به نام یگانه معبود بخشنده مهربان

یادگیری ماشین

Machine Learning

گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اصفهان

ترم دوم سال تحصیلی ۹۱ – ۹۲

ارائه دهنده : پیمان ادیبی

ماشین های بردار پشتیبان

Support Vector Machines

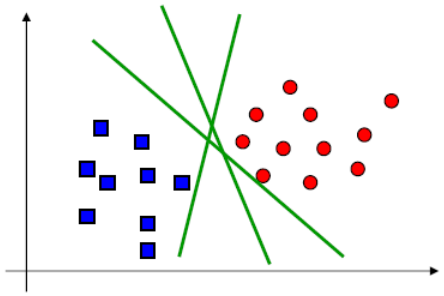
دسته بندی مجموعه های جدایی پذیر خطی

- اگر مجموعه نمونه‌های دو دسته جدایی پذیر خطی باشند، میتوان ابرصفحه ای برای جداسازی آنها بدست آورد. با یافتن وزنها بنحویکه:

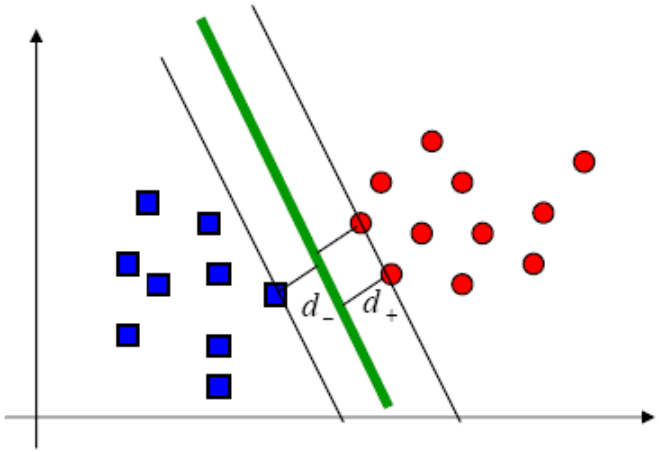
$$\mathbf{w}^T\mathbf{x}_i + w_0 \geq 0 \quad \text{For all i, such that} \quad y_i = +1$$

$$\mathbf{w}^T\mathbf{x}_i + w_0 \leq 0 \quad \text{For all i, such that} \quad y_i = -1$$

$$\longrightarrow \quad y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 0$$

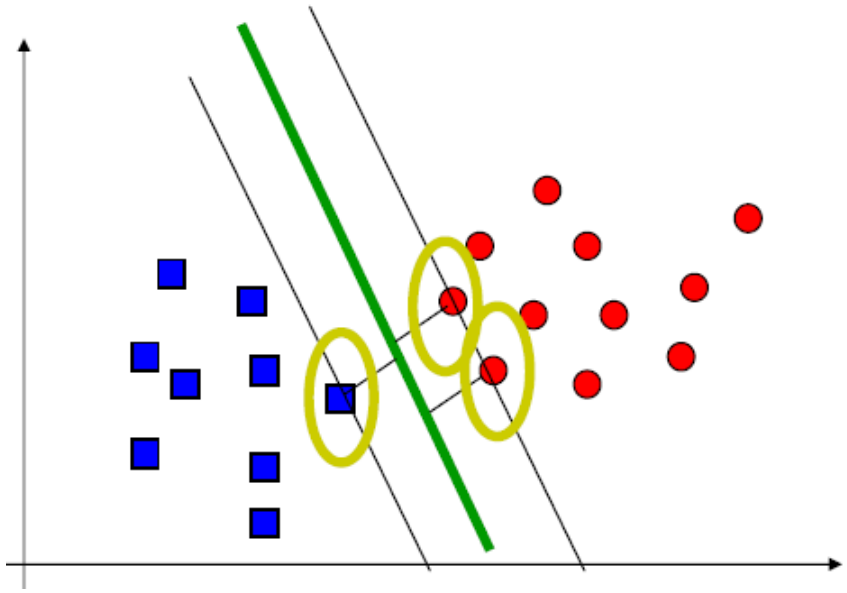


- سئوال: معمولاً ابرصفحه‌های متعددی برای این کار میتوان یافت. کدام یک را انتخاب کنیم؟
- پاسخ: ابر صفحه با بیشترین حاشیه (Maximum Margin) انتخاب شود.
- یعنی بیشترین فاصله $d_+ + d_-$ را داشته باشد، که d_+ کوتاهترین فاصله یک نمونه مثبت از ابرصفحه بوده و d_- کوتاهترین فاصله یک نمونه منفی از آن میباشد.



√484

یافتن ابرصفحه با بیشترین حاشیه



- برای ابرصفحه با بیشترین حاشیه، تنها نمونه های روی حاشیه مهم هستند (بر فواصل تأثیر میگذارند).
- به این نمونه ها بردارهای پشتیبان (**support vectors**) گفته میشود.
- زوج نمونه های آموزشی:

$$(\mathbf{x}_i, y_i) \qquad y_i \in \{+1, -1\}$$

- فرض میکنیم تمام نمونه‌ها شرایط زیر را برآورده میکنند:

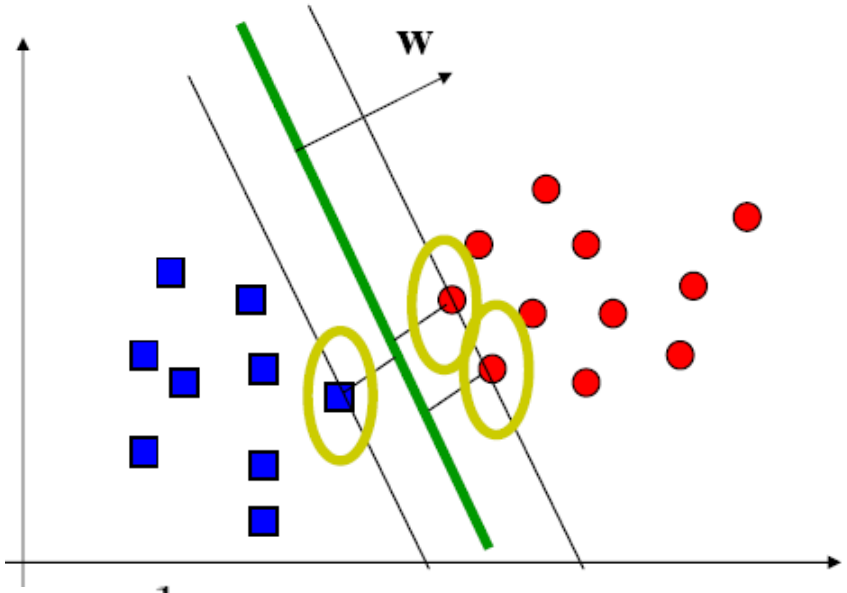
$$\mathbf{w}^T\mathbf{x}_i + w_0 \geq 1 \quad \text{for} \quad y_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + w_0 \leq -1 \quad \text{for} \quad y_i = -1$$

$$\text{combined as: } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 \geq 0 \quad \text{for all} \quad i$$

- نامساویها معرف دو ابرصفحه زیر هستند:

$$\mathbf{w}^T\mathbf{x}_i + w_0 = 1 \qquad \mathbf{w}^T\mathbf{x}_i + w_0 = -1$$

یافتن ابرصفحه با بیشترین حاشیه



- **حاشیه هندسی:** فاصله یک نقطه $\mathbf{x}$ از ابرصفحه:

$$\rho_{\mathbf{w},w_0}(\mathbf{x},y) = y(\mathbf{w}^T\mathbf{x} + w_0) / \|\mathbf{w}\|_{L2}$$

که $\mathbf{w}$ بردار عمود بر ابرصفحه و $\|\cdot\|_{L2}$ نرم اقلیدسی است.

- برای نقاطی که شرط $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 = 0$ را برقرار میسازند، فاصله میشود: $\frac{1}{\|\mathbf{w}\|_{L2}}$
- بنابراین پهنای حاشیه $d_+ + d_- = \frac{2}{\|\mathbf{w}\|_{L2}}$ میشود، که میخواهیم آنرا بیشینه کنیم.
- این کار با کمینه کردن $\|\mathbf{w}\|_{L2}^2 / 2 = \mathbf{w}^T\mathbf{w} / 2$ همراه با حفظ محدودیتهای $\left[y_i(\mathbf{w}^T\mathbf{x} + w_0) - 1\right] \geq 0$ بر روی نقاط داده انجام میگیرد.

یافتن ابرصفحه با بیشترین حاشیه

- **راه حل: وارد کردن محدودیتها در معیار بهینه سازی (روش لاگرانژ):**

$$J(\mathbf{w}, w_0, \alpha) = \|\mathbf{w}\|^2 / 2 - \sum_{i=1}^{n} \alpha_i \left[y_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 \right]$$

$$\alpha_i \geq 0 \Longrightarrow \textbf{Lagrange multipliers}$$

  - کمینه سازی بر حسب $\mathbf{w}, w_0$ (متغیرهای اصلی)
  - بیشینه سازی بر حسب $\boldsymbol{\alpha}$ ها (متغیرهای دوگان)

ضرایب لاگرانژ ارضاء محدودیتها را ایجاب میکند:

$$\text{If} \quad \left[y_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 \right] > 0 \Longrightarrow \alpha_i \to 0$$

$$\text{Else} \Longrightarrow \alpha_i > 0 \qquad \text{Active constraint}$$

  - قرار دادن مشتقها برابر با صفر:

$$\nabla_{\mathbf{w}} J(\mathbf{w}, w_0, \alpha) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \bar{0}$$

$$\frac{\partial J(\mathbf{w}, w_0, \alpha)}{\partial w_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

یافتن ابرصفحه با بیشترین حاشیه

- حال باید ضرایب لاگرانژ را از بیشینه سازی عبارت زیر بدست آوریم:

$$J(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) \quad \Leftarrow \quad \text{maximize}$$

البته با محدودیتهای: $\sum_{i=1}^{n} \alpha_i y_i = 0$ and for all i, $\alpha_i \geq 0$

- یک مسأله بهینه سازی درجه دو (Quadratic Programming) پاسخهای $\hat{\alpha}_i$ را میدهد (برای تمام **i** ها).
- بنابراین بردار پارامتر میتواند بصورت $\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i$ بدست آمده که $\hat{\alpha}_i$ ها پاسخ مسأله دوگان هستند.
- پارامتر w_0 نیز از شرایط Karush-Kuhn-Tucker (KKT) بدست می آید:

$$\hat{\alpha}_i [y_i (\hat{\mathbf{w}} \mathbf{x}_i + w_0) - 1] = 0$$

خواص پاسخها:

- برای تمام نقاطی که روی مرزهای حاشیه نیستند داریم: $\hat{\alpha}_i = 0$
- $\hat{\mathbf{w}}$ تنها ترکیبی خطی از بردارهای پشتیبان میشود.
- مرز تصمیم گیری:

$$\hat{\mathbf{w}}^T \mathbf{x} + w_0 = \sum_{i \in SV} \hat{\alpha}_i y_i (\mathbf{x}_i^T \mathbf{x}) + w_0 = 0$$

ماشین بردار پشتیبان

- مرز تصمیم گیری: $\hat{\mathbf{w}}^T\mathbf{x} + w_0 = \sum_{i \in SV} \hat{\alpha}_i y_i (\mathbf{x}_i^T \mathbf{x}) + w_0$
- تصمیم: $\hat{y} = \text{sign}\left[\sum_{i \in SV} \hat{\alpha}_i y_i (\mathbf{x}_i^T \mathbf{x}) + w_0 \right]$
- تصمیم گیری درباره یک $\mathbf{x}$ جدید نیاز به محاسبه ضرب داخلی آن با نمونه ها دارد $(\mathbf{x}_i^T \mathbf{x})$.
- بطور مشابه، بهینه سازی نیز وابسته به محاسبه $(\mathbf{x}_i^T \mathbf{x}_j)$ ها است:

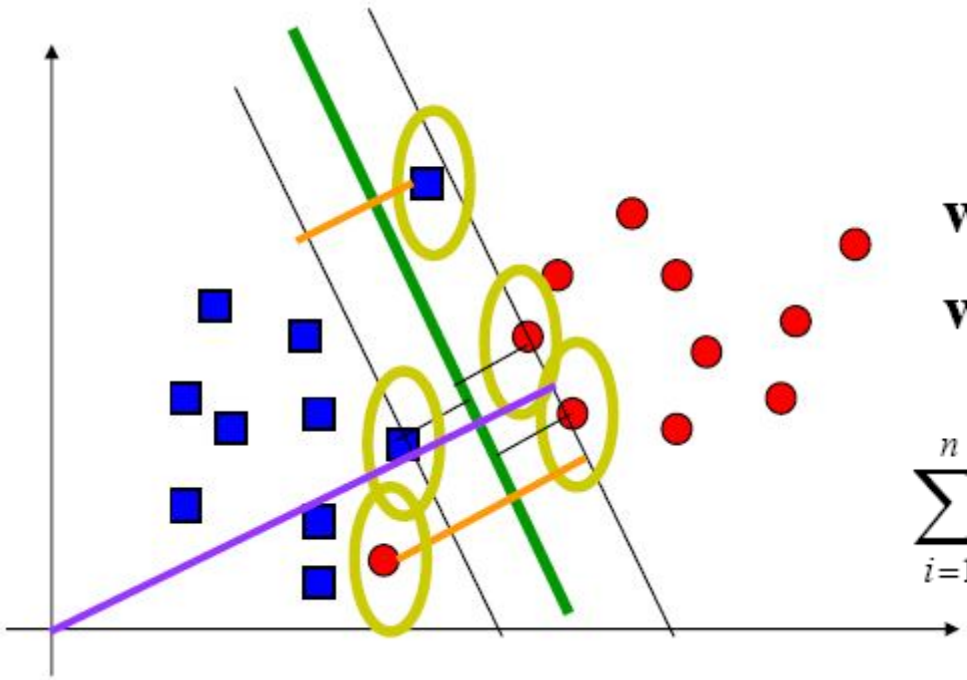
$$J(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

توسعه برای حالت جدایی ناپذیر خطی

- **آزاد سازی نسبی محدودیت ها**

با متغیرهای $\xi_i \geq 0$ **بشکل زیر:**

$$\mathbf{w}^T\mathbf{x}_i + w_0 \geq 1 - \xi_i \quad \text{for} \quad y_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + w_0 \leq -1 + \xi_i \quad \text{for} \quad y_i = -1$$



- **خطا بازاء** $\xi_i \geq 1$ **رخ میدهد.**
- **یک حد بالا برای تعداد خطا:** $\sum_{i=1}^{n} \xi_i$
- **یک جمله جریمه برای تعداد خطا در نظر میگیریم:**

$$\text{minimize} \quad \|\mathbf{w}\|^2 / 2 + C\sum_{i=1}^{n} \xi_i$$

Subject to constraints

که مقدار C توسط کاربر تعیین میشود. بزرگ بودن آن به معنای لحاظ نمودن جریمه بیشتر برای خطاست.

√448

توسعه برای حالت جدایی ناپذیر خطی

- تابع لاگرانژین برای فرم اصلی مسأله:

$$J(\mathbf{w}, w_0, \alpha) = \|\mathbf{w}\|^2 / 2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[y_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 + \xi_i \right] - \sum_{i=1}^{n} \mu_i \xi_i$$

- فرم دوگان مسأله، پس از جایگذاری $\mathbf{w}, w_0$ (با گرفتن $0 \le \alpha_i \le C$ و ξ_i ها حذف میشوند).

$$J(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i{}^T \mathbf{x}_j)$$

$$\text{Subject to: } \quad 0 \le \alpha_i \le C \quad \text{for all i, and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- پاسخ: $\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i$ (تفاوت با حالت جدایی پذیر: $0 \le \alpha_i \le C$)
- پارامتر w_0 نیز از شرایط **KKT** بدست می آید.
- تصمیم گیری مشابه قبل:

$$\hat{y} = \text{sign}\left[\sum_{i \in SV} \hat{\alpha}_i y_i (\mathbf{x}_i{}^T \mathbf{x}) + w_0 \right]$$

√448

ایجاد مرز غیرخطی

- در حالت خطی نیاز به محاسبه ضربهای داخلی $(\mathbf{x}_i^T \mathbf{x})$ داشتیم.
- برای تعمیم به حالت غیرخطی، ورودیها را به بردارهای ویژگی (معمولاً با ابعاد بیشتر) نگاشت میکنیم: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$

- حال روش SVM را بر روی بردارهای ویژگی انجام میدهیم. لذا ضربهای داخلی بردارهای ویژگی نمونه ها را انجام میدهیم:

$$\varphi(\mathbf{x})^T \varphi(\mathbf{x}')$$

- تابع هسته (Kernel function): تعریف میکنیم:

$$K(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x})^T \varphi(\mathbf{x}')$$

- نکته کلیدی: با انتخاب مناسب تابع هسته میتوان جدایی پذیری خطی را در فضای ویژگیها داشت، ولو آنکه جدایی پذیری خطی در فضای اصلی برقرار نباشد.

مثالی از تابع هسته

- فرض کنید $\mathbf{x} = [x_1, x_2]^T$ ، و نگاشت به یک مجموعه ویژگی درجه دو انجام میدهیم:

$$\mathbf{x} \to \boldsymbol{\varphi}(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1]^T$$

- تابع هسته متناظر با این فضای ویژگی:

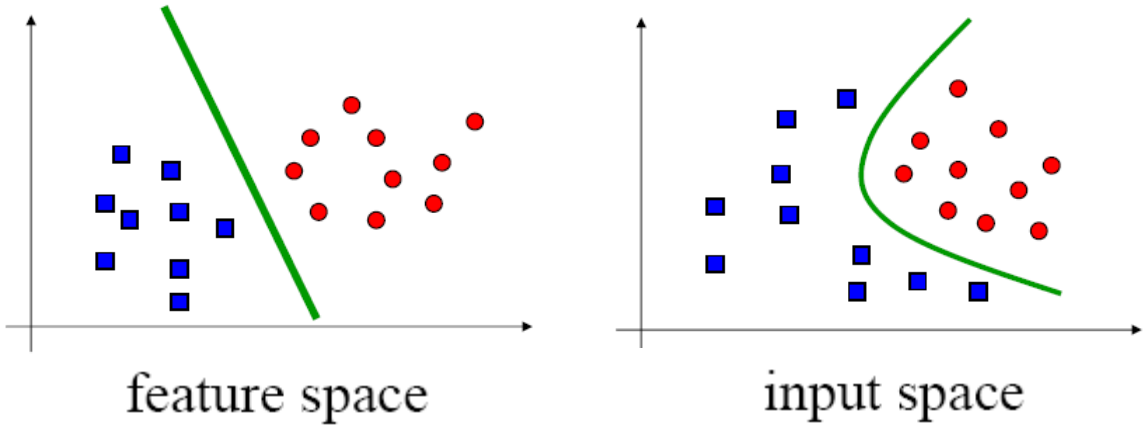
$$K(\mathbf{x}', \mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x}')^T \boldsymbol{\varphi}(\mathbf{x})$$

$$= x_1^2 x'^2_1 + x_2^2 x'^2_2 + 2x_1x_2x'_1 x'_2 + 2x_1x'_1 + 2x_2x'_2 + 1$$

$$= (x_1x'_1 + x_2x'_2 + 1)^2$$

$$= (1 + (\mathbf{x}^T\mathbf{x}'))^2$$

- جداسازی خطی در فضای ویژگی بطور ضمنی، جداسازی غیرخطی در فضای اصلی را موجب میشود:



√524

توابع هسته

- هسته خطی (Linear):

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$$

- هسته چندجمله ای (Polynomial):

$$K(\mathbf{x}, \mathbf{x}') = \left[1 + \mathbf{x}^T \mathbf{x}'\right]^k$$

- هسته پایه شعاعی (Radial basis):

$$K(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right]$$

- مقدار تابع هسته منعکس کننده نوعی شباهت (similarity) بین ورودیهای آن است.
- هسته ها میتوانند برای اشیاء پیچیده تر هم تعریف شوند. مثل رشته ها، گراف ها، تصویر ها، ... . لذا از SVM میتوان برای دسته بندی در حوزهای مختلف استفاده نمود.
- معیارهای انتخاب هسته:
  - معتبر بودن (valid)
  - خوب بودن (good)
  - مناسب بودن (appropriate)
  - کارآمد بودن (efficient)