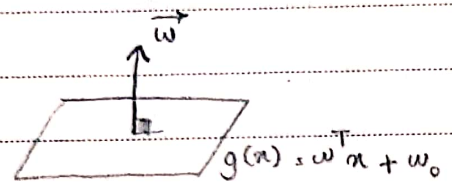
امیرمهدی انصاری پور 810198358

1) 1. اگر مرز تقسیم LDF باشد، آن را میتوان بصورت $g(x) = w^T x + w_0$ نوشت که $g(x)$ یک صفحه می باشد. جهت صفحه با بردار $\vec{w}$ (عمود بر $\vec{x}$) بدست می آید و موقعیت آن هم با w_0 تعیین می شود.



1) 2. چندین روش برای تحلیل نویز در SVM وجود دارد:

1) استفاده از Kernel: در صورت وجود رابطه غیرخطی میان نمونه‌ها، میتوان از یک $\phi(.)$ استفاده کرد که داده‌ها را به feature space دیگری می‌برد. با بکار بردن Kernel مناسب میتوان داده‌های نویزی را حذف کرد.

2) استفاده از term خطا: اگر رابطه loss function را بصورت زیر اصلاح کنیم:

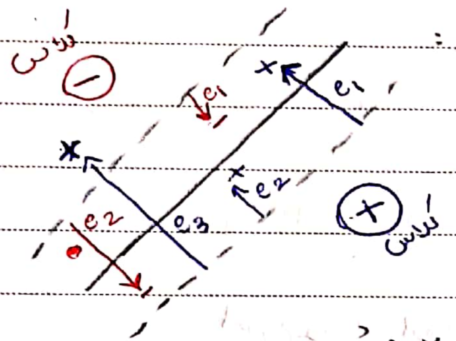
$$w^*, b^* = \underset{w,b}{\arg\min} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} e_i$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 \ \&\& \ e_i \geq 0 \quad (1 \leq i \leq n)$$

آنچه SVM سعی میکند همزمان:

1) $\|w\|^2$ را کاهش دهد، (که $\text{margin} = \frac{2}{\|w\|}$ افزایش یابد)

2) مجموع error ها را کاهش دهد، ارور برای هر sample بصورت فاصله آن sample تا margin کلاس تعریف میشود. به طور مثال:



(دقت شود در رابطه بهینه‌سازی صفحه قبلی هرچقدر C کوچکتر باشد، مدل تحمل misclassification بالاتری پیدا میکند، اصطلاحاً Soft SVM میشود.)

3) K-fold: با استفاده از روش K-fold که در آن دیتاست به K قسمت تقسیم میشود، توانایی مدل در شناسایی نویز بهتر میشود. همچنین hyperparameter های مدل نیز بخوبی ست میشوند که نویز را لحاظ کنند.

1) 3. چنانچه بخواهیم از SVM برای جداسازی داده‌هایی استفاده کنیم که بصورت خطی جداپذیر نیستند، باید با استفاده از یک تبدیل داده‌ها را به feature space دیگری ببریم. به آن تبدیل که با نماد $\Phi(\vec{x})$ نشان میدهیم kernel گوییم.

با توجه به رابطه‌های زیر:

$$margin = \frac{2}{\|w\|} \Rightarrow \arg\max\, margin = \arg\min_{w,b} \|w\|$$

$$, \quad \arg\min_{w,b} \|w\|^2 = \arg\max_{\lambda}\, g(\lambda)$$

$$, \quad g(\lambda) = \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \Phi^T(x_i)\, \Phi(x_j)$$

میان دو کرنل Φ_1 و Φ_2 ، حاصل $\Phi_1^T(x_i)\, \Phi_1(x_j) < \Phi_2^T(x_i)\, \Phi_2(x_j)$

در نتیجه به ازای Φ_1 ، $g(\lambda)$ مقدار بزرگتری می‌گیرد که معادل است با اینکه

margin مقدار بزرگتری بگیرد. پس به ازای Φ_1 ، margin بزرگتری نتیجه میشود.

2) الف:

$$\frac{\|x-y\|_2^2}{\sigma^2} = \frac{(x-y)^T(x-y)}{\sigma^2} = \frac{x^Tx}{\sigma^2} - \frac{2x^Ty}{\sigma^2} + \frac{y^Ty}{\sigma^2}$$

$$\Rightarrow K(x,y) = \exp\left(-\frac{x^Tx}{\sigma^2} + \frac{2x^Ty}{\sigma^2} - \frac{y^Ty}{\sigma^2}\right)$$

$$= \underbrace{\exp\left(-\frac{x^Tx}{\sigma^2}\right)}_{f(x)} \times \exp\left(\frac{2x^Ty}{\sigma^2}\right) \times \underbrace{\exp\left(-\frac{y^Ty}{\sigma^2}\right)}_{f(y)}$$

چنانچه ثابت کنیم $\exp\left(\frac{2x^Ty}{\sigma^2}\right)$ یک valid kernel است، طبق رابطه

$K_2(x,y) = f(x)\, k_1(x,y)\, f(y)$ (f هر تابع حقیقی) ، میتوان نتیجه گرفت کرنل

صورت مسئله نیز valid است.

$$\frac{2x^Ty}{\sigma^2} = \frac{\sqrt{2}}{|\sigma|} x^T \cdot \frac{\sqrt{2}}{|\sigma|} y = \phi^T(x) \cdot \phi(y)$$

if $k_1(x,y)$ is valid, then $k_2(x,y) = \exp(k_1(x,y))$ is also valid $\Rightarrow$

$$K(x,y) = \exp\left(\frac{2x^Ty}{\sigma^2}\right) \text{ is valid}$$

$$\Rightarrow K(x,y) = f(x) \exp\left(\frac{2x^Ty}{\sigma^2}\right) f(y) \text{ is valid}$$

$$\Rightarrow K(x,y) = \exp\left(\frac{-x^Tx}{\sigma^2}\right) \exp\left(\frac{2x^Ty}{\sigma^2}\right) \exp\left(\frac{-y^Ty}{\sigma^2}\right) \text{ is valid}$$

$$\Rightarrow K(x,y) = \exp\left(-\frac{1}{\sigma^2}\|x-y\|^2\right) \text{ is valid. } \checkmark$$

2 ب:

$$\left.\begin{aligned} k_1(x,y) &= \phi_1^T(x)\,\phi_1(y) \\ K_2(x,y) &= \phi_2^T(x)\,\phi_2(y) \end{aligned}\right\} \Rightarrow b\,k_1(x,y) + a\,K_2(x,y)$$

$$= \sqrt{b}\,\phi_1^T(x)\,\sqrt{b}\,\phi_1(y) + \sqrt{a}\,\phi_2^T(x)\,\sqrt{a}\,\phi_2(y)$$

$$= \begin{bmatrix} \sqrt{b}\,\phi_1(x) \\ \sqrt{a}\,\phi_2(x) \end{bmatrix}^T \cdot \begin{bmatrix} \sqrt{b}\,\phi_1(y) \\ \sqrt{a}\,\phi_2(y) \end{bmatrix} = \phi_{new}^T(x)\,\phi_{new}(y)$$

همانطور که دیده میشود، بشرطیکه a و b هر دو مثبت باشند، آنگاه "$K(x,y) = a\,k_1(x,y) + b\,K_2(x,y)$"

زیر رادیکال باید + باشند

valid است.

Subject:
Year. Month. Date. ( )

ج) 2 :

$K(x,y) = f(x)\, f(y)$ (f هر تابع حقیقی)

کرنل فوق valid است اگر :

$K(x,y) = f(x)\, \phi^T(x)\, \phi(y)\, f(y)$

اگر تعریف کنیم : $\phi(x) = I$: $\phi^T(x)\, \phi(y) = I^T I = I$ (symmetric)

$\Rightarrow K(x,y) = f(x)\, I\, f(y) = f(x)\, f(y)$ is valid ✓

د) 2 :

$K(x,y) = K_1(g(x), g(y))$

$K_1(x,y) = \phi^T(x)\, \phi(y) \Rightarrow K_1(g(x), g(y)) = \phi^T(g(x))\, \phi(g(y))$

$= \phi_{new}(x)^T\, \phi_{new}(y)$ is valid

ه) 2 :

$K_1(x,y) = \phi_1^T(x)\, \phi_1(y) \quad \phi_1(z) = [\phi_{10}(z), \phi_{11}(z), \dots, \phi_{1n}(z)]$

$K_2(x,y) = \phi_2^T(x)\, \phi_2(y) \quad \phi_2(z) = [\phi_{20}(z), \dots, \phi_{2m}(z)]$

$\Rightarrow K_1(x,y) = \sum_{i=0}^{m} \phi_{1i}(x)\, \phi_{1i}(y)$

$\Rightarrow K_2(x,y) = \sum_{j=0}^{n} \phi_{2j}(x)\, \phi_{2j}(y)$

$$\Rightarrow K_1(x,y) \cdot K_2(x,y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \overbrace{\phi_{1i}(x)\, \phi_{2j}(x)}^{c_{ij}(x)}\ \underbrace{\phi_{1i}(y)\, \phi_{2j}(y)}_{c_{ij}(y)}$$

$\sum_{i=0}^{m} \sum_{j=0}^{n} c_{ij}(x)\, c_{ij}(y) = C^T(x)\, C(y)$ is valid ✓

کرنل ما یک ماتریس $m \times n$ است که : $C_{ij}(z) = \phi_{1i}(z)\, \phi_{2j}(z)$

PAPCO

$$\text{loss fution} = \frac{1}{2}\|w\|^2 + \frac{c}{2}\sum_{i=1}^{N} \varepsilon_i^2 \quad s.t. \quad y_i(w^T x_i + b) \geq 1 - \varepsilon_i \qquad (3$$

اگر تعریف کنیم:

$$\vec{w}_{new} = [\vec{w}^T, \sqrt{c}\ \vec{\varepsilon_i}^T]^T$$

$$\vec{x}_{new\,i} = [\vec{x_i}, \frac{1}{\sqrt{c}}\ \vec{e_i}]^T \qquad b_{new} = b$$

در فرمول بالا، $\vec{e_i}$ یک بردار $[\ ]_{n\times 1}$ است که:

$$\vec{e_i} = [\ ]_{n\times 1} \ \&\& \ \forall j \neq i : e_{ij} = 0 \ \&\& \ e_{ii} = 1$$

(برای مثال، $e_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}$)

$$\frac{1}{2}\|w_{new}\|^2$$

حال داریم:

$$\boxed{\frac{1}{2} w_{new}^T w_{new}} = \frac{1}{2}\left[\|w\|^2 + c\|\varepsilon\|^2\right] =$$

$$\frac{1}{2}\|w\|^2 + \frac{c}{2}\sum_{i=1}^{N} \varepsilon_i^2$$

s.t.

از طرفی داریم:

$$\boxed{y_i(w_{new}^T x_{new\,i} + b_{new})} = y_i(w^T x_i + \underbrace{\varepsilon_i^T e_i}_{\varepsilon_i} + b)$$

Subject:
Year.　Month.　Date.　( )

حل الف:

$$y_i \left(w_{new}^{T} x_{new\,i} + b_{new} \right) \geqslant 1$$

$$\Longrightarrow \quad y_i \left(w^T x_i + b \right) \geqslant 1 - y_i \varepsilon_i$$

$$\Rightarrow \text{ if } y_i = 1 : \quad y_i \left(w^T x_i + b \right) \geqslant 1 - \varepsilon_i \quad \checkmark$$

$$\Rightarrow \text{ if } y_i = -1 : \quad y_i \left(w^T x_i + b \right) \geqslant 1 + \varepsilon_i \geqslant 1 - \varepsilon_i \quad \checkmark$$

پس با تعریف $\vec{w}_{new}$ و $\vec{x}_{new\,i}$ و b_{new} توانستیم تابع هزینه جدید را بصورت زیر بدست آوریم:

$$\text{loss - function} = \frac{1}{2} \left\| w_{new} \right\|^2$$

$$\text{s.t.} \quad y_i \left(w_{new}^{T} x_{i\,new} + b_{new} \right) \geqslant 1$$

Subject:

Year. Month. Date. ( )

.1 (5

$$\|\Phi(x_i)-\Phi(x_j)\|_2^2 = (\Phi(x_i)-\Phi(x_j))^T(\Phi(x_i)-\Phi(x_j))$$

$$= \Phi(x_i)\cdot\Phi(x_i) + \Phi(x_j)\cdot\Phi(x_j) - 2\,\Phi(x_i)\cdot\Phi(x_j)$$

$$= K\langle x_i, x_i\rangle + K\langle x_j, x_j\rangle - 2K\langle x_i, x_j\rangle =$$

$$\exp\left(-\frac{1}{2}\|\overset{0}{\cancel{x_i - x_i}}\|^2\right) + \exp\left(-\frac{1}{2}\|\overset{0}{\cancel{x_j - x_j}}\|^2\right)$$

$$-2\exp\left(-\frac{1}{2}\|x_i - x_j\|^2\right) = 1 + 1 - 2\exp\left(-\frac{1}{2}\|x_i - x_j\|^2\right) < 2$$

5) 2.

$$K(x,y) = (x^T y)^2 + 1 + 2x^T y$$

$$= \sum_i x_i y_i \sum_j x_j y_j + \sum_i \sqrt{2}\, x_i \sqrt{2}\, y_i + 1$$

$$= \sum_i \sum_j (x_i x_j)(y_i y_j) + \sum_i (\sqrt{2}\, x_i)(\sqrt{2}\, y_i) + 1$$

همانطور که دیده میشود، توانستیم x_i و y_i ها را طوری جدا کنیم که به فرم $\phi^T(x)\phi(y)$ برسیم. داریم:

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\phi(\vec{x}) = [x_1x_1, x_1x_2, \dots, x_1x_n,$$
$$x_2x_1, x_2x_2, \dots, x_2x_n,$$
$$x_3x_1, \dots, x_3x_3, \dots, x_3x_n,$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$x_nx_1, x_nx_2, \dots, x_nx_n,$$
$$\sqrt{2}x_1, \sqrt{2}x_2, \dots, \sqrt{2}x_n, 1]$$

همانطور که دیده میشود، برخی از term های بالا تکراری اند. مثلاً x_1x_2 و x_2x_1. پس باید تعدادشان را حساب کنیم:

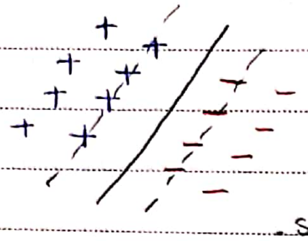
$$\underbrace{\frac{n^2-n}{2}}_{\text{تکراری ها}} + \underbrace{n}_{\substack{(x_i)^2 = \\ x_1x_1, \dots, x_nx_n}} + \underbrace{n}_{\sqrt{2}x_i} + \underbrace{1}_{1}$$

P4PCO

$$= \frac{n^2-n}{2} + 2n+1 = \frac{n^2+3n+2}{2} = \frac{1}{2}(n+2)(n+1)$$

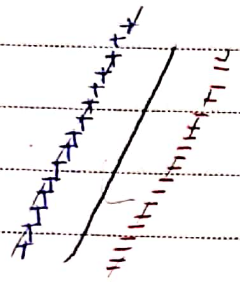
5) 3. حالت 2 support vector نیاز است، یکی از کلاس + و دیگری از کلاس -

که margin ما فاصله این دو sample میشود و خط جداکننده دقیقا



وسط این دو sample قرار می‌گیرد.) اما بیشتر از 2 support vector

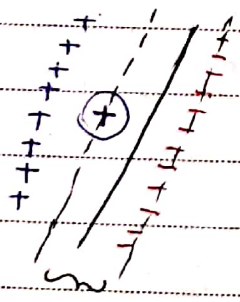
نیز قابل قبول است. برای مثال در شکل زیر $size(SV) = 8$



(در حالت حد بالایی برای SV وجود ندارد و حتی کل dataset

می‌تواند SV باشد $2 \leq SV \leq n$). مانند شکل زیر:

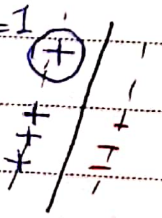
در صورتیکه یک داده جدید افزوده شود (مثلا با برچسب +)، 4 حالت وجود دارد:



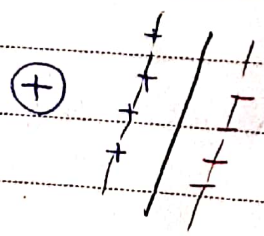
1) داده داخل margin بیفتد: در آنصورت SV کاهش می‌یابد (اما همچنان بزرگتر مساوی 2 است.)

new margin < old margin

SV+=1



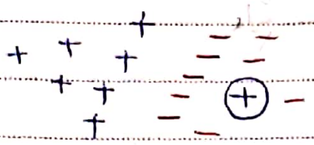
2) داده روی margin بیفتد: در آنصورت $SV += 1$ میشود (یک sample به آن افزوده میشود.)



3) داده خارج margin بیفتد و در ناحیه کلاس + قرار بگیرد:

در آنصورت SV تغییری نمی‌کند.

4) داده خارج margin بیفتد و در ناحیه



کلاس − قرار بگیرد: در آن صورت داده ها

بصورت خطی جداپذیر نیستند، و از آنجا که hard SVM هیچ گونه misclassification ی را

تحمل نمی کند، دیگر linear hard SVM قادر به کلاس بندی نمی باشد.

hard

6) با توجه به تعریف توزیع دوجمله ای ($Bin(n,p)$):

در n آزمایش برنولی مستقل (با احتمال موفقیت p) $P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$

احتمال آنکه x بار موفق شویم (متغیر تصادفی X برابر تعداد موفقیت هاست):

حال برای هرکدام از کلاس بندها میتوان گفت:

p_i = احتمال موفقیت کلاس بند iام = احتمال اینکه کلاس بند iام بتواند نمونه n را درست classify کند = 51%

احتمال اینکه از بین N کلاس بند: $\Rightarrow P(X=x)=\binom{N}{x}\left(\frac{51}{100}\right)^{x}\left(\frac{49}{100}\right)^{N-x}$

x تا شان درست پیش بینی کنند

$$\Rightarrow P\left(X\geq \frac{N+1}{2}\right)=\sum_{x=\frac{N+1}{2}}^{N}\binom{N}{x}\left(\frac{51}{100}\right)^{x}\left(\frac{49}{100}\right)^{N-x}$$

برای بدست آمدن accuracy مدل ensemble که برابر است با تعداد پیش بینی های درست به کل

پیش بینی ها، باید لااقل $\frac{N+1}{2}$ کلاس بند به درستی پیش بینی کنند.

Subject:
Year. Month. Date. ( )

1) $N=5 \Rightarrow \binom{5}{3}\left(\frac{51}{100}\right)^3\left(\frac{49}{100}\right)^2+\binom{5}{4}\left(\frac{51}{100}\right)^4\left(\frac{49}{100}\right)+\binom{5}{5}\left(\frac{51}{100}\right)^5 = 0.318+0.166+0.172=0.656$

2) $N=9 \Rightarrow \binom{9}{5}\left(\frac{51}{100}\right)^5\left(\frac{49}{100}\right)^4+\binom{9}{6}\left(\frac{51}{100}\right)^6\left(\frac{49}{100}\right)^3+\binom{9}{7}\left(\frac{51}{100}\right)^7\left(\frac{49}{100}\right)^2+\binom{9}{8}\left(\frac{51}{100}\right)^8\left(\frac{49}{100}\right)^1+\binom{9}{9}\left(\frac{51}{100}\right)^9 = 0.524$

3) ٪100 ، در واقع اگر خطای طبقه‌بندها از هم مستقل باشند آنگاه می‌توان N طبقه‌بند تعریف کرد $(N<\infty)$ که به‌ترتیب طبقه‌بند c_i، خطای طبقه‌بند c_{i-1} را پوشش دهد، پس با یک مقدار N بزرگ بالاخره به دقت ٪100 می‌رسیم. اما در واقعیت اینطور نیست و طبقه‌بندها به‌صورت مشابه عمل می‌کنند، پس خطایشان از هم مستقل نیست.

4) $N=5,\ p=0.5 \Rightarrow \underset{10}{\binom{5}{3}}(0.5)^5+\underset{5}{\binom{5}{4}}(0.5)^5+\underset{1}{\binom{5}{5}}(0.5)^5 = 0.5$

همانطور که دیده می‌شود، دقت هر کدام از طبقه‌بندها ٪50 است، یعنی random classifier اند (weak classifier باید دقتش کمی از 0.5 بیشتر باشد)، تجمیع شده این مدل‌ها نیز یک random classifier است. (افزایش N هیچ کمکی نمی‌کند.)