Support Vector Regression

1.Kernel Ridge Regression

对于任何包含正则项的L2-regularized linear model，它的最佳化解w都可以写成是z的线性组合形式，因此也就能引入核技巧，将模型kernelized化。

for any L2-regularized linear model

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

optimal $\mathbf{w}_* = \sum_{n=1}^{N} \beta_n \mathbf{z}_n$.
—any L2-regularized linear model can be **kernelized**!

Kernel Ridge Regression:

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

最佳解w_*肯定是z的线性组合
把$w_* = \sum_{n=1}^{N} \beta_n z_n$代入到ridge regression中，将z的内积用kernel替换，把求w_*的问题转化成求β_n的问题：

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

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

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

第一项可以看成是β_n的正则项
第二项可以看成是β_n的error function
求解该式最小化对应的β_n值，解决了kernel ridge regression问题。

求解β_n的问题可以写成如下形式：

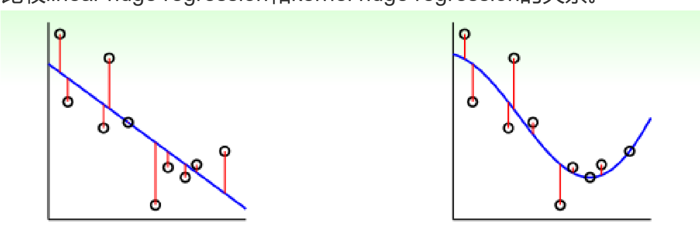
$$\begin{aligned} E_{\text{aug}}(\beta) &= \frac{\lambda}{N}\beta^T\mathrm{K}\beta + \frac{1}{N}\left(\beta^T\mathrm{K}^T\mathrm{K}\beta - 2\beta^T\mathrm{K}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right) \\ \nabla E_{\text{aug}}(\beta) &= \frac{2}{N}\left(\lambda\mathrm{K}^T\mathrm{I}\beta + \mathrm{K}^T\mathrm{K}\beta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

$E_{aug}(\beta)$是关于β的二次多项式，要对$E_{aug}(\beta)$求最小化解，这种凸二次最优化问题，只需要先计算其梯度，再令梯度为零即可。
令$\nabla E_{aug}(\beta)$等于零，得到：

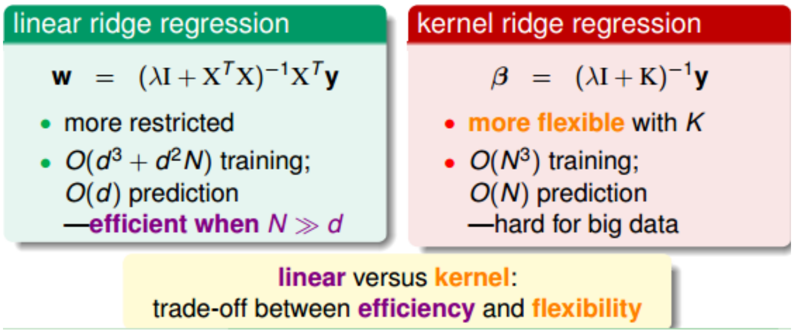
$$\beta = (\lambda I + K)^{-1} y$$

且$(\lambda I + K)$的逆矩阵的逆矩阵一定存在。因为核函数K满足Mercer's condition，它是半正定的，且$\lambda > 0$。

比较linear ridge regression和kernel ridge regression的关系。



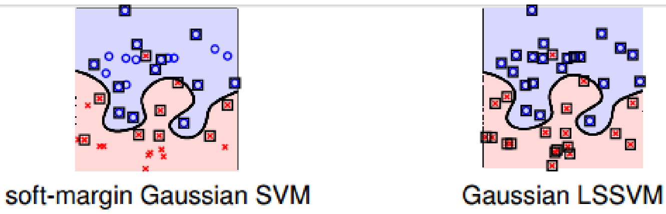
左	右
linear ridge regression	kernel ridge regression
线性模型，只能拟合直线	非线性模型，更灵活
训练复杂度$O(d^3+d^2N)$	训练复杂度$O(N^3)$
预测复杂度$O(d)$	预测复杂度$O(N)$



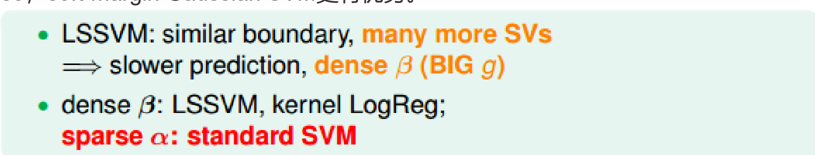
2.Support Vector Regression Primal

kernel ridge regression应用在classification上叫做least-squares SVM(LSSVM)

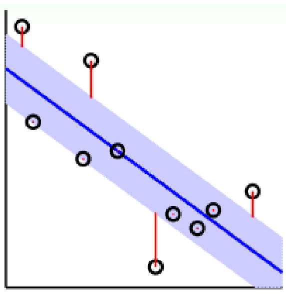
比较一下soft-margin Gaussian SVM和Gaussian LSSVM在分类上的差异：



左边soft-margin Gaussian SVM的SV不多，而右边Gaussian LSSVM中基本上每个点都是SV。
因为soft-margin Gaussian SVM中的α_n大部分是等于零，$\alpha_n > 0$的点只占少数，所以SV少。
而对于LSSVM，β的解大部分都是非零值，所以对应的每个点基本上都是SV。
SV太多会带来一个问题，就是做预测的矩$g(x)=\sum_{n=1}^{N}\beta_n K(x_n,x)$，如果$\beta_n$非零值较多，那么g的计算量也比较大，降低计算速度。
so，soft-margin Gaussian SVM更有优势。



可以通过一些方法得到sparse β，使得SV不会太多，从而得到和soft-margin SVM同样的分类效果。



引入一个叫做Tube Regression的做法，即在分类线上下分别划定一个区域（中立区）
如果数据点分布在这个区域内，则不算分类错误，只有误分在中立区域之外的地方才算error。
假定中立区的宽度为2ϵ，$\epsilon > 0$,那么error measure就可以写成：$err(y,s)=max(0,|s-y|-\epsilon)$，对应上图中红色标注的距离。

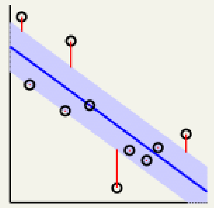
error measure:

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

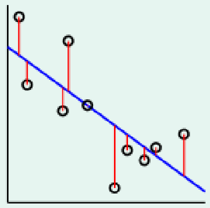
- $|s - y| \leq \epsilon$: 0
- $|s - y| > \epsilon$: $|s - y| - \epsilon$

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

把tube regression中的error与squared error做个比较：

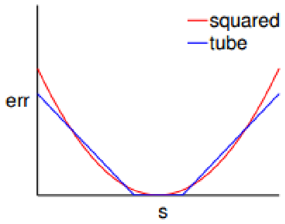


tube: $\text{err}(y, s) = \max(0, |s - y| - \epsilon)$



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

将err(y,s)与s的关系曲线分别画出来：



tube ≈ squared when $|s - y|$ small
& less affected by outliers

当|s-y|比较小即s比较接近y的时候，squared error与tube error是差不多大小的。
而在|s-y|比较大的区域，squared error的增长幅度要比tube error大很多。
error的增长幅度越大，表示越容易受到noise的影响，不利于最优化问题的求解。
所以，从这个方面来看，tube regression的这种error function要更好一些。

L2-Regularized Tube Regression:

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

上式，由于其中包含max项，并不是处处可微分的，所以不适合用GD/SGD来求解。
而且，虽然满足representer theorem，有可能通过引入kernel来求解，但是也并不能保证得到sparsity β。
从另一方面考虑，我们可以把这个问题转换为带条件的QP问题，仿照dual SVM的推导方法，引入kernel，得到KKT条件，从而保证解β是sparse的。

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**

所以，我们就可以把L2-Regularized Tube Regression写成跟SVM类似的形式：

will mimic standard SVM derivation:

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

已经有了Standard Support Vector Regression的初始形式，这还是不是一个标准的QP问题。
继续对该表达式做一些转化和推导：

mimicking standard SVM

$$\min_{b,\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n$$
$$\text{s.t. } |\mathbf{w}^T\mathbf{z}_n + b - y_n| \le \epsilon + \xi_n$$
$$\xi_n \ge 0$$

making constraints linear

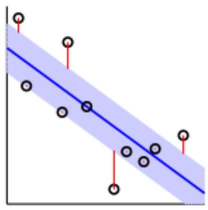
$$\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\left(\xi_n^\vee + \xi_n^\wedge\right)$$
$$-\epsilon - \xi_n^\vee \le y_n - \mathbf{w}^T\mathbf{z}_n - b \le \epsilon + \xi_n^\wedge$$
$$\xi_n^\vee \ge 0, \xi_n^\wedge \ge 0$$

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

SVR的标准QP形式包含几个重要的参数：C和ϵ。
C表示的是regularization和tube violation之间的权衡。
large C倾向于tube violation，small C则倾向于regularization。

ϵ表征了tube的区域宽度，即对错误点的容忍程度。
ϵ越大，则表示对错误的容忍度越大。
ϵ是可设置的常数，是SVR问题中独有的，SVM中没有这个参数。
另外，SVR的QP形式共有$\hat{d}+1+2N$个参数，2N+2N个条件。

- 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



3.Support Vector Regression Dual

先令拉格朗日因子$\alpha^\vee$和$\alpha^\wedge$，分别是与$\xi_n^\vee$和$\xi_n^\wedge$不等式相对应。

objective function		$\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\left(\xi_n^\vee + \xi_n^\wedge\right)$
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	$-\epsilon - \xi_n^\vee \le y_n - \mathbf{w}^T\mathbf{z}_n - b$

然后，与SVM一样做同样的推导和化简，拉格朗日函数对相关参数偏微分为零，得到相应的KKT条件：

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

通过观察SVM primal与SVM dual的参数对应关系，直接从SVR primal推导出SVR dual的形式

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

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

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

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

SVR dual形式下推导的解w为：

$$w=\sum_{n=1}^{N}(\alpha_n^{\wedge}-\alpha_n^{\vee})z_n$$

相应的complementary slackness为：

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

对于分布在tube中心区域内的点，满足$|w^Tz_n+b-y_n|<\epsilon$，此时忽略错误，$\xi_n^{\vee}$和$\xi_n^{\wedge}$都等于零。
则complementary slackness两个等式的第二项均不为零，必然得到$\alpha_n^{\wedge}=0$和$\alpha_n^{\vee}=0$，即$\beta_n=\alpha_n^{\wedge}-\alpha_n^{\vee}=0$。

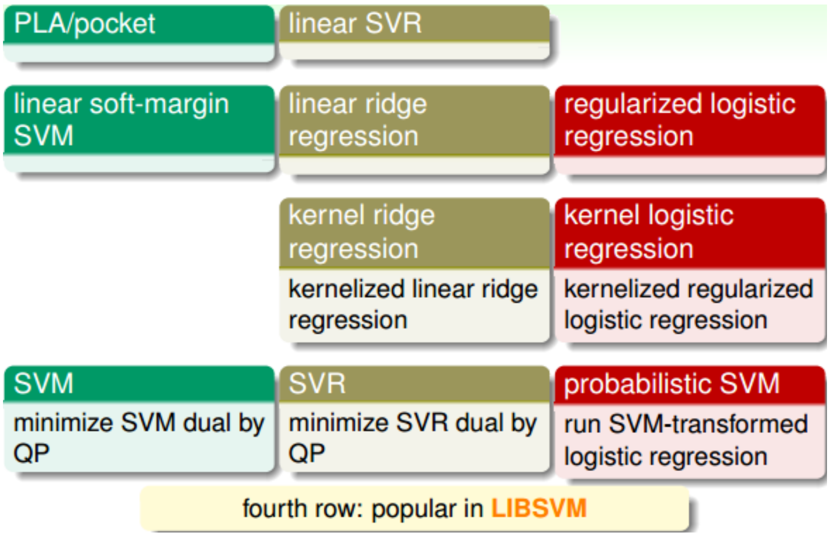
所以，对于分布在tube内的点，得到的解$\beta_n=0$，是sparse的。
而分布在tube之外的点，$\beta_n\neq 0$。
至此，我们就得到了SVR的sparse解。

4.Summary of Kernel Models

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

second row: popular in **LIBLINEAR**

上图中相应的模型也可以转化为dual形式，引入kernel，整体的框图如下：


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