Recap of Soft-SVM

$$\min_{w,b} \frac{\|w\|^2}{2} + C\|\xi\|_1$$

subject to (A) $y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$

(B) $\xi_i \geq 0 \quad \forall\, i \in [m]$

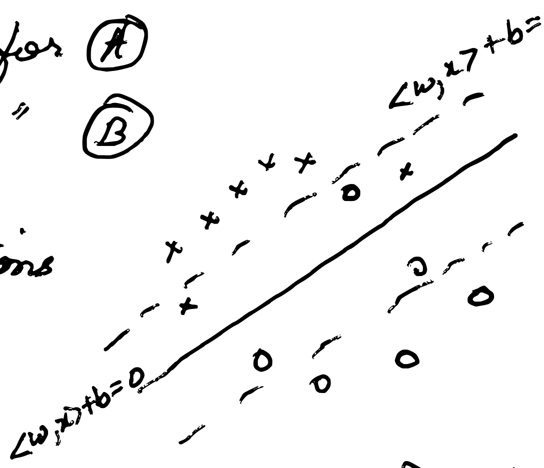
From KKT conditions

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

x_i : support vectors for $\alpha_i \neq 0$

α_i : Dual variables for (A)

β_i : " " " (B)



Other KKT conditions

$$\alpha_i + \beta_i = C$$

$\alpha_i = 0$ or $y_i(\langle w, x_i\rangle + b) = 1 - \xi_i$

$\beta_i = 0$ or $\xi_i = 0$ ←

$\alpha_i \neq 0 \Rightarrow y_i(\langle w, x_i\rangle + b) = 1 - \xi_i$

$\xi_i = 0$

$y_i(\langle w, x_i\rangle + b) = 1$

x_i are on marginal hyperplane

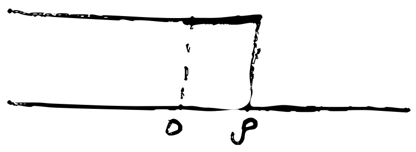
$\xi_i > 0$

$\beta_i = 0$

$\alpha_i = C$

$y_i(\langle w, x_i\rangle + b) = 1 - \xi_i$

$$\gamma(x,y,w,b) = y(\langle w,x\rangle + b)$$



$$l_{hinge}(\underbrace{(x,y)}_{z},(w,b)) = \max\left\{0,\ 1-\frac{\gamma}{\rho}\right\}$$

$\rho = 1$

$$l_{quad\text{-}hinge} = (l_{hinge})^2$$

$$l_{ramp}((x,y),(w,b))$$

$$= \min\left\{1,\ \max\left\{0,\ 1-\frac{\gamma}{\rho}\right\}\right\}$$

$$\boxed{1_{\{\gamma((x,y),(w,b))\leq 0\}} \leq l_{ramp}((x,y),(w,b)) \leq 1_{\{\gamma((x,y),(w,b))\leq \rho\}}}$$

$(\rho > 0)$

$$\hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} l_{ramp}(z_i,h)$$

$z_i = (x_i, y_i)$

$$h \in \mathcal{H} = \{h(\cdot,w,b) : h(x,w,b) = \langle w,x\rangle + b\}$$

$$R_\rho(h) = \mathbb{E}_{S\sim\mathcal{D}^m} \hat{R}_{S,\rho}(h)$$

(Form of) Generalization bound with $\Pr \geq 1-\delta$ over $S \sim \mathcal{D}^m$

$$R_\rho(h) \leq \hat{R}_{S,\rho}(h) + \sqrt{\frac{f(\mathcal{H},\mathcal{D})}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}$$

$$R(h) \leq \hat{R}_{S,\rho}(h) + \cdots$$

(generalization for 0-1 loss)

$$\leq \hat{R}_{S,\rho,hinge}(h) + \cdots$$

Generalization bound based on Rademacher Complexity

Koltchinskii & Panchenko 2002
Bartlett and Mendelson 2002

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\,\mathrm{Rad}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

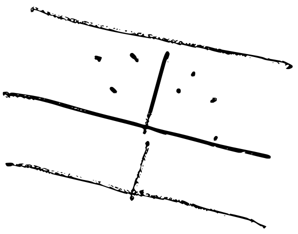
for

$$\mathcal{H} = \{ x \to \langle \omega, x \rangle : \|\omega\| < \Lambda \}$$

(0-1) $\forall h \in \mathcal{H}$

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\frac{\Lambda r}{\sqrt{m}} + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

$\frac{\Lambda r}{\rho}$ is small and at the same time $\hat{R}_{S,\rho}(h)$ (hinge loss) is small $\Rightarrow$ good generalization



Want:

$$\mathrm{Rad}_S(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{m}}$$

where $r = \sup_{x \sim D} \|x\|$ ✓

$\Lambda = \sup_{w} \|w\|$ ✓

$$\mathrm{Rad}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)$$

$\sigma = \{\sigma_1, \dots, \sigma_m\}$ iid

$\mathcal{H} = \{h(\cdot, w) : h(x, w) = \langle w, x\rangle, \|w\| \leq \Lambda\}$

Proof

$$\mathrm{Rad}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma \sup_{\|w\|\leq\Lambda} \sum_{i=1}^{m} \sigma_i \langle w, x_i\rangle$$

$$= \frac{1}{m} \mathbb{E}_\sigma \sup_{\|w\|<\Lambda} \langle w, \sum_{i=1}^{m} \sigma_i x_i\rangle$$

$(|\langle a,b\rangle| \leq \|a\|\,\|b\|)$

$$\leq \frac{\Lambda}{m} \mathbb{E}_\sigma \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \rightarrow (*)$$

[(Jensen's inequality: $\mathbb{E}f(X) \geq f(\mathbb{E}X)$ when f is convex)

$\mathbb{E}X^2 \geq (\mathbb{E}X)^2$

$X = \left\|\sum_{i=1}^{m} \sigma_i x_i\right\|$

$$\left(\mathbb{E}_\sigma \left\|\sum_{i=1}^{m}\sigma_i x_i\right\|^2\right)^{1/2} \geq \left(\mathbb{E}_\sigma \left\|\sum_{i=1}^{m}\sigma_i x_i\right\|\right) \rightarrow (*)]$$

$$\mathrm{Rad}_S(\mathcal{H}) \leq \frac{\Lambda}{m}\left(\mathbb{E}_\sigma \left\|\sum_{i=1}^{m}\sigma_i x_i\right\|^2\right)^{1/2}$$

$$= \frac{\Lambda}{m}\left(\mathbb{E}_\sigma \sum_{j,i=1}^{m} \sigma_i\sigma_j \langle x_i, x_j\rangle\right)^{1/2}$$

(σ_i iid $\quad \mathbb{E}[\sigma_i\sigma_j] = \mathbb{E}[\sigma_i]\,\mathbb{E}[\sigma_j] = 0$)

$$\leq \frac{\Lambda}{m}\left(\sum_{i=1}^{m} \|x_i\|^2\right)^{1/2}$$

$$\leq \frac{\Lambda}{m}(r^2 m)^{1/2} = \frac{r\Lambda}{\sqrt{m}}$$

$$\boxed{\mathrm{Rad}_S(\mathcal{H}) \leq \frac{r\Lambda}{\sqrt{m}}}$$

Rademacher complexity

$$\mathrm{Rad}_m(\mathcal{H}) = \mathbb{E}_{S\sim D^m} \mathrm{Rad}_S(\mathcal{H})$$

Generalization of SVM:

$$R(h) \le \underline{\hat{R}_{S,\rho}(h)} + \frac{2}{\rho}\underline{\sqrt{\frac{r^2\Lambda^2}{m}}} + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

for all $h \in \mathcal{H} = \{x \to \langle w, x \rangle : \|w\| \le \Lambda\}$

and $\|x\| < r$

with probability at least $1-\delta$ over $S \sim D^m$

$$\frac{r\Lambda}{\rho} \quad \text{vs} \quad \hat{R}_{S,\rho}(h)$$

Activity tasks

→ trains a SVM

Components

→ Training data

→ $\mathcal{H}$

→ Loss ←

→ SGD ;

→ hyperparameters
(problem definition:
algorithm/optimizer)

Code

c : loss

lr : algorithm