MA4270 Lecture 9.

1) Recap of dual SVM formulation

2) Model Selection

3) Structural Risk Management → Upper Bound (SRM)

Primal SVM with slackness & offset.

$$\min_{\underline{\theta}, \theta_0, \underline{\xi}} \quad \frac{1}{2}\|\underline{\theta}\|^2 + C\sum \xi_t$$

$$\text{s.t} \quad y_t(\langle \underline{\theta}, \underline{\phi}(x_t)\rangle + \theta_0) \geq 1 - \xi_t$$

$$\xi_t \geq 0$$

Lagrangian

↓

$$\mathcal{L}(\underline{\theta}, \theta_0, \underline{\xi}; \underline{\alpha}, \underline{\lambda}) = \frac{1}{2}\|\underline{\theta}\|^2 + C\sum \xi_t + \sum \alpha_t (1 - \xi_t - y_t(\langle \underline{\theta}, \underline{\phi}(x_t)\rangle + \theta_0)) - \sum \lambda_t \xi_t$$

KKT Condition:

① Stationarity:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \underline{\theta}} = 0 \Rightarrow \underline{\theta} = \sum \alpha_t y_t \underline{\phi}(x_t) \\ \frac{\partial \mathcal{L}}{\partial \theta_0} = 0 \Rightarrow \sum \alpha_t y_t = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_t} = 0 \Rightarrow C = \alpha_t + \lambda_t \end{cases}$$

$C = \alpha_t + \lambda_t \rightarrow \alpha_t \in [0, C]$

② feasibility

③ CS

Substitue stationarity conditions into Lagrangian:

$$\phi(\underline{\alpha}, \underline{\lambda}) = \min_{\underline{\theta}, \theta_0, \underline{\xi}} \mathcal{L}(\underline{\theta}, \theta_0, \underline{\xi}; \underline{\lambda}, \underline{\alpha})$$

$\underline{\theta}, \theta_0, \underline{\xi}$

$$= -\frac{1}{2}\sum_{t,s} \alpha_t \alpha_s y_t y_s \langle \underline{\phi}(\underline{x}_t), \underline{\phi}(\underline{x}_s) \rangle + \sum_t \alpha_t$$

→ Dual Maximization Prob:

$$\max_{\underline{\alpha}} \; -\frac{1}{2}\sum_{t,s} \alpha_t \alpha_s y_t y_s \langle \underline{\phi}(\underline{x}_t), \underline{\phi}(\underline{x}_s) \rangle + \sum \alpha_t$$

s.t : $\alpha_t \in [0, C]$

$\sum \alpha_t y_t = 0$ ← $\frac{\partial L}{\partial \theta_0} = 0$

⇓

Stationarity Condition

$$\Leftrightarrow \min_{\underline{\alpha}} \; \frac{1}{2}\sum \alpha_t \alpha_s y_t y_s \langle \underline{\phi}(\underline{x}_t), \underline{\phi}(\underline{x}_s) \rangle - \sum \alpha_t$$

s.t $\alpha_t \in [0, C]$

$\sum \alpha_t y_t = 0$

$$\Leftrightarrow \boxed{\begin{aligned} &\min_{\underline{\alpha}} \; \frac{1}{2}\underline{\alpha}^T H \underline{\alpha} - \underline{\alpha}^T \underline{1} \\ &\text{s.t} \quad \alpha_t \in [0, C] \\ &\qquad y^T \alpha = 0 \end{aligned}}$$

[Rmk]: i) (P) is a QP in at least D variables.

$$\min_{\underline{\theta}, \theta_0, \underline{\xi}} \; \frac{1}{2}\|\theta\|^2 + C\sum \xi_t$$

s.t $y_t(\langle \underline{\theta}, \underline{\phi}(\underline{x}_t) \rangle + \theta_0) \geq 1 - \xi_t$

$\xi_t \geq 0$

$\underline{\theta} \in \mathbb{R}^D$

can be very BIG.

$\phi: \mathbb{R}^d \to \mathbb{R}^D$

(D) is a QP in only n variables

→ $\alpha_1, \ldots, \alpha_n \in [0, C]$

small dataset

Dual is cheaper if $D \gg n$ (modern dataset)

ii) Dual only involves the $\langle \phi(x_t), \phi(x_s) \rangle$

$K(x_t, x_s) = \langle \phi(x_t), \phi(x_s) \rangle$.

E.g. If we choose RBF Kernel $K(x, x') = \exp(-\gamma \|x - x'\|^2)$

$\Rightarrow \gamma$: "kernelscale" in Matlab.

① if $\gamma \to 0^+$, $\lim_{\gamma \to 0^+} K(x, x') = 1$ for all $x, x' \in \mathbb{R}^d$

$$K = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$$ → 都一样

check the dual prob: $(\sum \alpha_t y_t)(\sum \alpha_s y_s) = 0$!

$$\min \ \frac{1}{2} \sum \alpha_t \alpha_s y_t y_s - \sum \alpha_t$$

$$s.t: \ \alpha_t \in [0, C], \ \boxed{\sum \alpha_t y_t = 0}$$

$\Leftrightarrow$ $\min: -\sum \alpha_t$

$s.t: \ \alpha_t \in [0, C]$

$\sum \alpha_t y_t = 0$

if: # of '+' = # of '−'

$\Rightarrow \alpha_t = C$ for every t!

→ guarantee by # of '+' = # of '−'

Every data point is SV!

Model Selection : ⇒New:

Consider 2 candidate kernel for classification/regression

⇓

$K_1(x,x') = 1+\langle x,x'\rangle$ & $K_2(x,x') = (1+\langle x,x'\rangle)^2$

K_2 is "richer" than K_1 !

⇓

For K_1, the discriminant func (or decision boundary) are functions such that :

$$f(x;\theta) = \theta^T \underline{\phi}^{(1)}(x)$$

$$= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots$$

Eg. In 1 dim $f(x;\theta) = \theta_0 + \theta_1 x_1$

For K_2, the discriminant func (or decision boundary) are functions such that :

$$f(x;\theta) = \theta^T \underline{\phi}^{(2)}(x)$$ ⇒ 一共 $\binom{d+2}{2}$ 个参数

$$= \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d + \cdots$$

$$+ \sqrt{2}\theta_{12} x_1 x_2 + \sqrt{2}\theta_{13} x_1 x_3 + \cdots + \theta_{dd} x_d^2 + \cdots$$

Informally, whatever discriminant fn K_1 can produce, K_2 can also produce.

⇓

More precisely, $\mathcal{F}_1 \subseteq \mathcal{F}_2$.

$$\mathcal{F}_i = \{ f(\cdot;\underline{\theta}) : \underline{\theta} \in \mathbb{R}^{d_i} \}$$

the disri. functions space of K_i

$\forall f_1 \in \mathcal{F}_1$, there $\exists\ f_2 \in \mathcal{F}_2$ s.t $f_1 = f_2$!

choose special $\theta_1 \in \mathbb{R}^{d_1}$ ← $\theta_2 \in \mathbb{R}^{d_2}$

Problem: Select a kernel function from a set $\{K_i : i=1,2,\dots,m\}$ where models are associated to $f^{\sim}$ classes $\mathcal{F}_i$.

$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_m$.

↘ Nobody tells us use which F_i

$\mathcal{D}_n = \{(x_t, y_t)\}_{t=1}^{n}$

Given that we have to use functions in $\mathcal{F}_i$, find parameters $\underline{\theta}$ corresponding to $\hat{f}_i \in \mathcal{F}_i$, s.t

↓

$$J(\theta) = \sum_t \text{loss}(y_t, f(x_t; \underline{\theta})) + \lambda_n \|\theta\|^2$$

Note: this loss should be corresponding to the certain classifier

is minimized over over $\underline{\theta}$ s.t $f(\cdot; \underline{\theta}) \in \mathcal{F}_i$

$\hat{f}_i(x) = f(x; \hat{\theta}_i) \rightsquigarrow \hat{f}_i \in \mathcal{F}_i$.

$\mathcal{F}_1, \dots, \mathcal{F}_n$

$\hat{f}_1, \dots, \hat{f}_n$

$\hat{\theta}_1, \dots, \hat{\theta}_n$

$\hat{\theta}_i$ → r.v. → depend on $\mathcal{D}_n$

Each parameter of $\underline{\theta}$ has an associated RISK

Expected Risk

RISK : $R(\theta) = \mathbb{E}_{(x,y)\sim P}[\text{Loss}_{0-1}(y, f(x;\theta))]$

$\hat{y}$

target: y; our prediction over f, θ

(x,y) has joint dist. $P(x,y) = P$.

Unknown Prob dist

$$\text{Loss}_{0-1}(y, \hat{y}) = \begin{cases} 1, & y \neq \hat{y} \\ 0, & y = \hat{y} \end{cases}$$

→ if we know P ⇒ analytically solve!

⇓

But actually we don't know!

→ What we can do?

$$J(\theta) = \sum \text{Loss}(y_t, f(x_t, \theta)) + \lambda_n \|\theta\|^2$$

Answer: $\hat{\theta}_i = \underset{\theta \in \mathbb{R}^{d_i}}{\text{argmin}} \{ J(\theta) : f(\cdot;\theta) \in \mathcal{F}_i \}$.

$R(\hat{\theta}) = R(\hat{f}_i)$: still a random variable

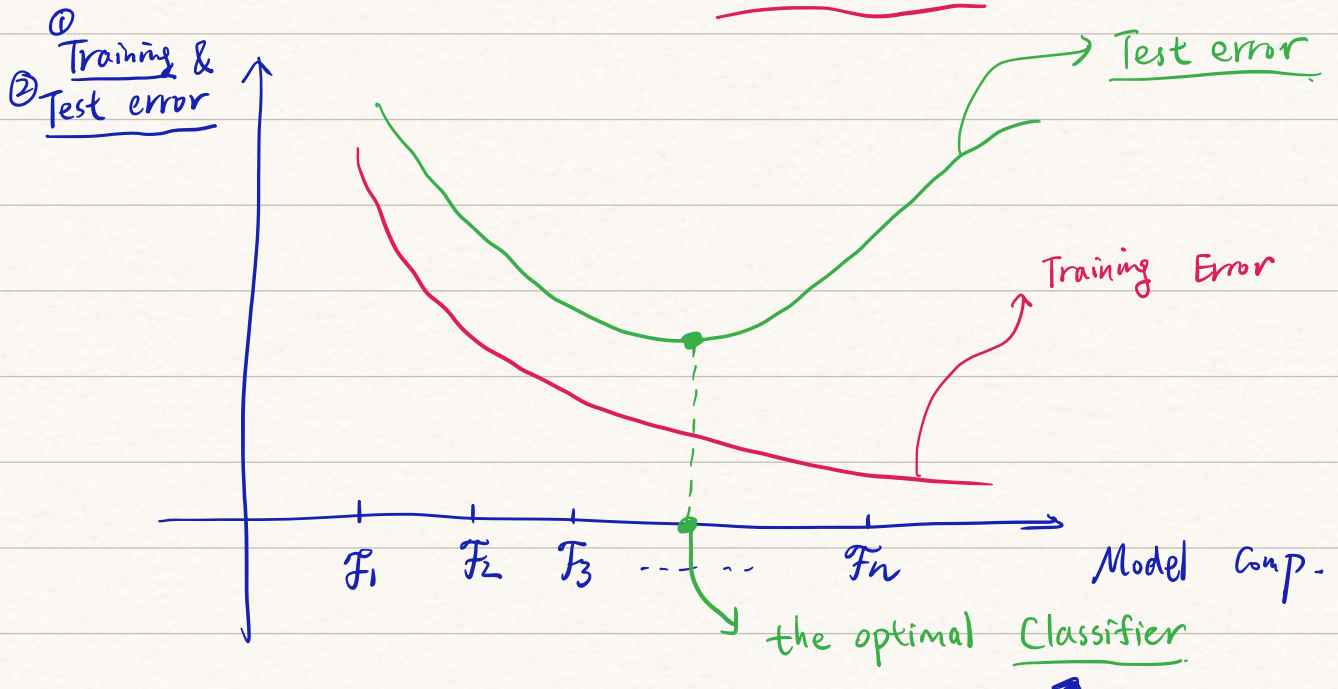
risk associated to the function learned from

⇓

$(x_t, y_t) \sim P$ (i.i.d)

⇓

$\hat{\theta}$ is random



Q: How to select/learn the "sweet spot"?

↓

correct model order

A: Relate the expected risk $R(\hat{f}_i)$ to the emprical

risk $\hat{R}_n(\hat{f}_i)$

$$R(\hat{f}_i) = \mathbb{E}_{(x,y)\sim P}\left[\text{Loss}_{01}(y, \hat{f}_i(x)) \right]$$

expected risk (unable to calculate)

Unknown

Empirical Risk → obtained from data

$$\hat{R}_n(\hat{f}_i) = \frac{1}{n} \sum_{t=1}^{n} \text{Loss}_{01}(y_t, \hat{f}_i(x_t))$$

Suppose we can approximate $R(\hat{f}_i)$ → (unknown expected risk) with the empirical risk (known), then we can minimize $\hat{R}_n(\hat{f}_i)$ instead!

⇓

Goal: Show that with probability $\geq 1-\delta$ (small!),

$$R(\hat{f}_i) \leq \hat{R}_n(\hat{f}_i) + \epsilon(n, \mathcal{F}_i, \delta)$$

Exp. Risk Emp. Risk.

A uniform Bound → for every $f \in \mathcal{F}$

Use this Upper Bound to select Models

Now perform model selection: → strategy.

$$i^* = \underset{i}{\text{argmin}} \left\{ \hat{R}_n(\hat{f}_i) + \epsilon(n, \mathcal{F}_i, \delta) \right\}$$

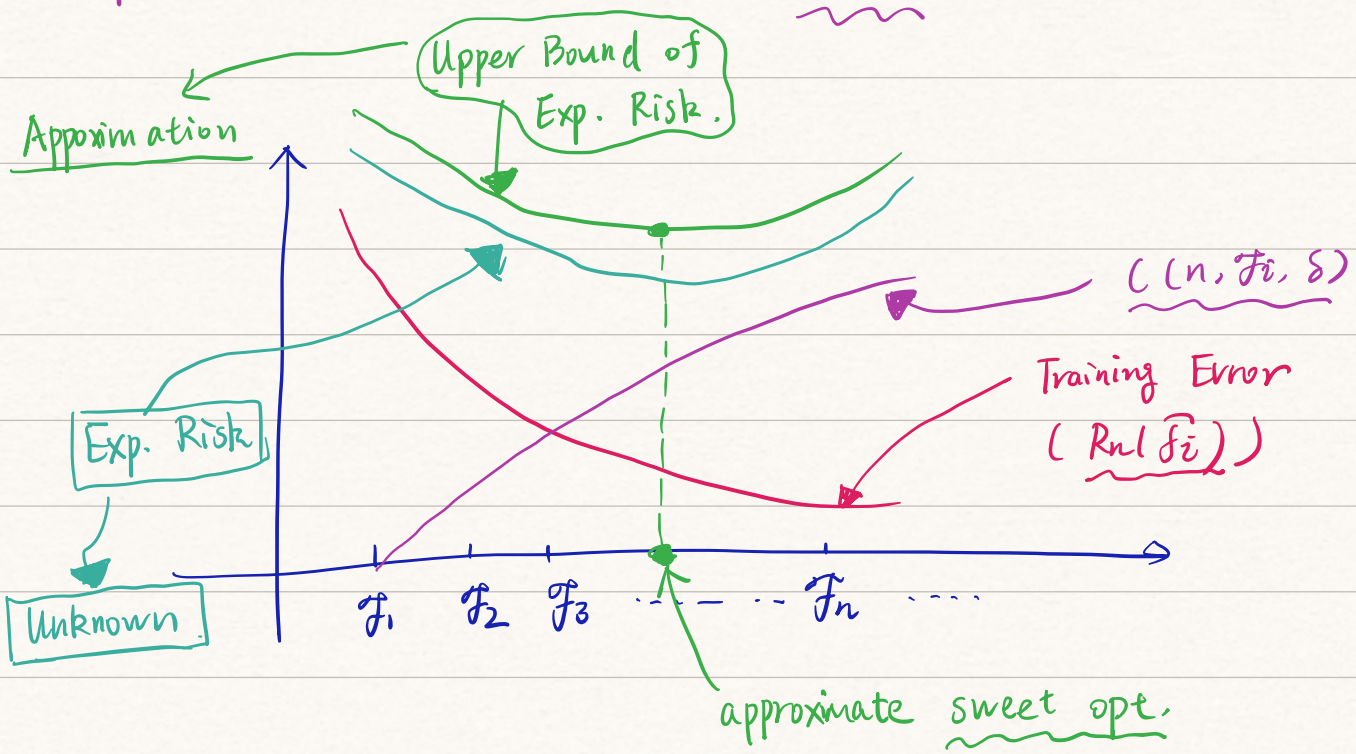
↓ Model Complexity

⇓

This Principle → Structural Risk Minimization (SRM)

排连续复杂的模型

cf: Minimizing $\hat{R}_n(\hat{f}_i)$ over i is known as Empirical Risk Minimization (ERM).



Will derive a result for the case in which $|\mathcal{F}_i| < \infty$

$i = 1, 2, \ldots, m$

use VC dimension to extend this to $|\mathcal{F}_i| = \infty$.

$$\mathcal{F}_i = \{ f : \mathbb{R}^d \to \{\pm 1\} : f(x; \theta) = \text{sgn}(\langle \theta, x \rangle), \theta \in \mathbb{R}^d \}$$

→ Not a valid func. class because $|\mathcal{F}_i| = \infty$

Looking for a bound s.t.

$$\forall f \in \mathcal{F}_i, \quad \underbrace{R(f)}_{\text{Exp. Risk}} \leq \underbrace{R_n(f)}_{\text{Emp. Risk}} + C(n, \mathcal{F}_i, \delta) \text{ happens with prob. } \geq 1-\delta$$

[Lemma 1: $A_1, \ldots, A_k$, $\Pr(\bigcup A_i) \leq \sum \Pr(A_i)$

Two Facts!

Lemma 2: $X_1, \dots, X_n$ i.i.d $EX_1 = \mu$. $X_i \in [0,1]$ a.s.

$$\Pr\left(\left| \frac{1}{n} \sum X_i - \mu \right| > \varepsilon \right) \leq 2 e^{-n\varepsilon^2}$$

Consider: $\Pr\left(\max_{f \in \mathcal{F}_i} |R(f) - R_n(f)| > \varepsilon \right) := p$.

this is because $|\mathcal{F}_i| < \infty$

This is the prob. that at least one classifier in $\mathcal{F}_i$ has expected risk derivating from emp. risk by $> \varepsilon$

$$p = \Pr\left(\bigcup_{f \in \mathcal{F}_i} \{ |R(f) - R_n(f)| > \varepsilon \} \right)$$

$$\leq \sum_{f \in \mathcal{F}_i} \Pr\left(|R(f) - R_n(f)| > \varepsilon \right) \leftarrow \text{Lemma 1}.$$

($R(f)$: Exp. Risk; $R_n(f)$: Emp. Risk.)

Associate to each $(X_t, y_t) \sim P$ a r.v.

$$S_t := \mathbb{1}\{ y_t f(X_t) \leq 0 \} \quad t = 1, \dots, n$$

Note:

① $S_t = 1 \Leftrightarrow f$ makes an error on (X_t, y_t)

② $S_t = 0 \Leftrightarrow f$ makes no error on (X_t, y_t)

$$\Rightarrow R_n(f) = \frac{1}{n} \sum_t \text{Loss}_{0-1}(y_t, f(X_t)) = \frac{1}{n} \sum_t S_t$$

$$R(f) = \mathbb{E}[S_t] = \Pr(y_t f(X_t) \leq 0)$$

$$= \mathbb{E}_{(x,y)\sim P}\left[\text{Loss}_{0\text{-}1}(y, f(x))\right]$$

$$\Pr\left(\left|R(f) - R_n(f)\right| > \varepsilon\right)$$

$$= \Pr\left(\left|\mathbb{E}(S_t) - \frac{\sum S_t}{n}\right| > \varepsilon\right) \qquad S_t \in [0,1]$$

$$\leq 2e^{-n\varepsilon^2} \qquad \leftarrow \text{Lemma 2}$$

All in all, by Lemma 1 & 2 & $|\mathcal{F}_i| < \infty$

$$\Pr\left(\max_{f\in\mathcal{F}_i} \left|R(f) - R_n(f)\right| > \varepsilon\right) \leq |\mathcal{F}_i|\, 2e^{-n\varepsilon^2} = \delta$$

choose $\Rightarrow \varepsilon := \sqrt{\frac{\log|\mathcal{F}_i| + \log\left(\frac{2}{\delta}\right)}{2n}}$ $\Rightarrow$ with prob δ

$\Leftrightarrow \max_{f\in\mathcal{F}_i} \left|R(f) - R_n(f)\right| \leq \varepsilon = \sqrt{\frac{\log|\mathcal{F}_i| + \log\left(\frac{2}{\delta}\right)}{2n}}$ with prob $1-\delta$

Rewrite:

$$\left|R(f) - R_n(f)\right| \leq \sqrt{\frac{\log|\mathcal{F}_i| + \log\left(\frac{2}{\delta}\right)}{2n}} \qquad \forall f \in \mathcal{F}_i$$

$$\Rightarrow R(f) \leq R_n(f) + \sqrt{\frac{\log|\mathcal{F}_i| + \log\left(\frac{2}{\delta}\right)}{2n}} \qquad \forall f \in \mathcal{F}_i$$

$$C(n, \mathcal{F}_i, \delta)$$

goes down as $i \to \infty$; goes up as $i \to \infty$

#

Q: → $R(f)$ → fixed number
→ $R(\tilde{f}_i)$ → r.v.
→ $\hat{R}_n(f)$ → r.v. ?

① $R(f)$ → exp. risk.

$\mathbb{E}[\text{Loss}_{01}(y, \hat{f}(x))]$ → if $\tilde{f}$ is attained by: minimize loss + penalty

sample dataset → random → $\tilde{f}$ (r.v.)

Can we understand as: we do the things in the probability space? ⇒ r.v

② $\mathcal{D}_n = \{(x_t, y_t)\}_{t=1}^n$ (r.v, r.v) — theoretical

$\xrightarrow{\text{real world}}$ $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^n$ (observation)

Emp. risk.

$\hat{R}_n(f) = \frac{1}{n}\sum \text{Loss}_{01}(y_t, f(x_t))$

⇓

can be calculated.

$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \ \frac{1}{n}\sum \text{loss}(y_t, f(x_t)) + \lambda \|\theta\|^2$

Conclusion:

↳ Given δ, $R(f) \le \hat{R}_n(f) + C(n, \mathcal{F}_i, \delta)$ — Compensate for $\mathbb{E}$

$R(\hat{\delta}_i) \Rightarrow$ ① $\mathbb{E}$ X

② $\hat{\delta}_i$ X

$\rightarrow \hat{R}_n(\hat{\delta}_i) \Rightarrow \hat{\delta}_i$ X

Compensate for $\hat{\delta}_i$

$\hat{R}_n(\hat{\delta}_i)$

$\Downarrow \hat{\delta}_i = \text{argmin}\{\quad\}$