Support Vector Regression

上一篇介绍对于logistic regression的kernel方法，这次介绍对于一般的回归使用kernel的方法

一、Kernel Ridge Regression

我们采用L2-正则化的时候需要解决这样一件事情。配合表示定理我们转化为关于β的问题，这里直接用的是square-error。

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 $\mathbf{w}$

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

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

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

Solving Kernel Ridge Regression

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

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

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

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

- $(\cdot)^{-1}$ always exists for $\lambda > 0$, because 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'

我们使用kernel matrix的test时候第一个矩阵是X_train,第二个是X_test，因为最终得到的g是用X_train里面的模型进行线性组合得到的，所以训练的时候需要用X_train,X_train，测试的时候用的是X_train,X_test

对比一下两个模型：

Linear versus Kernel Ridge Regression

linear ridge regression

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

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

kernel ridge regression

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

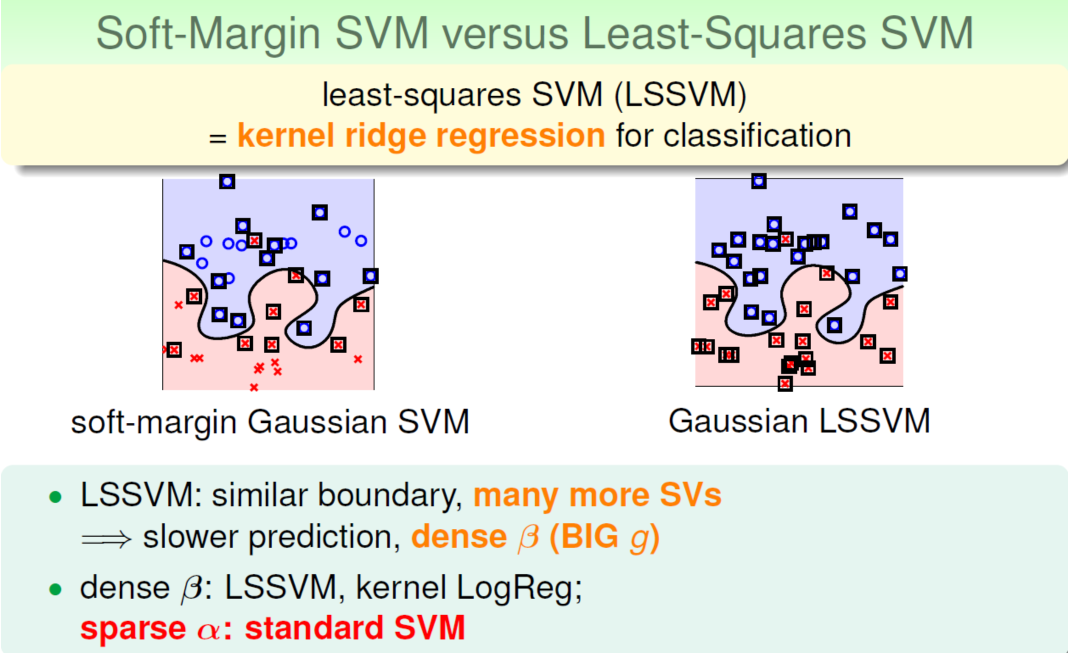
- **more flexible** with K
- $O(N^3)$ training; $O(N)$ prediction —hard for big data

linear versus **kernel**: trade-off between **efficiency** and **flexibility**

二、Support Vector Regression Primal

这里引入tube regression对于SVM standard进行微调

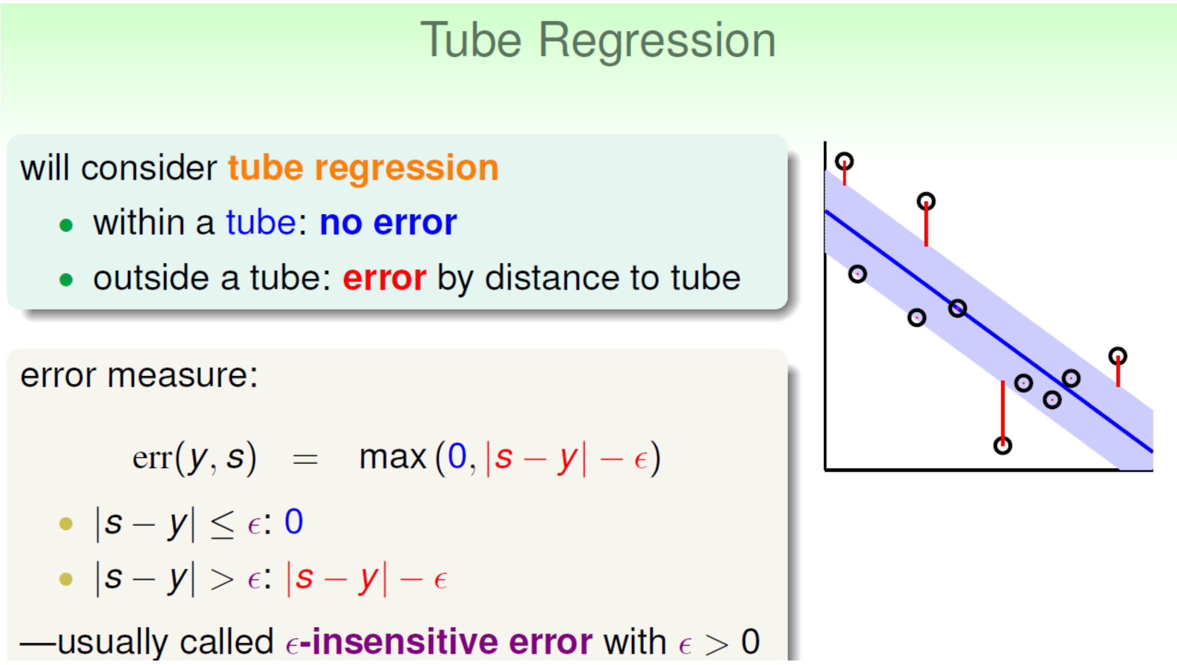
1. 对比一下Soft-Margin和Least Square两种情况的模型：



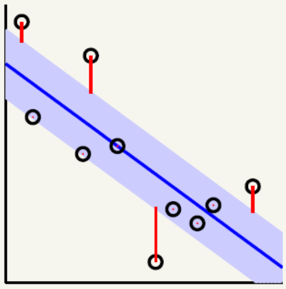
这里我们的β比较dense，我们希望有和standard SVM一致的sparse 特性，利用SV改进loss function

2. 考虑一下Tube Regression:

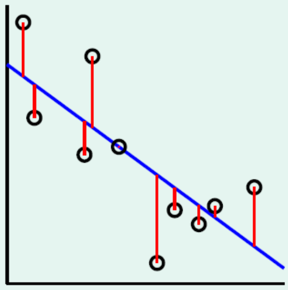
我们把cost function重新定义



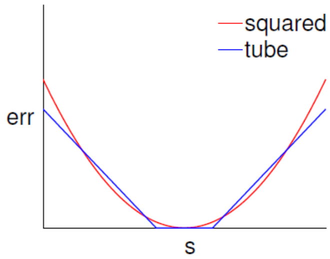
Tube versus Squared Regression



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.

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

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

standard SVM

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

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

will mimic standard SVM derivation:

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

我们模仿标准SVM的模型得到现在的模型

注意我们对于regularization的最小化是对于loss function而对于SVM本质上是对于margin的最小化，所以两者有差别！！！

Standard Support Vector Regression Primal

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

mimicking standard SVM

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

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

$$\xi_n \ge 0$$

making constraints linear

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

$$-\epsilon - \xi_n^{\vee} \le y_n - \mathbf{w}^T\mathbf{z}_n - b \le \epsilon + \xi_n^{\wedge}$$

$$\xi_n^{\vee} \ge 0, \xi_n^{\wedge} \ge 0$$

Support Vector Regression (SVR) primal:
minimize regularizer + (upper tube violations $\xi_n^{\wedge}$ & lower violations $\xi_n^{\vee}$)

这就是SVR primal，我们使用新的loss function而不是square或者一般的svm

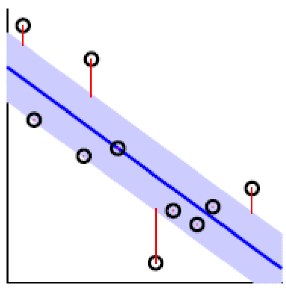
Quadratic Programming for SVR

$$\min_{b,\mathbf{w},\xi^{\vee},\xi^{\wedge}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\left(\xi_n^{\vee} + \xi_n^{\wedge}\right)$$

$$\text{s.t.} \quad -\epsilon - \xi_n^{\vee} \le y_n - \mathbf{w}^T\mathbf{z}_n - b \le \epsilon + \xi_n^{\wedge}$$

$$\xi_n^{\vee} \ge 0, \xi_n^{\wedge} \ge 0$$

- 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?

ϵ是可选的，这个是比standard多的变量

三、Support Vector Regression Dual

我们用使用Dual模型对它进行改进，同之前我们引入拉格朗日因子

Support Vector Regression — Support Vector Regression Dual

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

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

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

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

Some of the KKT Conditions

- $\frac{\partial\mathcal{L}}{\partial w_i} = 0$: $\mathbf{w} = \sum_{n=1}^{N} \underbrace{(\alpha_n^\wedge - \alpha_n^\vee)}_{\beta_n}\mathbf{z}_n$; $\frac{\partial\mathcal{L}}{\partial b} = 0$: $\sum_{n=1}^{N}(\alpha_n^\wedge - \alpha_n^\vee) = 0$
- complementary slackness:
$$\begin{aligned}\alpha_n^\wedge(\epsilon + \xi_n^\wedge - y_n + \mathbf{w}^T\mathbf{z}_n + b) &= 0\\ \alpha_n^\vee(\epsilon + \xi_n^\vee + y_n - \mathbf{w}^T\mathbf{z}_n - b) &= 0\end{aligned}$$

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

利用KKT条件我们推导一下。

我们看一下Dual相似性：

SVM Dual and SVR Dual

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

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

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

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

similar QP, **solvable by similar solver**

回到之前关于sparsity的解决

Sparsity of SVR Solution

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

$$\begin{aligned} \alpha_n^{\wedge}(\epsilon + \xi_n^{\wedge} - y_n + \mathbf{w}^T\mathbf{z}_n + b) &= 0 \\ \alpha_n^{\vee}(\epsilon + \xi_n^{\vee} + y_n - \mathbf{w}^T\mathbf{z}_n - b) &= 0 \end{aligned}$$

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

SVR: allows **sparse** β

SVR提供了sparse β

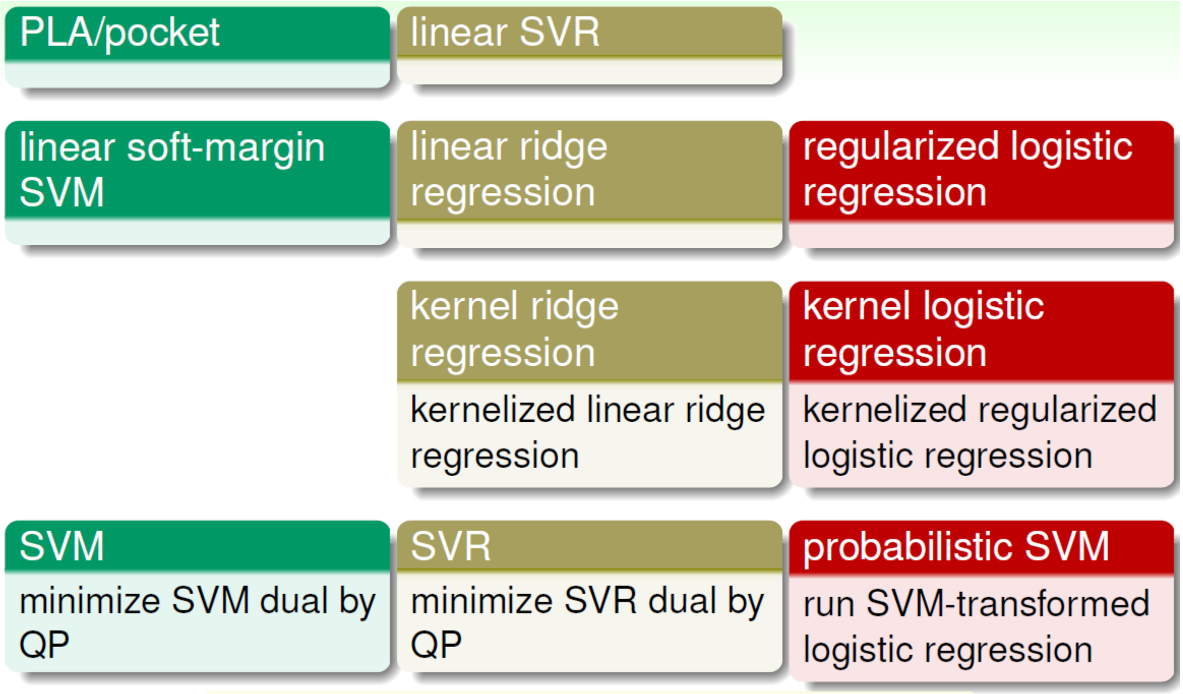
四、对于核方法的总结

1. 总结一下

首先是linear部分：

PLA/pocket minimize $\mathrm{err}_{0/1}$ specially	**linear SVR** minimize regularized $\mathrm{err}_{\mathrm{TUBE}}$ by QP	
linear soft-margin SVM minimize regularized $\widehat{\mathrm{err}}_{\mathrm{SVM}}$ by QP	**linear ridge regression** minimize regularized $\mathrm{err}_{\mathrm{SQR}}$ analytically	**regularized logistic regression** minimize regularized $\mathrm{err}_{\mathrm{CE}}$ by GD/SGD

然后我们有kernel形式的:



probabilistic SVM是2-level的方法，先SVM再logistic regression

2. 选择：

- 第一排不怎么用
- 第三排不怎么用，dense data!