فهرست مطالب

۱ مقدمه

در این گزارش به بررسی سوالات تمرین پنجم درس یادگیری ماشین می‌پردازیم.

۲ سوال ۱

۱.۲ خطوط جداساز

ابتدا تابع لاگرانژ را می‌سازیم:

$$\mathcal{L} = \frac{w_1^2 + w_2^2}{2} + \lambda_1(1 - (w_1 + b)) + \lambda_2(1 - (w_2 + b)) + \lambda_3(1 + (-w_1 + b)) \tag{۱}$$

حال شرایط KKT را بررسی می‌نماییم:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_1} = 0 \\ \frac{\partial \mathcal{L}}{\partial w_2} = 0 \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \\ \lambda_j(1 - y_j(w^T x_j)) = 0; \qquad \lambda_j \geq 0,\ (1 - y_j(w^T x_j)) \leq 0 \end{cases} \implies \begin{cases} w_1 - \lambda_1 - \lambda_3 = 0 \\ w_2 - \lambda_2 = 0 \\ -\lambda_1 - \lambda_2 + \lambda_3 = 0 \\ \lambda_1(1 - (w_1 + b)) = 0 \\ \lambda_2(1 - (w_2 + b)) = 0 \\ \lambda_3(1 - (-w_1 + b)) = 0 \end{cases} \tag{۲}$$

مشاهده می‌نماییم که در نظر گرفتن $b = 0, w_1 = 1, w_2 = 1$ موجب برآورده شدن سه معادله آخر دستگاه معادلات ۲ می‌گردد. حال برای برآورده شدن سه معادله اول بدست می‌آید:

$$\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 1$$

بنابراین همانطور که مشاهده می‌گردد، جوابهای محاسبه شده تماما شروط KKT را برآورده می‌نمایند و با توجه به محدب بودن حتما این نقطه بهینه سراسری است. بنابراین می‌توان گفت خطوط جدا ساز به صورت زیر می‌باشند:

$$\rightarrow \boxed{\begin{matrix} l_1 : x_2 + x_1 = 1 \\ l_2 : x_2 + x_1 = -1 \end{matrix}} \tag{۳}$$

۲.۲ SV

دیتاهایی که دارای ضرایب ناصفر هستند SV ها را تشکیل می‌دهند. در واقع می‌توان گفت:

$$\begin{aligned} &SV = \{X_2, X_3\} = \{[0,1]^T, [-1,0]^T\} \\ &\rightarrow W = \lambda_2 y_2 X_2 + \lambda_3 y_3 X_3 = [0,1]^T + (-1) \times [-1,0]^T = [1,1] \end{aligned} \tag{۴}$$

۳ سوال ۲

۱.۳ ۱

برای مقایسه‌ی مساله‌ی دوگان [1] و مساله‌ی اولیه [2] می‌توان موارد زیر را در نظر داشت:

۱. باید توجه داشت که مساله‌ی دوگان مستقل از بعد فضای ورودی است. در واقع همانطور که در روابط ذکر شده در ارائه‌ی درس می‌توان مشاهده نمود، در فرموله کردن مسئله به روش دوگان تنها ضرب داخلی بین بردارهای ویژگی مطرح می‌گردد که حاصل یک اسکالر است. بنابراین عملا مسئله به فرمی مستقل از ابعاد تبدیل می‌گردد و از نظر محاسبه این موضوع بهینه خواهد بود.

۲. همانطور که در ارائه‌ی درس نیز مطرح شد، تنها ضریب α_i بردارهایی که بر روی صفحات جدا کننده قرار دارند، غیر صفر می‌باشند و بقیه ضرایب صفر خواهند بود. به این بردارها با ضرایب غیر صفر که نقش تعیین کننده در مسئله طبقه بندی دارند بردار پشتیبان [3] گفته می‌شود. استفاده از دوگان این فرصت را فراهم می‌نماید که در مسائل کلان، تنها از تعداد مجدودی از بردارها برای طبقه بندی استفاده شود که این امر موجب افزایش سرعت، کاهش هزینه حافظه و محاسبه می‌گردد.

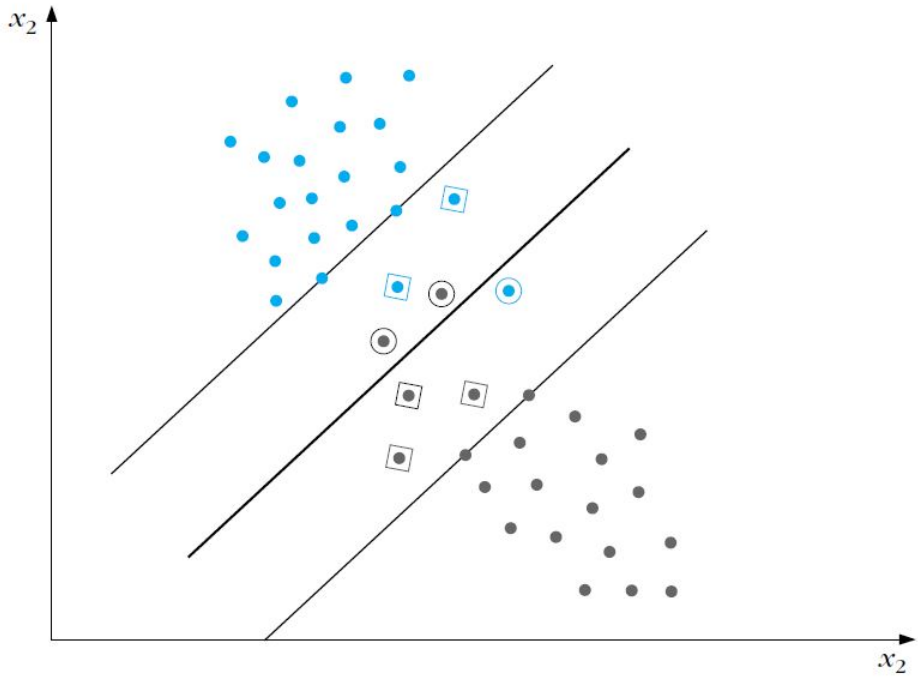
۳. استفاده از این فرم مسئله می‌تواند شرایط استفاده از توابع کرنل را فراهم کند که تکنیک موثر برای حل مسائل غیرخطی می‌باشد.

۴. استفاده از مسئله دوگان، از نظر مجاسبات عددی بسیار پر اهمیت است. در واقع این مسئله QP را می‌توان با استفاده از روش‌های مناسبی حل کرد، برای مثال استفاده از روش Coordinate Descent می‌تواند با حجم مناسبی از محاسبات به جواب زیر بهینه دست پیدا کند.

۲.۳ ۲

همانطور که می‌دانیم در مسئله Hard Margin سعی می گردد نمونه‌ها با استفاده از دو ابر صفحه $|w^T x+b| = 1$ ابر صفحه از یکدیگر به طور کامل جدا شده و تا حد امکان این حاشیه [4] ایجاد شده توسط دو صفحه بیشینه گردد. اما Soft Margin اجازه می‌دهد تعدادی از نمونه‌ها درون این ناحیه حاشیه‌ای قرار داشته باشند و یا حتی تعدادی از نمونه‌ها به اشتباه طبقه بندی شوند. در واقع Soft Margin می‌پذیرد که در ازای تعدادی محدود خطا در طبقه بندی، عمده‌ی داده‌های کلاس را به صورت مقاوم و همراه با حاشیه امن بالا طبقه بندی نماید. به طور کلی در مواقعی که داده‌ها به صورت خطی جداپذیر نیستند و یا در مواقعی که تعداد اندکی داده‌ی پرت در هر دو کلاس وجود دارد روش Soft Margin موفق عمل می‌نماید و کلیت داده‌ها را خوبی تشخیص می‌دهد. برای مثال شکل ۱ نمونه‌ای از مواردی را نشان می‌دهد که در آن داده‌ها به صورت خطی جداپذیر نیستند و عملا استفاده از Soft Margin انتخاب مناسب‌تری نسبت به Hard Margin می‌باشد.

[1] Dual Problem
[2] Primal Problem
[3] Support Vector
[4] Margin



شکل ۱: نمونه‌ای از موارد استفاده SOFT MARGIN

۳.۳ ۳

ضریب C به مانند یک پیچ تنظیم کننده تعینن می‌کند که طبقه به چه میزان باید به برآورده کردن یک حاشیه مناسب توجه کند و چه میزان به قرار گرفتن نمونه‌ها در خارج از حاشیه ایجاد شده و یا جلوگیری از طبقه‌بندی اشتباه توجه نماید. به خصوص در طبقه بندی مواردی که به صورت خطی جدا پذیر نیستند، افزایش C موجب کاهش سایز حاشیه و کاهش تعداد نمونه با تشخیص اشتباه و یا نمونه قرار گرفته در حاشیه می‌گردد. همچنین کاهش C موجب افزایش اندازه حاشیه (Margin) و در نتیجه آن افزایش تعداد نمونه‌ها با تشخیص اشتباه و ... می‌گردد.

۴ سوال۳

۱.۴ مفهوم کلی

همانطور که مطرح شد، یکی از روش‌های حل مسائل غیرخطی انتقال این مسائل به بعد بالاست. می‌توان فضای جدید ویژگی‌ها را طوری تعیین نمود که پیچیدگی‌های غیرخطی آن کاهش پیدا کند و عملا به صورت خطی حل پذیر باشد. مشکل روش انتقال مسئله به بعدهای بالاتر آنست که عملا موجب افزایش هزینه ذخیره‌سازی می‌گردد. برای حل این مشکل، توجه می‌گردد که باتوجه به روش SVM عملا به حاصل ضرب داخلی نمونه‌ها در فضای جدید که بیانگر شباهت نمونه‌ها در این فضا می‌باشد احتیاج است. بنابراین می‌توان تنها با بررسی ضرب داخلی نمونه‌ها تمامی اطلاعات مورد نیاز را جمع آوری کرد. این ضرب داخلی در فضای جدید می‌توان به صورت تابع $K(x,y) : R^d \times R^d \to R$ تعریف بشود. به این تابع هسته [۵] گفته می‌شود. مطابق قضیه‌ی Mercer این تابع متقارن و مثبت معین می‌باشد.

۲.۴ اثبات ناتساوی

از خاصیت مثبت نیمه معین بودن استفاده می‌شود:

$$\forall x, y \in \mathbb{R}^n : \; 0 \le K(\lambda y - x, \lambda y - x) = \lambda^2 K(y,y) - 2\lambda K(y,x) + K(x,x) = f(\lambda) \tag{۵}$$

می‌دانیم در چندجمله‌ای درجه دوم $ax^2 + b^x + c$ ، شرط آنکه چند جمله‌ای همیشه ناصفر باشد آنست که:

$$\begin{aligned} a &> 0 \\ b^2 - 4ac &\le 0 \end{aligned} \tag{۶}$$

حال برای آنکه چندجمله‌ای درجه دوم $f(\lambda)$ در معادله ۵ همواره مثبت باشد از شروط مطرح شده در ۶ استفاده می‌نماییم:

$$\to (2K(x,y))^2 - 4K(y,y)K(x,x) < 0 \to \boxed{K(x,y)^2 \le K(x,x)K(y,y)} \tag{۷}$$

۳.۴ اثبات تساوی میانگین

$$\begin{aligned} &||\mu_\phi|| = \sqrt{\mu_\phi^T \mu_\phi} \\ &\xrightarrow{\mu_\phi = \frac{\phi(x_1)+\ldots+\phi(x_Q)}{Q}} ||\mu_\phi|| = \frac{1}{Q}\sqrt{\Sigma_{m=1}^{Q}\Sigma_{n=1}^{Q}\phi(x_m)^T\phi(x_n)} = \frac{1}{Q}\sqrt{\Sigma_{m=1}^{Q}\Sigma_{n=1}^{Q}K(x_m,x_n)} \\ &\to \boxed{||\mu_\phi|| = \frac{1}{Q}\sqrt{\Sigma_{m=1}^{Q}\Sigma_{n=1}^{Q}K(x_m,x_n)}} \end{aligned} \tag{۸}$$

[5] Kernel

۵ سوال ۴

۱.۵ ۱

ابتدا ماتریس K که درایه‌ی i,j آن برابر با $K(x_i,x_j)$ می‌باشد را بررسی می‌نماییم:

$$K_{n\times n} = \begin{pmatrix} f(x_1)K_1(x_1,x_1)f(x_1) & \cdots & f(x_1)K_1(x_1,x_n)f(x_n) \\ \vdots & \ddots & \vdots \\ f(x_n)K_1(x_n,x_1)f(x_1) & \cdots & f(x_n)K_1(x_n,x_n)f(x_n) \end{pmatrix} = FK_1F^T$$
$$s.t: F = \begin{pmatrix} f(x_1) & 0 & 0 & \cdots & 0 \\ 0 & f(x_2) & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots & \\ 0 & 0 & \cdots & 0 & f(x_n) \end{pmatrix} \tag{۹}$$

حال مثبت معین بودن K را بررسی می‌نماییم. توجه داریم با توجه به کرنل بودن K_1 می‌دانیم که ماتریس متناظر K_1 متقارن و مثبت معین است. بنابراین داریم:

(۱۰)

$$\text{Symmetri}: K = FK_1F^T = FK_1^TF^T = (FK_1F^T)^T = K^T$$
$$\text{PD}: \forall X \in \mathbb{R}^d: X^TKX = X^TFKF^TX = (F^TX)^TK(F^TX) \xrightarrow{F^TX=Y\in R^d} X^TKX = Y^TKY \geq 0$$
$$\rightarrow \boxed{K \geq 0,\ K^T = K}$$

۲.۵ ۲

در این بخش از گزاره‌ی سوم و چهارم همین سوال استفاده می‌گردد. در واقع روند پیشرفت این گزارش بدین صورت است که ابتدا گزاره‌ی سوم چهارم مستقلا اثبات شده، سپس با استفاده از آن به اثبات گزاره‌ی دوم پرداخته می‌شود.

از بسط تیلور استفاده می‌نماییم:

$$K(x,y) = 1 + \Sigma_{i=1}^{\infty}\frac{K_1^i(x,y)}{i!}$$

مطابق گزاره‌ی چهارم می‌دانیم در صورتی که $K_1(x,y)$ یک کرنل باشد، در این صورت 1 و $K_1^i(x,y)$ نیز کرنل هستند و از طرفی مطابق گزاره‌ی سوم، جمع کرنل‌های 1 و $K_1^i(x,y)$ خود یک کرنل جدید است و بدین ترتیب حکم اثبات می‌گردد:

$$K(x,y) = 1 + \Sigma_{i=1}^{\infty}\frac{K_1^i(x,y)}{i!} \xrightarrow[\text{3rd statement: Sum of Kernels is a kernel}]{\text{4th statement:}K_1(x,y)^i\text{is Kernel}} K(x,y)\text{ is a Kernel} \tag{۱۱}$$

۳.۵ ۳

$$K_{n\times n} = \begin{pmatrix} K_1(x_1,x_1)+K_2(x_1,x_1) & \cdots & K_1(x_1,x_n)+K_2(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K_1(x_n,x_1)+K_2(x_n,x_1) & \cdots & K_1(x_n,x_n)+K_2(x_n,x_n) \end{pmatrix} = K_1 + K_2$$

$$\text{Symmetri} : K_{n\times n} = K_{1n\times n} + K_{2n\times n} = K^T_{1n\times n} + K^T_{2n\times n} = (K_{1n\times n} + K_{2n\times n})^T = K^T_{n\times n} \quad (۱۲)$$

$$\text{PD} : \forall X \in \mathbb{R}^d : X^T K_{n\times n} X = X^T(K_{1n\times n} + K_{2n\times n})X = X^T K_{1n\times n} X + X^T K_{2n\times n} X$$

$$\xrightarrow[X^T K_{1n\times n} X>0,\ X^T K_{2n\times n} X>0]{K_{1n\times n}>0,\ K_{2n\times n}>0} X^T K_{n\times n} X = X^T K_{1n\times n} X + X^T K_{2n\times n} X > 0$$

$$\boxed{K^T = K,\ K_{n\times n} > 0}$$

۴.۵ ۴

برای اثبات این گزاره از قضیه‌ی Schur Product استفاده می‌نماییم. این قضیه بیان می‌کند که ضرب عضوبه‌عضو [۶] دو ماتریس مثبت معین یک ماتریس مثبت معین است.

$$K_{n\times n} = \begin{pmatrix} K_1(x_1,x_1)\times K_2(x_1,x_1) & \cdots & K_1(x_1,x_n)\times K_2(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K_1(x_n,x_1)\times K_2(x_n,x_1) & \cdots & K_1(x_n,x_n)\times K_2(x_n,x_n) \end{pmatrix} = K_{1n\times n} \circ K_{2n\times n}$$

$$\text{Symmetri} : K_{n\times n} = K_{1n\times n} \circ K_{2n\times n} = K^T_{1n\times n} \circ K^T_{2n\times n} = (K_{1n\times n} \circ K_{2n\times n})^T = K^T_{n\times n} \quad (۱۳)$$

$$\text{PD} : K_{n\times n} = K_{1n\times n} \circ K_{2n\times n} \xrightarrow[K_{1n\times n}, K_{2n\times n} \geq 0]{\text{Schur Product}} K_{n\times n} \geq 0$$

$$\boxed{K = K^T,\ K_{n\times n} \geq 0}$$

۵.۵ ۵

$$K_{n\times n} = \begin{pmatrix} x_1^T A x_1 & \cdots & x_1^T A x_n \\ \vdots & \ddots & \vdots \\ x_n^T A x_1 & \cdots & x_n^T A x_n \end{pmatrix} = X^T A X; \qquad X^T_{n\times d} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \quad (۱۴)$$

$$\text{Symmetri} : K_{n\times n} = X^T A X = X^T A^T X = (X^T A X)^T = K^T_{n\times n}$$

$$\forall x \in \mathbb{R}^d : x^T X^T A X x = y^T A y; \qquad y \in \mathbb{R}^d, y = Xx$$

[6] Elementwise

۶ سوال ۵

۱.۶ الف

برای حل این بخش از رابطه‌ی اثبات شده در سوال ۳ استفاده می‌نماییم:

$$||\phi(x_1) - \phi(x_2)||^2 = ||\phi(x_1)||^2 + ||\phi(x_2)||^2 - 2\phi(x_1)^T\phi(x_2)$$
$$= K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2) = exp(0) + exp(0) + exp(-0.5||[1, 1] - [3, 4]||) = 1.996 = dist^2$$
$$\rightarrow \boxed{dist = 1.412} \tag{۱۵}$$

۲.۶ ب

$$K(x, y) = (x^Ty + 1)^2 =$$
$$(x_1y_1 + ... + x_dy_d + 1)^2 = 1 + \Sigma_{i=1}^d(x_iy_i)^2 + 2\Sigma_{i=1}^d\Sigma_{j=i+1}(x_iy_i)(x_jy_j) + \Sigma_{i=1}^d 2x_iy_i \tag{۱۶}$$

همانطور که مشاهده می‌شود، عبارت بالا از جمع $1 + d + \frac{d\times(d-1)}{2} + d = \frac{(d+1)(d+2)}{2}$ جمله مستقل از هم(در فضای چند جمله‌ای‌های دو فرمی) تشکیل شده است. هر کدام از اجزا مستقل را اگر h بنامیم، می‌توان آن‌ها را به فرم زیر نوشت:

$$h(x, y) = f(x)g(y) : \begin{cases} f_i(x) = x_i^2, \;\; g_i(y) = y_i^2; \;\; i \in I_1 \\ f_{ij}(x) = x_ix_j, \;\; g_{ij}(y) = y_iy_j \;\; i, j \in I_2 \\ f_i(x) = 1, \;\; g_i(y) = 1 \;\; i \in I_3 \end{cases} \tag{۱۷}$$

که مقصود از I_1, I_2, I_3 مجموعه‌های اندیسگذار هستند. با توجه به مستقل بودن جملات در میابیم که هر جمله باید جداگانه و توسط $f(x), g(y)$ مشخصی تولید گردد که هر $f(x)$ بیانگر یکی از درایه‌های $\phi(x)$ و هر $g(y)$ بیانگر یک درایه‌ی $\phi(y)$ می‌باشد که در اثر ضرب داخلی به صورت زیر در می‌آیند:

$$< \phi(x), \phi(y) > \; = \Sigma_i\phi(x)\phi(y) \tag{۱۸}$$

با توجه به مطالب بیان شده، و از آنجا که تعداد جملات مجزا برابر با $\frac{(d+1)(d+2)}{2}$ می‌باشد بنابراین $\phi(x)$ باید دارای $\frac{(d+1)(d+2)}{2}$ جز باشد، یا به عبارتی دیگر:

$$\phi(x) \in \mathbb{R}^{\frac{(d+1)(d+2)}{2}} \tag{۱۹}$$

۷ سوال ۶

۱.۷ ۱

در رویکرد Generative هدف تخمین توزیع داده‌های کلاس‌های مختلف است. این مدل‌ها معمولا به توزیع‌های احتمال توام منجر می‌گردند. اما در مدل‌های Discriminative سعی می‌گردد مرزهای طبقه‌بندی تشخیص داده شود. در واقع در این رویکرد سعی می‌گردد توزیع داده‌ها به شرط هر کلاس و درون هر کلاس تخمین زده شود تا بدین ترتیب مشکل بالانس نبودن داده‌های کلاس‌های مختلف مرتفع بشود.

۲.۷ ۲

فرض کنید با دیتای مورد بحث به صورت کلی شامل N کلاس مختلف می‌باشد.

۱. One vs Rest در این رویکرد، به تعداد کلاس‌ها(N) طبقه بند می‌سازیم که هر طبقه بند سعی در جداسازی دیتای یک کلاس نسبت به کلاس‌های دیگر دارد. این روش سریع است اما مشکل آن ایسنت که منجر به تولید هیسترزیس می‌شود و عملا یک ناحیه‌ی ابهام می‌سازد. همچنین از دیگر مشکلات این رویکرد می‌توان به حجم نسبتا بالای محاسباتی به دلیل لزوم ساخت وآموزش تعداد زیادی طبقه‌بند اشاره کرد.

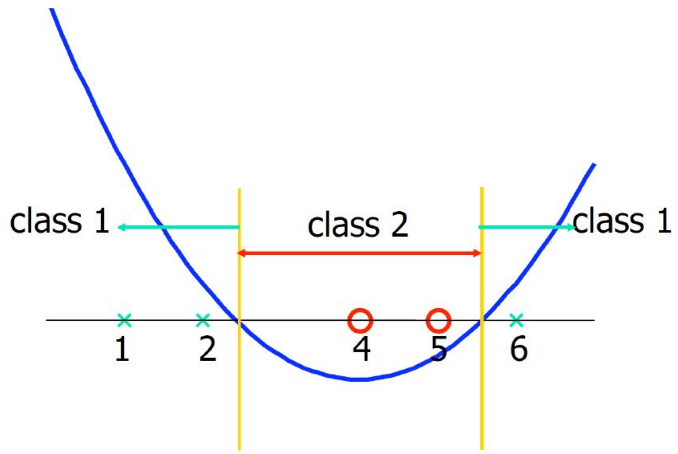
۲. One vs One در این رویکرد به تعداد $\frac{N(N-1)}{2}$ طبقه بند ساخته می‌شود. در واقع سعی می‌گردد به تعداد هر انتخاب دوتایی از کلاس‌های موجود، یک طبقه‌بند ساخته بشود. سپس در انتها رای گیری بین طبقه بندهای موجود صورت می‌گیرد و کلاسی که بیشترین رای را می‌آورد انتخاب می‌گردد. این روش نسبت به روش قبل ابهام کمتری دارد اما هزینه محاسباتی زیادی دارد.

۳. Linear Machine در این روش می‌توان از ابتدا طبقه‌بند را برای مسئله چند کلاسه طرح نمود. در واقع در این روش سعی می‌گردد تعدادی تابع افتراق ساز [7] تعریف بشود، و سپس به محاسبه پارامترها بر اساس قیدهای جدید پرداخت. در واقع مرزبندی‌ها بر اساس تساوی و نا تساوی‌های توابع مذکور بدست می‌آید. همچنین عملکرد بهنیه‌ی توابع مذکور به خصوص در مسائل که ذاتا با توابع Affine سرو کار داریم در صورت تعریف نواحی محدب [8] مزیت این روش بهینه بودن جواب نهایی و واضح بودن نتایج انتهایی است اما محاسبه پارامترها و برآورد کردن قیدهای مختلف عملا حجم محاسبات را افزایش می‌دهد و موجب سختی محاسبه می‌گردد. همچنین در بسیاری از مسائل با توجه به سهولت حل مسئله به صورت Convex نواحی با یک ناحیه‌ی محدب تقریب زده می‌شود که اغلب این جواب‌ها بهینه‌ی تام نیستند.

[7] Discriminant Function
[8] Convex

۳ ۳.۷

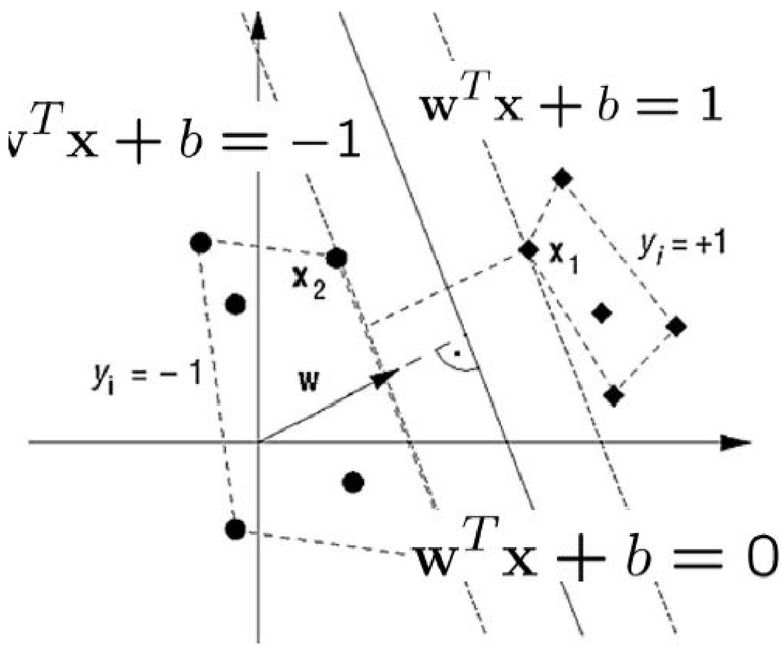
در مسائل طبقه بندی، به منظور حذف وابستگی‌ها و تمیز کردن داده و انتخاب بهینه ویژگی‌ها و به طور کلی Feature Conditioning عمدتا کاهش بعد صورت می‌گیرد. اما در مواردی برای کاهش پیچیدگی مسئله، باز کردن Fold های منیفلدهای هندسی پیچیده، می‌توان با انتقال منیفلد به فضای با بعد بالاتر به شکل هندسی ساده‌تری رسید که توسط طبقه‌بندهای خطی قابل تحلیل باشد. برای مثال در شکل ۲ مشاهده می‌گردد که نقاط ابتدایی که در فضای یک بعدی قرار دارند به صورت خطی قابل جداسازی نیستند اما پس از انتقال این دیتاها به فضای دو بعدی با استفاده از تبدیل مناسب، عملا می‌توان دیتاها را با استفاده زا یک خط جدا کرد.



شکل ۲: انتقال مسئله به بعد بالا می‌تواند موجب حل پذیری مسئله به صورت خطی شود

۴ ۴.۷

کوتاه ترین خط متصل کننده‌ی دو Convex Hull بر ابرصفحه‌ی محاسبه شده عمود است و در نیمه‌ی راه این خط را قطع می‌نماید. در شکل ۳ مثالی از این موضوع آورده شده.



شکل ۳: کوتاه‌ترین خط واصل بر ابرصفحه بهینه عمود است

۵ ۵.۷

$$
\begin{aligned}
&Min\frac{1}{2}||W||^2\\
&s.t\, 1 - y_i(w^T x_i + b) \leq 0\\
&\mathcal{L} = \frac{1}{2}w^T w + \Sigma_{i=1}^{n}\alpha_i(1 - y_i(w^T x_i + b))\\
&\to \frac{\partial \mathcal{L}}{\partial w} = w + \Sigma_{i=1}^{n}\alpha_i(-y_i)x_i = 0 \to \boxed{w = \Sigma_{i=1}^{n}\alpha_i y_i x_i}\\
&\frac{\partial}{\partial b}\mathcal{L} = \boxed{\Sigma_{i=1}^{n}\alpha_i y_i = 0}\\
&\xrightarrow{w=\Sigma_{i=1}^{n}\alpha_i y_i x_i} \mathcal{L} = \frac{1}{2}\Sigma_{i=1}^{n}\Sigma_{j=1}\alpha_i y_i x_i^T \Sigma_{j=1}^{n}\alpha_j y_j x_j + \Sigma_{i=1}^{n}\alpha_i(1 - y_i(\Sigma_{j=1}^{n}\alpha_j y_j x_j^T x_i + b))\\
&= \frac{-1}{2}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} + \Sigma_{i=1}^{n}\alpha_i\\
&\to \boxed{max.\ W(\alpha) = \Sigma_{i=1}^{n}\alpha_i - \frac{1}{2}\Sigma_{i=1,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j}\\
&\boxed{\Sigma_{i=1}^{n}\alpha_i y_i = 0, \qquad \alpha_i \geq 0}
\end{aligned}
\tag{۲۰}
$$

۶ ۶.۷

خیر تفاوتی ایجاد نمی‌شود و تنها بر روی ضرایب لاگرانژ کرانی مطابق با ضریب رگوله قرار می‌گیرد.