تکلیف هشت درس شناسایی الگو

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۲۲ آذر ۱۴۰۴

حل سوال ۵: طراحی دسته‌بند SVM

داده‌های اولیه عبارتند از:

$$x_1 = (0,0), \quad y_1 = +1$$
$$x_2 = (4,4), \quad y_2 = -1$$

الف) طراحی SVM خطی و تعیین بردارهای پشتیبان

۱. تعیین بردارهای پشتیبان: چون تنها دو داده داریم و این دو داده از دو کلاس متفاوت هستند، هر دو داده روی حاشیه (Margin) قرار می‌گیرند تا فاصله حاشیه حداکثر شود. بنابراین، هر دو نقطه x_1 و x_2 **بردارهای پشتیبان (Support Vectors)** هستند.

۲. یافتن مرز تصمیم گیری: در SVM خطی با دو نقطه متقارن (یا کمینه)، مرز تصمیم گیری عمودمنصف پاره‌خط واصل بین دو بردار پشتیبان خواهد بود.

- نقطه وسط m:

$$m = \frac{x_1 + x_2}{2} = \frac{(0,0)+(4,4)}{2} = (2,2)$$

- بردار هادی خط واصل v:

$$v = x_2 - x_1 = (4-0, 4-0) = (4,4)$$

- بردار نرمال صفحه جداکننده w: جهت بردار وزن w باید به سمت کلاس مثبت $(+1)$ باشد. بنابراین w هم‌راستا با بردار $x_1 - x_2$ (از منفی به مثبت) است:

$$w \propto x_1 - x_2 = (-4,-4)$$

برای سادگی می‌توان معادله خطی که از $(2,2)$ می‌گذرد و بر $(1,1)$ عمود است را نوشت:

$$1(x-2) + 1(y-2) = 0 \Rightarrow x + y - 4 = 0$$

حال باید علامت نامساوی را چک کنیم. برای کلاس $y = +1$ (نقطه $0,0$):

$$0 + 0 - 4 = -4 < 0$$

این جهت منفی است، پس برای اینکه خروجی مثبت باشد باید کل معادله را در -1 ضرب کنیم:

$$-x - y + 4 > 0$$

یا به فرم استاندارد $w_1x^{(1)} + w_2x^{(2)} + b \geq 0$:

$$-1x_1 - 1x_2 + 4 \geq 0$$

(دقت کنید که در فرمول سوال x_1 و x_2 به عنوان مولفه‌های بردار ورودی x در نظر گرفته شده‌اند).

تکمیل دسته‌بند:

$$h(x) = \begin{cases} +1 & \text{fi} \quad -\mathbf{1}x_1 + -\mathbf{1}x_2 + \mathbf{4} \geq 0 \\ -1 & otherwise \end{cases}$$

ب) اضافه شدن داده $x_3 = (-1, -1)$ با $y_3 = +1$

ما در مرحله قبل مرز تصمیم گیری $g(x) = -x_1 - x_2 + 4$ را به دست آوردیم. حال داده جدید را در این معادله چک می کنیم:

$$g(x_3) = -(-1) - (-1) + 4 = 1 + 1 + 4 = 6$$

مقدار 6 بسیار بزرگتر از 1 است (شرط حاشیه $g(x) \geq 1$ بود). این یعنی داده x_3 به درستی دسته‌بندی شده و فاصله آن از مرز بسیار بیشتر از فاصله x_1 تا مرز است.

نتیجه:

- **بردارهای پشتیبان:** تغییری نمی کنند (همان x_1 و x_2). x_3 بردار پشتیبان نیست.
- **مرز تصمیم گیری:** تغییری نمی کند و همان مرز حالت (الف) باقی می‌ماند.
- **وزن بردارهای پشتیبان:** چون مسئله بهینه‌سازی روی همان نقاط قبلی حل می‌شود، وزن‌ها (α) **تغییری نمی کنند.**

پ) اضافه شدن داده $x_4 = (2, 2)$ با $y_4 = +1$

نقطه $x_4 = (2, 2)$ دقیقا روی مرز تصمیم گیری قبلی قرار دارد (وسط x_1 و x_2). اما کلاس آن $+1$ است. اکنون نزدیک‌ترین داده‌ها به هم از دو کلاس مختلف، $x_4 = (2, 2)$ با کلاس $+1$ و $x_2 = (4, 4)$ با کلاس -1 هستند. داده‌های x_1 و x_3 اکنون دورتر از حاشیه جدید قرار می گیرند و تاثیری در ساخت مرز ندارند.

تحلیل:

- **بردارهای پشتیبان:** بردارهای جدید x_4 و x_2 هستند.
- **مرز تصمیم گیری:** عمودمنصف پاره‌خط بین $(2, 2)$ و $(4, 4)$ خواهد بود.

$$\text{Midpoint} = \frac{(2, 2) + (4, 4)}{2} = (3, 3)$$

معادله جدید (با حفظ جهت به سمت کلاس مثبت):

$$-(x - 3) - (y - 3) = 0 \Rightarrow -x - y + 6 = 0$$

بنابراین مرز تصمیم گیری **تغییر کرده است.**

- **وزن بردارهای پشتیبان:** فاصله بین دو کلاس (Margin Width) کاهش یافته است. در حالت (الف) فاصله x_1 تا x_2 برابر $\sqrt{32}$ بود. در اینجا فاصله x_4 تا x_2 برابر $\sqrt{8}$ است. رابطه وزن α با فاصله داده‌ها در حالت دو نقطه‌ای متناسب با معکوس مجذور فاصله است ($\alpha \propto \frac{1}{\|x_+ - x_-\|^2}$). چون فاصله کمتر شده است، وزن‌های α (قدر مطلق بردار وزن w) باید **بزرگتر** شوند تا شیب تندتری برای جدا کردن نقاط نزدیک ایجاد کنند.
پاسخ: وزن بردارهای پشتیبان نسبت به حالت قبل **بزرگتر** هستند.

ت) اضافه شدن داده $x_5 = (-3, -3)$ با $y_5 = -1$

بیایید توزیع داده‌ها را روی خط $y = x$ (که همه داده‌ها روی آن هستند) بررسی کنیم:

- $x_5 = -3 \rightarrow$ کلاس -1
- $x_3 = -1 \rightarrow$ کلاس $+1$
- $x_1 = 0 \rightarrow$ کلاس $+1$
- $x_4 = 2 \rightarrow$ کلاس $+1$
- $x_2 = 4 \rightarrow$ کلاس -1

الگو به صورت (منفی - مثبت - منفی) است. داده‌های کلاس مثبت توسط داده‌های کلاس منفی محاصره شده‌اند.
بررسی کرنل‌ها:

- **کرنل خطی:** قادر به جداسازی نیست (چون داده‌ها به صورت خطی جداپذیر نیستند).
- **کرنل چندجمله‌ای (درجه $\geq$ ۲):** بله، قادر است. یک چندجمله‌ای درجه ۲ (سهمی) می‌تواند دو ریشه داشته باشد و بازه‌ی بین ریشه‌ها را مثبت و خارج آن‌ها را منفی در نظر بگیرد.
- **کرنل گاوسی (RBF):** بله، قادر است. این کرنل توانایی مدل‌سازی مرزهای تصمیم‌گیری پیچیده و محلی را دارد و می‌تواند ناحیه وسط را به عنوان کلاس مثبت جدا کند.

نتیجه نهایی: کرنل‌های **چندجمله‌ای (با درجه $n \geq 2$)** و **کرنل گاوسی** قادر به جداسازی این داده‌ها هستند.