = 0.000, \xi_1^A = 0.058$

Reference Answer: 3

$y_1 - \mathbf{w}^T \mathbf{z}_1 - b = -0.108 < -0.05$, 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.

Lagrange Multipliers α^\wedge & α^\vee

objective function $\frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\xi_n^\vee + \xi_n^\wedge)$

Lagrange multiplier α_n^\wedge for $y_n - \mathbf{w}^T \mathbf{z}_n - b \leq \epsilon + \xi_n^\wedge$

Lagrange multiplier α_n^\vee for $-\epsilon - \xi_n^\vee \leq y_n - \mathbf{w}^T \mathbf{z}_n - b$

Some of the KKT Conditions

• $\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = 0: \mathbf{w} = \sum_{n=1}^N \underbrace{(\alpha_n^\wedge - \alpha_n^\vee)}_{\beta_n} \mathbf{z}_n$; $\frac{\partial \mathcal{L}}{\partial b} = 0: \sum_{n=1}^N (\alpha_n^\wedge - \alpha_n^\vee) = 0$

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

standard dual can be derived
 using the same steps as Lecture 4

SVM Dual and SVR Dual

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \xi_n \\ \text{s.t.} \quad & y_n (\mathbf{w}^T \mathbf{z}_n + b) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\xi_n^{\wedge} + \xi_n^{\vee}) \\ \text{s.t.} \quad & 1(y_n - \mathbf{w}^T \mathbf{z}_n - b) \leq \epsilon + \xi_n^{\wedge} \\ & 1(\mathbf{w}^T \mathbf{z}_n + b - y_n) \leq \epsilon + \xi_n^{\vee} \\ & \xi_n^{\wedge} \geq 0, \xi_n^{\vee} \geq 0 \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m K(\mathbf{x}_n, \mathbf{x}_m) \\ & - \sum_{n=1}^N 1 \cdot \alpha_n \\ \text{s.t.} \quad & \sum_{n=1}^N y_n \alpha_n = 0 \\ & 0 \leq \alpha_n \leq C \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N (\alpha_n^{\wedge} - \alpha_n^{\vee})(\alpha_m^{\wedge} - \alpha_m^{\vee}) k_{n,m} \\ & + \sum_{n=1}^N ((\epsilon - y_n) \cdot \alpha_n^{\wedge} + (\epsilon + y_n) \cdot \alpha_n^{\vee}) \\ \text{s.t.} \quad & \sum_{n=1}^N 1 \cdot (\alpha_n^{\wedge} - \alpha_n^{\vee}) = 0 \\ & 0 \leq \alpha_n^{\wedge} \leq C, 0 \leq \alpha_n^{\vee} \leq C \end{aligned}$$

similar QP, **solvable by similar solver**

Sparsity of SVR Solution

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

- 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 $|\mathbf{w}^T \mathbf{z}_n + b - y_n| < \epsilon$
 - $\implies \xi_n^{\wedge} = 0$ and $\xi_n^{\vee} = 0$
 - $\implies (\epsilon + \xi_n^{\wedge} - y_n + \mathbf{w}^T \mathbf{z}_n + b) \neq 0$ and $(\epsilon + \xi_n^{\vee} + y_n - \mathbf{w}^T \mathbf{z}_n - b) \neq 0$
 - $\implies \alpha_n^{\wedge} = 0$ and $\alpha_n^{\vee} = 0$
 - $\implies \beta_n = 0$
- SVs ($\beta_n \neq 0$): **on or outside tube**

SVR: allows **sparse** β

Fun Time

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 $2N$

Fun Time

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 α^\vee , and another N in α^\wedge .

Map of Linear Models

PLA/pocket

minimize
 $\text{err}_{0/1}$ specially

linear SVR

minimize regularized
 err_{TUBE} by QP

linear soft-margin
SVM

minimize regularized
 $\widehat{\text{err}}_{\text{SVM}}$ by QP

linear ridge
regression

minimize regularized
 err_{SQR} analytically

regularized logistic
regression

minimize regularized
 err_{CE} by GD/SGD

second row: popular in **LIBLINEAR**

Map of Linear/Kernel Models

PLA/pocket

linear SVR

linear soft-margin
SVM

linear ridge
regression

regularized logistic
regression

kernel ridge
regression

kernel logistic
regression

kernelized linear ridge
regression

kernelized regularized
logistic regression

SVM

minimize SVM dual by
QP

SVR

minimize SVR dual by
QP

probabilistic SVM

run SVM-transformed
logistic regression

fourth row: popular in **LIBSVM**

Map of Linear/Kernel Models

PLA/pocket

linear SVR

linear soft-margin
SVM

linear ridge
regression

regularized logistic
regression

kernel ridge
regression

kernel logistic
regression

SVM

SVR

probabilistic SVM

first row: less used due to **worse performance**
third row: less used due to **dense β**

Kernel Models

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? :-)**

Fun Time

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

Fun Time

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