SVM & SVR

Support Vector Machine & Support Vector Regression

- Maximum margin classifier
- Primal and dual forms
- Linearly seperable data
- Not linearly seperable data
- Slack variables and hinge loss
- Epsilon-insensitive loss

When we use SVM:

Sparse data
(include under-determined systems)
$N_c < (3\text{-}10) \times$ d.o.f

Adding constraints to learning problem

Non linear mapping to expanded feature space

Use kernels $k(\underline{x}, \underline{x}')$

I. Concept (non-augmented notation & space)

① Non-linear mapping $\underline{x} \xrightarrow{\text{N.L. mapping}} \underline{u} = \underline{\phi}(\underline{x})$ ← Assume data is linearly seperable

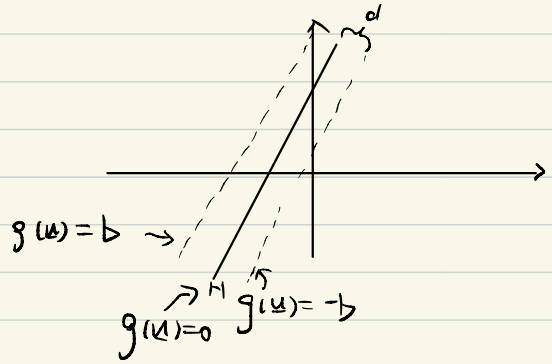
② Optimized version of using a margin

$g(\underline{x}) = \underline{w}^T\underline{x} + w_0 \geq b > 0$ for $\underline{x} \in S_1$

$g'(\underline{u}) = \underline{w}'^T\underline{u} + w_0 \geq b' > 0$ for $\underline{u} \in S_1$ // $\phi(\underline{x})$ of $k(\underline{x}, \underline{x}')$

unaugmented

$$\Rightarrow g(\underline{u}_n) = \begin{cases} \underline{w}^T\underline{u}_n + w_0 \geq b > 0 & \forall \underline{u}_n \in S_1 \\ \underline{w}^T\underline{u}_n + w_0 \leq -b < 0 & \forall \underline{u}_n \in S_2 \end{cases}$$



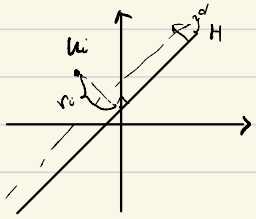
$$d = d_S(H, u_0) = \frac{g(u)}{\|w\|} = \frac{b}{\|w\|}$$

$g(u) = w^T u + w_0$

augmented reflected

$g(u) = w^{(+)T} z_i u_i^{(+)} > 0 \Rightarrow u_i$ is correctly classified.

$\therefore z_i(w^T u_i + w_0) > 0 \ \forall i \Rightarrow$ all data points correctly classified.



Distance:

$r_i = d_s(H, u_i) = \frac{z_i(w^T u_i + w_0)}{\|w\|} > 0$

want: $r_i \geq d > 0$

$\therefore$ Require: (1) $\begin{cases} \frac{z_i(w^T u_i + w_0)}{\|w\|} \geq d > 0 & \forall i \\ \text{with maximum possible } d \end{cases}$

assume L.S. in u space ← for training data set

Maximum margin classifier

II. SVM Learning

RHS of (1): $d = \frac{b}{\|w\|}$

max $d \Rightarrow$ max b, & min $\|w\|$

if Fix b then minize. $\|w\|$ for given b

$\therefore$ (1) $\Rightarrow$ Require $z_i(w^T u_i + w_0) \geq 1 \ \forall i$, with minimal $\|w\|$

Use Lagrange Optimization

(2) $\begin{cases} \text{Minimize } J(w) = \|w\|^2 \\ \text{subject to constraints: } z_i(w^T u_i + w_0) - 1 \geq 0 \ \forall i \end{cases}$

Note: Use λ instead of μ for langrange multipliers

Primal Form:

(2)' $L(w, w_0, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i [z_i(w^T u_i + w_0) - 1]$

$\lambda_i \geq 0, \ \forall i$

$\lambda_i [z_i(w^T u_i + w_0) - 1] = 0 \ \forall i$ } KKT conditions

$z_i(w^T u_i + w_0) - 1 \geq 0 \ \forall i$

Primal form of lagrangian

$w^* \rightarrow$ solution

Dual form

Easier to use if express its dual representation

Solve: $\begin{cases} \nabla_w L(w, w_0, \lambda) = 0 \\ \text{and } \frac{\partial}{\partial w_0} L(w, w_0, \lambda) = 0 \end{cases}$ for w^* $\Rightarrow$ to get Min(w) & Min(w_0)

(2'') $L_D(\lambda) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j z_i z_j u_i^T u_j + \sum_{i=1}^{N} \lambda_i$

with: $\sum_{i=1}^{N} a_i z_i = 0$

$\lambda_i \geq 0, \ \lambda_i [z_i(w^{*T} u_i + w_0) - 1] = 0 \ \forall i$

$w^* = \sum_{i=1}^{N} \lambda_i z_i u_i$

$z_i(w^T u_i + w_0) - 1 \geq 0$

← DUAL FORM

$\Rightarrow L_D(\underline{\lambda}) = L(\underline{w}^*, w_0^*, \underline{\lambda})$ = L optimized w.r.t. $\underline{w}, w_0$

Comments.

1. $L_D(\underline{\lambda}) \Rightarrow$ only need to optimize w.r.t. $\underline{\lambda}$; yields solution $\underline{\lambda}^*$
2. In $L_D(\underline{\lambda})$, $\underline{u}$ appears in terms of what form?
   $\lambda_i \lambda_j (z_i u_i)^T (z_j u_j) \rightarrow$ form of a kernel
   $\Rightarrow$ Can define nonlinear transformation by a choice of kernel $k(\underline{x}_i, \underline{x}_j)$

$\nabla_{\underline{w}} L = \frac{1}{2} \cdot 2\underline{w} - \sum_{i=1}^{N} \lambda_i z_i u_i = 0$

$\therefore \underline{w}^* = \sum_{i=1}^{N} \lambda_i z_i u_i$ — ①

$\therefore \frac{\partial L}{\partial w_0} = \frac{-\sum_{i=1}^{N} \lambda_i z_i w^T u_i + \lambda_i z_i w_0 + \lambda_i}{\partial w_0} = \sum_{i=1}^{N} \lambda_i z_i = 0$ — ②

substitute ① ② into ①,

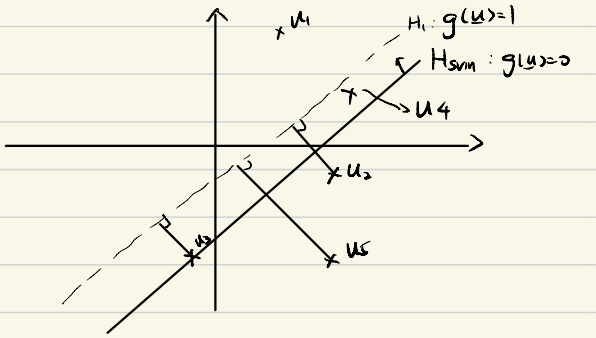
$L_D = \frac{1}{2} \| \sum_{i=1}^{N} \lambda_i z_i u_i \|^2 - \sum_{i=1}^{N} (\lambda_i z_i w^T u_i + \lambda_i z_i w_0 - \lambda_i)$

$= \frac{1}{2} \| \sum_{i=1}^{N} \lambda_i z_i u_i \|^2 - \sum_{i=1}^{N} \lambda_i z_i (\sum_{j=1}^{N} \lambda_j z_j u_j)^T u_i + \sum_{i=1}^{N} \lambda_i$

$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j z_i z_j u_j^T u_i + \sum_{i=1}^{N} \lambda_i$

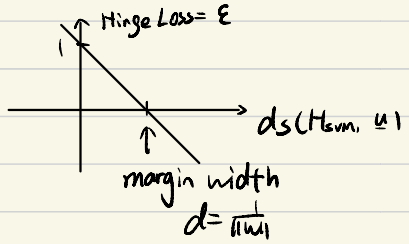
→ steps from (2') to (2'')

→ Introduce slack variables ξ_i, such that (ξ is xi)
$\xi_i = 0$ if $\underline{u}_i$ is on correct side of margine boundary
ξ_i = normalized distance to boundary, if on incorrect side of margin boundary
$\xi_i = 1$ if $\underline{u}_i$ is on decision boundary



$\xi_1 = 0$
$\xi_2 = 2$
$\xi_3 = 1$
$\xi_4 = 0.5$
$\xi_5 = 3.5$

Loss function:



Now, (2) becomes:

(3) $\begin{cases} \text{Minimize} & J = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{constraints:} & z_i (w^T u_i + w_0) \geq 1 - \xi_i \quad \forall i \end{cases}$

$\therefore$

L becomes

(3') $L(\underline{w}, w_0, \underline{\xi}, \underline{\lambda}, \underline{\mu}) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \lambda_i [z_i (\underline{w}^T \underline{u}_i + w_0) - 1 + \xi_i] - \sum_{i=1}^{N} \mu_i \xi_i$

$\lambda_i \geq 0 \quad \lambda_i^* [z_i (\underline{w}^{*T} \underline{u}_i + w_0) - 1 + \xi_i] = 0 \quad \forall i$

$\mu_i \geq 0 \quad \mu_i^* \xi_i^* = 0 \quad \forall i$

$z_i (\underline{w}^T \underline{u}_i + w_0) - 1 + \xi_i \geq 0, \ \xi_i \geq 0 \quad \forall i$

↑ Primal

From L, we can derive L_D:

$\Rightarrow L_D(\underline{\lambda})$ has same L_D for seperable case. except constraint $\lambda_i \geq 0 \ \forall i \Rightarrow 0 \leq \lambda_i \leq C$

$$\therefore L_D(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j \underbrace{z_i z_j}_{z_i z_j \ \phi(x_i)\phi(x_j)} \underline{u}_i^T \underline{u}_j$$

$\therefore k(x_i, x_j) = \phi^T(x_i)\phi(x_j)$

$\Rightarrow$ we can

① choose a kernel fuction

② subtitute kernel function

$\nabla_w L = \frac{1}{2} \cdot 2\underline{w} - \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i = 0$

$\therefore \underline{w}^* = \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i$ ①

$\therefore \frac{\partial L}{\partial w_0} = -\sum_{i=1}^{N} \lambda_i z_i \underline{w}^T \underline{u}_i + \lambda_i z_i w_0 + \lambda_i z_i \varepsilon_i = \sum_{i=1}^{N} \lambda_i z_i = 0$ ②

$\frac{\partial L}{\partial \varepsilon_i} = C \sum_{i=1}^{N} \varepsilon_i^0 + C - \sum_{i=1}^{N} \lambda_i z_i \varepsilon_i^0 - \lambda_i - \sum_{i=1}^{N} \mu_i \varepsilon_i^0 - \mu_i$

$= C - \lambda_i - \mu_i = 0$ ③

Substitute ①②③

$L = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j z_i z_j \underline{u}_i^T \underline{u}_j + \sum_{i=1}^{N} \lambda_i + C \sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \lambda_i z_i \varepsilon_i - \sum_{i=1}^{N} \mu_i \varepsilon_i$

$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j z_i z_j \underline{u}_i^T \underline{u}_j + \sum_{i=1}^{N} \lambda_i + \varepsilon_i (C - \lambda_i - \mu_i)$

$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j z_i z_j \underline{u}_i^T \underline{u}_j + \sum_{i=1}^{N} \lambda_i$

with ① $C - \mu_i - \lambda_i = 0$

② $\mu_i \geq 0$

③ $\lambda_i \geq 0$

④ $\mu_i \varepsilon_i = 0 \ \forall i$

⑤ $\lambda_i [z_i(\underline{w}^T \underline{u}_i + w_0) - 1 + \varepsilon_i] = 0$

SVR: Support Vector Regression

SVM adapted to regression problems $\Big\}$ $\underline{x} \rightarrow$ NL mapping $\phi(\underline{x})$ or $k(\underline{x}, \underline{x}') \rightarrow \underline{u} \rightarrow$ [Linear regressor] $\rightarrow \hat{y}$

Motivation: to have a regression model suitable to sparse data (e.g. high D, low N)

In Ridge Regression, we used a criterion fcn: MSE + l_2 regularizer

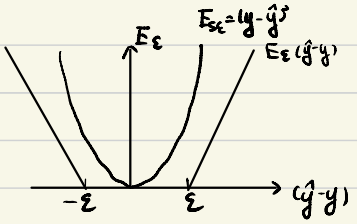
$$J_{RR}(\underline{w}) = \frac{1}{n} \sum_{n=1}^{N} \underbrace{[\hat{y}(\underline{x}_n) - y_n]^2}_{\text{use "}\varepsilon\text{-insensitive" loss instead of squared-error loss}} + \lambda \|\underline{w}\|_2^2, \ \lambda \geq 0$$

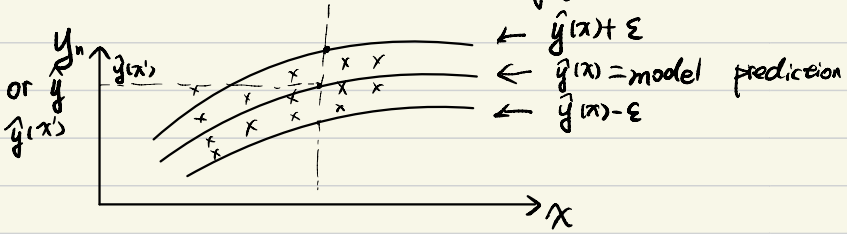
keep L_2 regularizer

For SVR:

ε-insensitive-loss E_ε.

$$E_\varepsilon(\hat{y} - y) = \begin{cases} 0 & \text{if } |\hat{y} - y| < \varepsilon \\ |\hat{y} - y| - \varepsilon & \text{if } |\hat{y} - y| \geq \varepsilon \end{cases}$$





- 0 error for predictions $\hat{y}(\underline{x}_n)$ within ε of known output y_n, means learning algorithm is less likely to overfit to noise or small variations in data.