Lecture 12 FML April 1, 21

Multiclass classification

$\mathcal{X}$ input space, $\mathcal{Y}$ output space

K no of classes

Sample $S = (x_1, y_1), \ldots, (x_m, y_m)$

y_i ← labels

$x_i \overset{iid}{\sim} D$ D unknown

$y_i = f(x_i)$ f unknown

- we want to know f from sample S.

$\mathcal{H}$ = Hypothesis class.

Set of functions from which we search for f

$R(h) = E_D[1\{h(x) \neq f(x)\}]$

↑ generalization error

$h \in \mathcal{H}$.

- Aim is to get $h \in \mathcal{H}$ which has min $R(h)$.
- We don't know D
  ∴ we don't know R
- $\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq f(x_i)\}$

$d_H(h(x_i), f(x_i))$

↑ distance

Binary classification via SVM & kernels

$\mathcal{Y} = \{-1, 1\}$

$h(x) = \text{sgn}(w \cdot x + b)$

(w, b)

$= \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x + b)\right)$

$= \underset{y \in \{-1,1\}}{\text{argmax}} \; (y \cdot (w \cdot x + b))$

Kernel. K. $\Phi : x \to \mathbb{H}$.

$K(x, x') = \langle \Phi(x), \Phi(x') \rangle$

$\langle \Phi(x), \Phi(x') \rangle = K(x, x')$.

$h(x) = \underset{y \in \{-1,1\}}{\text{argmax}} \langle w_y, \Phi(x) \rangle$

$w_y = \sum_{i=1}^{m} y \alpha_i y_i \Phi(x_i)$

For $K \geq 2$ classes

$h(x) = \underset{y \in \mathcal{Y} = \{1, 2, \ldots K\}}{\text{argmax}} h(x, y)$

$h(x, y) =$ score function

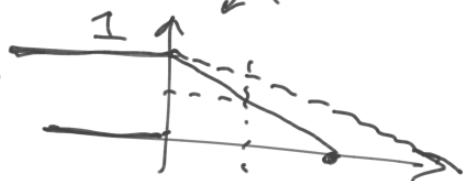
$h \in \mathcal{H}$.

Margin based

$h \in H$, $\rho_h(x,y) = h(x,y) - \max_{y' \neq y} h(x,y')$

(x,y) — h misclassifies x if $\rho_h(x,y) \leq 0$

For a given $\rho > 0$ define margin loss fn $\Phi_\rho(u)$



$\hat{R}_\rho(h)$ = Empirical margin loss

$= \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(\rho_h(x_i, y_i))$

$\Pi_1(H) = \{x \mapsto h(x,y), y \in Y, h \in H\}$.

Theorem Let H be a set of functions $X \times Y \to \mathbb{R}$. Fix $\rho > 0$. Then for any $\delta > 0$, with prob $\geq 1-\delta$, $\forall h \in H$

$$R_\rho(h) \leq \hat{R}_\rho(h) + \frac{2|Y|^2}{\rho} R_m(\Pi_1(H)) + \left(\frac{\log 1/\delta}{2m}\right)^{1/2}.$$

Theorem A family of fn $G : Z \to [0,1]$ $\leftarrow$

$z, z_1, z_2, \ldots$ i.i.d r.v.s with values in Z. Then for any $\delta > 0$, with prob $\geq 1-\delta$, $\forall\, g \in G$

$$E[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2R_m(G) + \left(\frac{\log\frac{1}{\delta}}{2m}\right)^{\frac{1}{2}}$$

$$E[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{R}_S(G) + 3\left(\frac{\log\frac{1}{\delta}}{2m}\right)^{1/2}.$$

Lemma (Talagrand) $\Phi : \mathbb{R} \to \mathbb{R}$, l-lip

$\mathcal{H}$ Hypothesis set of real valued fn.

Then

$$\hat{R}_S(\Phi \circ \mathcal{H}) \leq l\,\hat{R}_S(\mathcal{H}).$$

$$Rm(\Phi \circ \mathcal{H}) \leq l\, Rm(\mathcal{H}).$$

Proof of Theorem

$\tilde{\mathcal{H}} = \{(x,y) \mapsto \rho_h(x,y),\ h \in \mathcal{H}\}$

$\hat{\mathcal{H}} = \{\Phi_\rho \circ f,\ f \in \tilde{\mathcal{H}}\}$

From theorem given above with prob $\geq 1-\delta$, $\forall\, h \in \mathcal{H}$

$$R_\rho(h) = E\{\Phi_\rho(\rho_h(x,y))\] \leq \hat{R}_\rho(h) + 2Rm(\hat{\mathcal{H}}) + \left(\frac{\log\frac{1}{\delta}}{2m}\right)^{\frac{1}{2}}$$

From Talagrand Lemma

since Φ_ρ is $\frac{1}{\rho}$ – Lip fn.

$$Rm(\hat{\mathcal{H}}) \leq \frac{1}{\rho} Rm(\tilde{\mathcal{H}})$$

It is shown in Mohri's Book

$$Rm(\mathcal{H}) \leq K^2 Rm(\Pi_1(H))$$

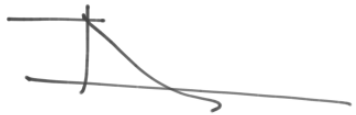
$$R_\rho(h) \leq \hat{R}_\rho(h) + \frac{2K^2}{\rho} Rm(\Pi_1(H)) + \ldots$$

$R(h)$ = generalization error

$= E_D\{ 1_{h(x) \neq f(x)} \}$

$R_\rho(h) = E_D\{ \Phi_\rho(x,y) \}$.

error occurs when $\rho_h(x,y) \leq 0$.



$\therefore \quad R(h) \leq R_\rho(h) \leq \cdots$

$$R_\rho(h) \leq \hat{R}_\rho(h) + \frac{2k^2}{\rho} R_m(\Pi_1(H)) + \left(\frac{\log \frac{1}{\delta}}{2m}\right)^{\frac{1}{2}}$$

$4k$.

We compute $R_m(\Pi_1(H))$ for SVM kernels.

Ex For k class

K, Φ

$$\underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; w_y \cdot \Phi(x)$$

$\mathcal{Y} \to \{-1, 1\}$.

$\Phi : \mathcal{X} \to \mathbb{H}$.

$W = (w_1, w_2, \cdots, w_k)$

$w_i \in \mathbb{H}$.

For $p \geq 1$

$$\|W\|_{\mathbb{H},p} \triangleq \left(\sum_{l=1}^{k} \|w_l\|_{\mathbb{H}}^p \right)^{1/p}.$$

$$\mathcal{H}_{k,p} = \{ (x,y) \mapsto \langle w_y, \Phi(x) \rangle, \; \|W\|_{\mathbb{H},p} \leq \Lambda \}$$

Prop Let k be PDS kernel with $k(x,x) \le r^2$ for some $r > 0$. Then for $m \ge 1$

$$R_m(\pi_1(\mathcal{H}_{k,p})) \le \left(\frac{r^2 \Lambda^2}{m}\right)^{1/2}$$

Proof: S sample size $(x_1, y_1), \ldots, (x_m, y_m)$

$\|w\|_{\mathcal{H},p} \le \Lambda$, $\|w\|_{\mathcal{H},p} = \left(\sum_{l=1}^{k} \|w_l\|_{\mathcal{H}}^p\right)^{\frac{1}{p}}$.

$\therefore \left(\|w_l\|_{\mathcal{H}}^p\right)^{1/p} \le \Lambda$ //

$$R_m(\pi_1(\mathcal{H}_{k,p})) = \frac{1}{m} E_{\sigma,s}\left[\sup_{g \in G} \sum_{i=1}^{m} \sigma_i g(z_i)\right]$$

$$= \frac{1}{m} E_{\sigma,s}\left[\sup_{\substack{y \in Y \\ \|w\| \le \Lambda}} \sum_{i=1}^{m} \sigma_i \langle w_y, \Phi(x_i)\rangle\right]$$

$$= \frac{1}{m} E_{\sigma,s}\left[\sup_{y \le y} \langle w_y, \sum_{i=1}^{m} \sigma_i \Phi(x_i)\rangle\right]$$

$$\le \frac{1}{m} E_{\sigma,s}\left[\sup_{y} \|w_y\|_{\mathcal{H}} \left\|\sum_{i=1}^{m} \sigma_i \Phi(x_i)\right\|_{\mathcal{H}}\right]$$

$$\le \frac{\Lambda}{m} E_{\sigma,s}\left[\left\|\sum_{i=1}^{m} \sigma_i \Phi(x_i)\right\|_{\mathcal{H}}\right] \quad \text{(Concavity)}$$

$$\le \frac{\Lambda}{m} E_{\sigma,s}\left[\left\|\sum_{i=1}^{m} \sigma_i \Phi(x_i)\right\|_{\mathcal{H}}^2\right]^{1/2}$$

$$= \frac{\Lambda}{m} E_{\sigma,s}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i \sigma_j \Phi(x_i) \Phi(x_j)\right]^{\frac{1}{2}} \quad E[|XY|] \le E[|X|^2]^{\frac{1}{2}} E[|Y|^2]^{\frac{1}{2}}$$

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

$$= \frac{\Lambda}{m} E_S\left[\sum_{i=1}^{m} k(x_i, x_i)\right]^{\frac{1}{2}}$$

$$\le \frac{\Lambda}{m} m^{\frac{1}{2}} r = \left(\frac{r^2 \Lambda^2}{m}\right)^{\frac{1}{2}}$$

$E[\sigma_i \sigma_j] = 0$ $i \ne j$

For kernel SVM multi class classifier

$$R_\rho(h) \le \hat{R}_\rho(h) + \frac{2k^2}{\rho}\left(\frac{r^2\Lambda^2}{m}\right)^{\frac{1}{2}} + \sqrt{\frac{\log 1/\delta}{2m}}$$

$\|W\|_{\mathcal{H},p} \le \Lambda$

$k(\cdot,\cdot) \le r^2$

To get a good classifier we sol

$$\min_W \hat{R}_\rho(h) + \frac{1}{2}\|W\|^2_{\mathcal{H},p}$$

If the sol is W^* then the classifier is

$$x \longmapsto \arg\max_{y \in \mathcal{Y}} \langle W^*_y, \Phi(x)\rangle$$

— This is convex opt prob

— Primal, dual prob are in fact QP.

— complexity can be large for k large.

$p \ge 1$.

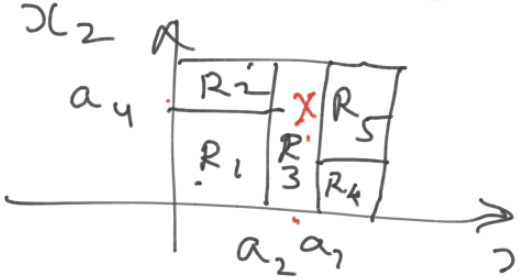
Take $p = 2$

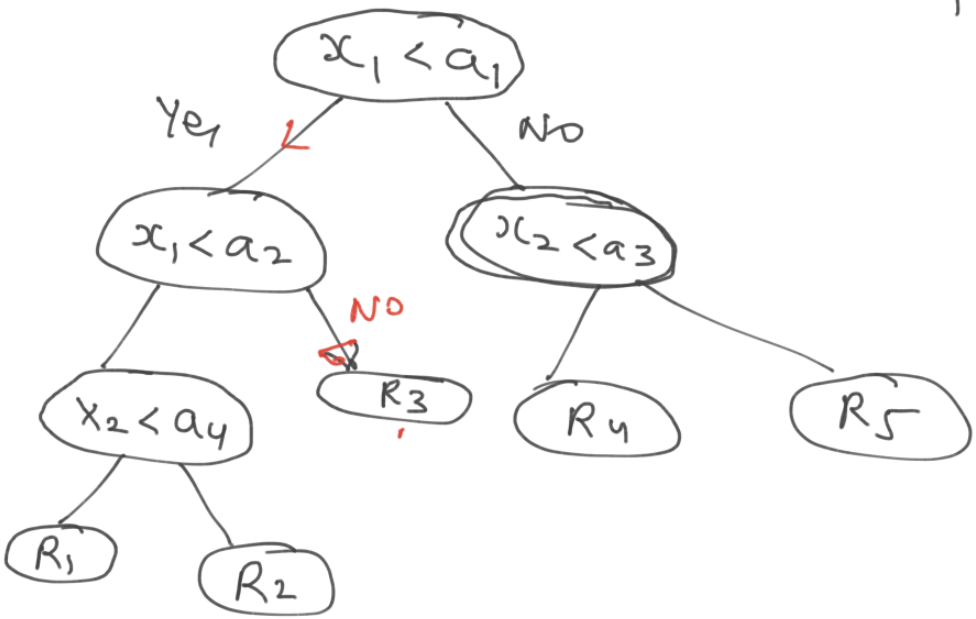
$$\hat{R}_\rho(h) = \frac{1}{m}\sum_{i=1}^{m} \xi_i$$

$$\xi_i = \max\{1 - W_{y_i}\cdot\Phi(x_i) - \max_{y' \ne y_i} W_{y'}\cdot\Phi(x_i), 0\}$$

Decision Tree classifiers

Ex





- Decision tree can be used also for regression, clustering
- Performance of decision tree is not the best — SVM, NN can give better performance
  - can boost performance via Random forest to get the best performing systems.
- Advantages of decision trees:
  - Fast to learn and evaluate
  - can treat numerical and categorical data naturally.

- Scales well with large data set
- Easy to explain.
- Decision tree need not be binary tree. But Binary is preferred.
  - Computationally easier to learn
  - For other trees the splitting of data may be too fast

Learning Decision tree:
- obtaining optimal tree is NP hard.
- VC dim of a binary tree with n nodes and N features (dim of $X = N$) is $O(n \log N)$
- ∴ VC dim of a general family of decision trees is ∞.
- n nodes, N features. binary tree: with prob $\geq 1-\delta$

$$R_D(h) \leq R_S(h) + \sqrt{\frac{(n+1)\log_2(N+3) + \log\frac{2}{\delta}}{2m}}$$

$R_D(h)$: generalization error for tree h; $R_S(h)$: training error

Greedy algorithm to learn a "good" decision tree