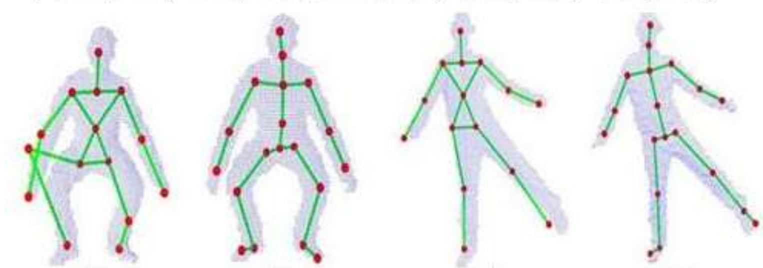
بسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

گروه سیستم‌های دیجیتال



آزمایشگاه یادگیری و بینایی ماشین

دستور کار آزمایش چهارم: طبقه بندی به روش بردارهای پشتیبان

زمان لازم برای انجام آزمایش: حداکثر یک جلسه

طبقه بند SVM

در این آزمایش به طبقه‌بندی به روش Support Vector Machines (SVM) می‌پردازیم. طبقه‌بند SVM یکی از روش‌های مهم و قدرتمند طبقه‌بندی است. در ادامه به توضیح این روش پرداخته و سپس به آزمایش آن می‌پردازیم.

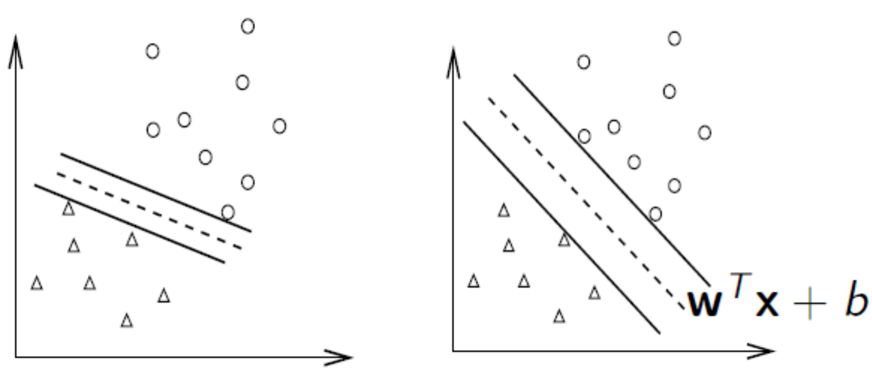
فرض کنید که N تا داده آموزش x^i در اختیار داریم:

$$x^i, \quad i = 1, \dots, N$$

فرض کنید که دو کلاس داریم:

$$y^i = \begin{cases} -1 & \text{if } x^i \text{ in class 1} \\ +1 & \text{if } x^i \text{ in class 2} \end{cases}$$

اگر داده‌های این دو کلاس را رسم کنیم، ابرصفحه‌های (Hyperplane های) زیادی برای تفکیک آنها در فضای آماری می‌توان پیشنهاد داد:



همان‌طور که در شکل بالا دیده می‌شود، معادله ابرصفحه را به صورت زیر در نظر می‌گیریم:

$$\text{hyperplane:}\, W^T x + b = 0$$

این ابرصفحه را با توجه داده‌های کلاس‌ها و شکل بالا می‌توان به صورت زیر تعیین علامت کرد:

$$\begin{cases} (W^T x^i) + b > 0 & if\ y_i = 1 \\ (W^T x^i) + b < 0 & if\ y_i = -1 \end{cases}$$

ولی ما علاقه داریم که یک رابطه به جای دو رابطه داشته باشیم. لذا دو رابطه بالا را به صورت زیر خلاصه می‌کنیم:

$$y^i\left(W^T x^i + b\right) \geq 1 \qquad i = 1, \dots, N$$

همان‌طور که از شکل بالا دیده می‌شود، به نظر می‌رسد که نمودار سمت راست بهتر توانسته عمل تفکیک را انجام دهد. علت آن را می‌توان در این دید که این روش، بیشترین فاصله را نزدیک‌ترین داده‌ها به مرز دارد. در شکل بالا، خط نقطه‌چین، مرز است و دو خط اطرافش هم موازی آن و گذرنده از نزدیکترین داده‌ها به مرز هستند. این دو خط همان خطوط $\left(W^T x^i + b\right) = \pm 1$ است. فاصله بین این دو خط برابر است با:

$$d = \frac{2}{||W||} = \frac{2}{\sqrt{W^T W}}$$

ما قصد داریم که این فاصله ماکزیمم شود که معادل این است که مقدار $\frac{1}{2} W^T W$ مینیمم شود. لذا در کل مساله بهینه‌سازی زیر را داریم:

$$\min_{W,b} \frac{1}{2} W^T W \qquad (\text{subject to} \quad y^i\left(W^T x^i + b\right) \geq 1 \qquad i = 1, \dots, N)$$

اگر مساله بهینه‌سازی بالا را حل کنیم، خواهیم داشت:

$$W = \sum_{i=1}^{N} \alpha_i y^i x^i$$

$$b = \text{average}(y^i - W^T x^i)$$

در عبارت بالا α_i ضریب لاگرانژ است که عددی مثبت است و از حل مساله بهینه سازی بدست می آید. اگر x_i دارای ابعاد M×1 باشد، W هم دارای ابعاد M×1 است. و b هم یک عدد اسکالر است.

برای تخمین کلاس یک داده تست x به صورت زیر عمل می‌کنیم:

$$y = \text{sign}(W^T x + b)$$

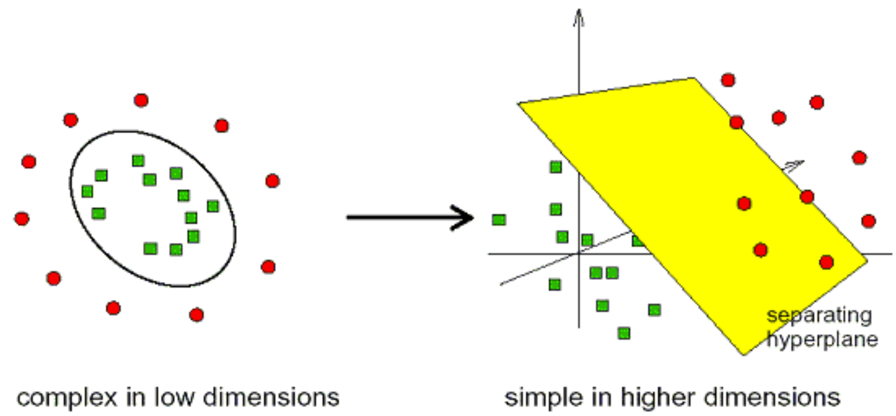
داده‌ای که دارای $y = -1$ باشد، به کلاس ۱ و داده‌ی $y = 1$ به کلاس ۲ تعلق دارد.

آزمایش طبقه بندی بین سه حالت خنثی، تعجب و خوشحال را روی دیتاست Cohn-Kanade با استفاده از روش SVM و با حاشیه نرم انجام دهید. در روش SVM با حاشیه نرم، بردارهای پشتیبان

می توانند از مرز حاشیه عبور کنند که میزان این عبور با پارامتر C کنترل می شود. از آنجا که ۳ کلاس در این آزمایش وجود دارد لازم است طبقه بندهای جداگانه بین هر حالت و بقیه حالتها بسازید. برچسب نهایی یک نمونه ی دلخواه تست، توسط خروجی آن طبقه بندی مشخص خواهد شد که ادعا میکند با بیشترین قاطعیت برچسب این نمونه را تشخیص داده است[۱]. از مجموعه های neutral، happy، surprise و disgust به ترتیب ۱۲۰، ۶۰ ، ۶۰ و ۳۰ تصویر را به صورت تصادفی به عنوان نمونه های آموزش و مابقی تصاویر را به عنوان نمونه های تست جدا کنید. با آموزش یک فضای PCA با استفاده از داده های آموزش، تمام داده های آموزش و تست را کاهش بعد بدهید. پس از آموزش و تست طبقه بند، با تشکیل یک ماتریس درهم ریختگی[۲]، نتیجه طبقه بندی را گزارش نمایید. علاوه بر دقت تشخیص، معیارهای Precision و Recall[3] را نیز محاسبه کنید. دقت نمایید که هنگام آموزش طبقه بند نیاز به تنظیم پارامتر C دارید. برای این کار، لازم است تا بهترین مقدار را برای این پارامتر از میان چند مقدار از پیش تعیین شده با انجام cross-validation بر روی مجموعه آموزش پیدا کنید. در این قسمت، لازم است cross-validation را **خودتان پیاده سازی کنید و مجاز به استفاده از تابع آماده برای انجام آن نیستید**. مقدار C بدست آمده را گزارش کنید و نتیجه طبقه بندی بر روی مجموعه تست را با استفاده از این مقدار و دیگر مقادیری که از پیش تعیین کرده بودید باهم مقایسه کنید.

حال، سوال مهمی مطرح می‌شود. اگر داده‌ها به صورت خطی قابل تفکیک نباشند چه کار باید کرد. نمونه‌ای از این توزیع داده‌ها را در زیر مشاهده می‌کنید. همان‌طور که در شکل زیر دیده می‌شود، اگر این داده‌ها را به ابعاد بالاتر ببریم (ابعاد آنها را افزایش دهیم)، می‌توانیم آنها را با یک ابرصفحه خطی جدا کنیم.

[1] One Versus Rest (OVR)
[2] https://en.wikipedia.org/wiki/Confusion_matrix
[3] https://en.wikipedia.org/wiki/Precision_and_recall



به این افزایش ابعاد داده، کرنل (kernel) گفته می‌شود. برای افزایش بُعد داده، آن را تحت تبدیل زیر قرار می‌دهیم:

$$\phi(x) = (\phi_1(x), \phi_2(x), \ldots)$$

برای مثال، اگر x سه‌بعدی ($x = (x_1, x_2, x_3)$) باشد، می‌توان $\phi(x)$ را ۱۰ بعدی به صورت زیر ساخت:

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)$$

پس به جای x، $\phi(x)$ را قرار می‌دهیم:

$$W = \sum_{i=1}^{N} \alpha_i y^i \phi(x^i)$$

پس برای تشخیص کلاس داده تست x داریم:

$$y = \text{sign}(W^T\phi(x) + b) = \text{sign}(\sum_{i=1}^{N} \alpha_i y^i \phi(x^i)^T \phi(x) + b)$$

عبارت $\phi(x^i)^T\phi(x)$ را کرنل می‌خوانند و رابطه بالا به صورت زیر نوشته می‌شود.

$$y = \text{sign}(\sum_{i=1}^{N} \alpha_i y^i K(x^i, x) + b)$$

می توان نشان داد که در حل مساله بهینه سازی نیز فقط به مقادیر کرنل بین دو بدوی نمونه ها نیاز است و به خود بردارهای ویژگی احتیاجی نیست.

کرنل‌های گوناگونی موجود است. سه نمونه از آنها را در زیر مشاهده می‌کنید:

Linear Kernel: $K\left(x^i, x^j\right) = {x^i}^T x^j$

Polynomial Kernel: $K\left(x^i, x^j\right) = \left(1 + {x^i}^T x^j\right)^p$

Gaussian (Radial Basis Function (RBF)) Kernel: $K\left(x^i, x^j\right) = \exp\left(-\frac{\left|\left|x^i - x^j\right|\right|^2}{2\sigma^2}\right)$

رابطه b هم به طور مشابه قابل محاسبه است که به صورت زیر می‌شود.

$$b = \text{average}\left(y^i - W^T\phi\left(x^i\right)\right) = \text{average}\left(y^i - \sum_{j=1}^{N} \alpha_j y^j\, K(x^i, x^j)\right)$$

طبقه بندی احساس را این بار با Kernel SVM انجام دهید. برای این منظور، با انجام cross-validation بر روی مجموعه آموزش، کرنل rbf با سیگمای مناسب را انتخاب کنید و نتیجه را با حالت SVM خطی مقایسه و تحلیل نمایید. دقت نمایید که هم پارامتر C و هم پارامتر سیگما نیاز به تنظیم دارند. در این قسمت برای تنظیم این دو پارامتر می توانید از تابع آماده GridSearchCV استفاده نمایید.

پیش گزارش

دو کلاس ۱ و ۱- داریم و می‌خواهیم روشی برای طبقه‌بندی داده‌هایشان پیدا کنیم.

$Class_1 = [(1, 1), (-1, -1)]$

$Class_{-1} = [(1, -1), (-1, 1)]$

- آیا این داده‌ها به صورت خطی جدایی پذیرند؟ خیر.
- $\varphi(x) = (1, x_1, x_2, x_1 x_2)$ و $y(x) = w^T * \varphi(x)$ را تعریف کنید. (x_1 بعد اول x و x_2 بعد دوم آن است.) w را چنان تعیین کنید که داده‌ها در این فضای جدید به صورت خطی جداپذیر باشند. (y بتواند داده‌ها را طبقه‌بندی کند.)
- کرنل $K(x, x') = \varphi(x)\varphi(x')$ را طوری پیدا کنید که ویژگیهای تابع تبدیل $\varphi(x)$ یا ضرایبی از آنها را شامل باشد. چرا این تابع کرنل می تواند دو کلاس بالا را از هم تفکیک کند؟

$$y^i \left(w^T \phi(x^i) \right) \geq 1$$

$$1 \times \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \geq 1 \Rightarrow w_1 + w_2 + w_3 + w_4 \geq 1$$

$$1 \times \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \times \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \geq 1 \Rightarrow w_1 - w_2 - w_3 + w_4 \geq 1$$

$$-1 \times \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \geq 1 \Rightarrow -w_1 - w_2 + w_3 + w_4 \geq 1$$

$$-1 \times \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \end{bmatrix} \times \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \geq 1 \Rightarrow -w_1 + w_2 - w_3 + w_4 \geq 1$$

Solve the opti —— $\Rightarrow$ $\boxed{W = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}$

$k(x, x') = \phi(x)\phi(x') = 1 + x_1 x_1' + x_2 x_2' + x_1 x_2 x_1' x_2'$

the kernel can extract more complicated featers (non-linear) and we can classify data points in a higher dimensional space by a linear boundry.