PS5841

Data Science in Finance & Insurance

SVC & SVM

Yubo Wang

Spring 2022

Columbia University

Geometry (2D)

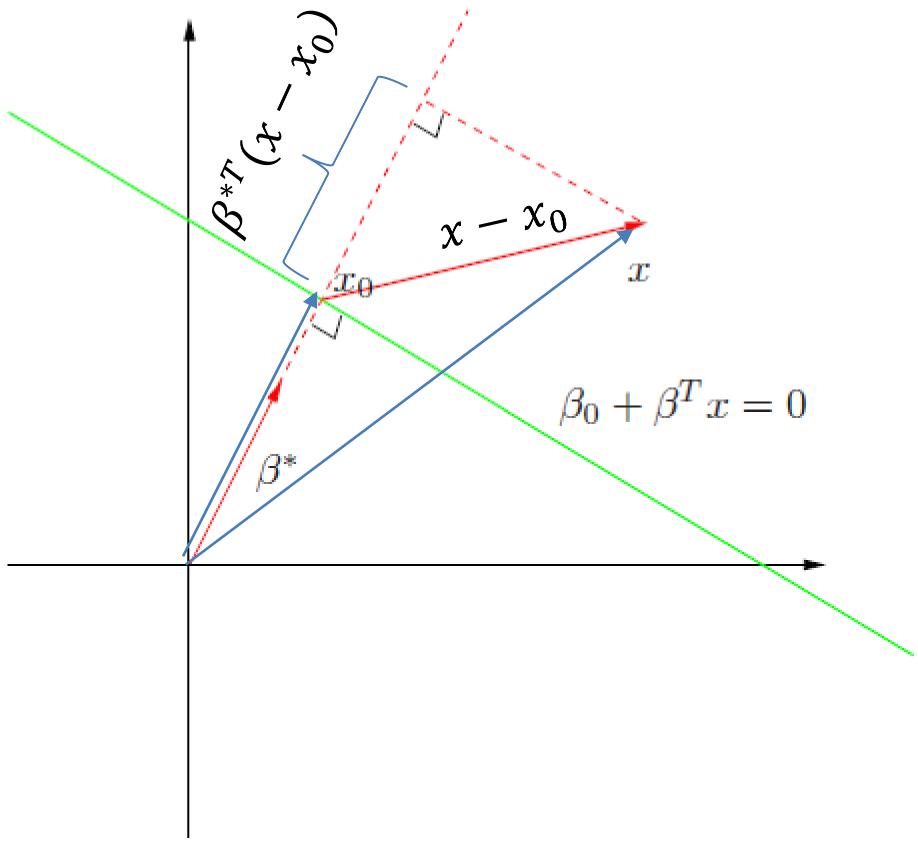
- Perpendicular lines

$$slope_1 \times slope_2 = -1$$

- Example
  - The vector (β_1, β_2) has slope $\frac{\beta_2}{\beta_1}$
  - The line $\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$ has slope $-\frac{\beta_1}{\beta_2}$
  - Are perpendicular to each other

COLUMBIA UNIVERSITY

Geometry (higher dimensions)



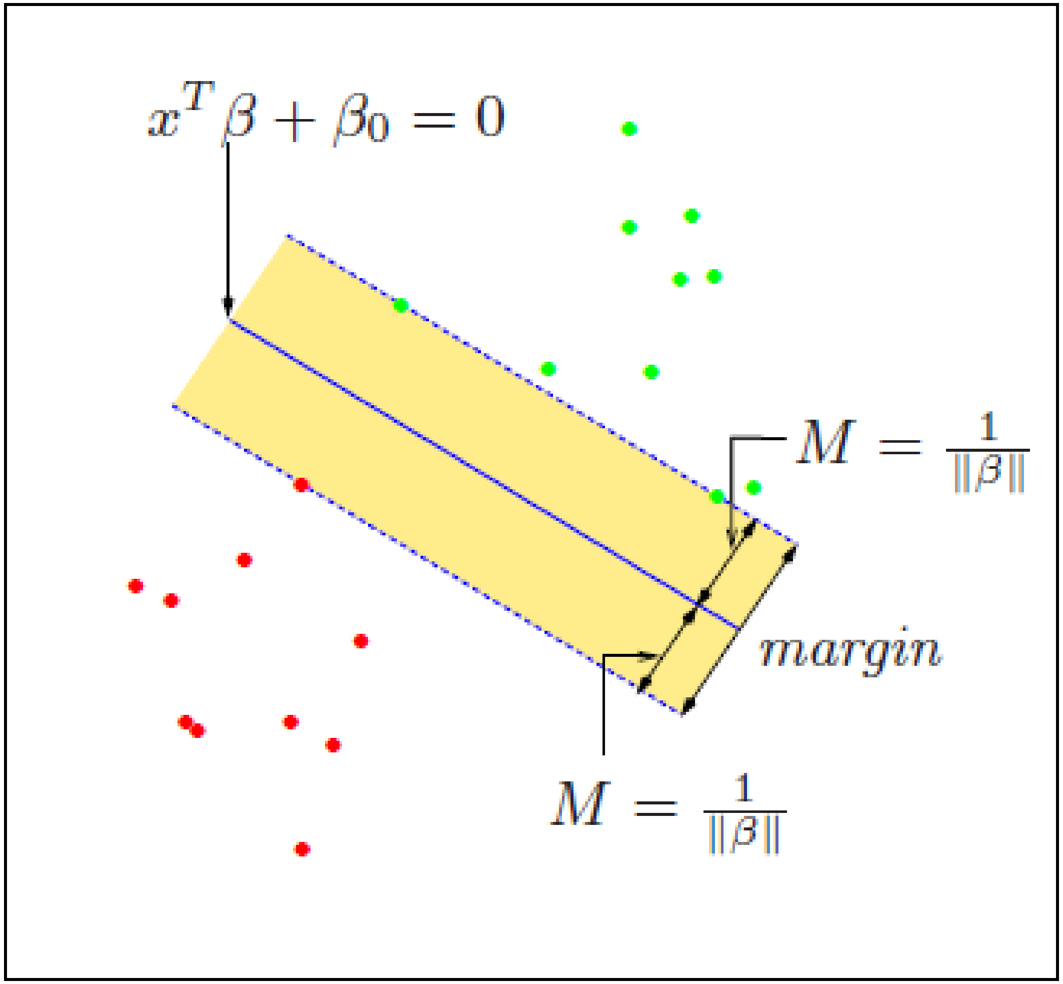
β^* is a unit vector

Geometry (higher dimensions)

- $\boldsymbol{\beta}$ is normal to the hyperplane L defined by
$$\{\boldsymbol{x} | \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} = 0\}$$
- For $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on L,
$$\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$$
- The signed distance from $\boldsymbol{x}$ to L is
$$\left(\frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||}\right)^T (\boldsymbol{x} - \boldsymbol{x}_0) = \frac{1}{||\boldsymbol{\beta}||}(\boldsymbol{\beta}^T\boldsymbol{x} - \boldsymbol{\beta}^T\boldsymbol{x}_0) = \frac{1}{||\boldsymbol{\beta}||}(\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0)$$
- $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \beta_0$ is proportional to the signed distance from $\boldsymbol{x}$ to L

Columbia University

Separable Case



Optimal Separating Hyperplane

- Label y_i indicates where $\boldsymbol{x}_i$ is in relation to L

$$y_i = \begin{cases} +1, & \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} > 0 \\ -1, & \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} < 0 \end{cases}$$

- $\boldsymbol{x}_i$ is at least a distance M from L

$$y_i \frac{1}{\|\boldsymbol{\beta}\|}\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq M \rightarrow \ y_i\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq \|\boldsymbol{\beta}\| M$$

- The optimal separating hyperplane is the maximum margin hyperplane that maximizes M

COLUMBIA UNIVERSITY

Optimization Problem

- Maximizing M is equivalent to minimizing $\|\boldsymbol{\beta}\|$. WLOG, set $\|\boldsymbol{\beta}\| = \frac{1}{M}$

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2$$

subject to

$$y_i\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq 1, \qquad \forall i$$

COLUMBIA UNIVERSITY

Optimization (1)

- Lagrange primal

$$L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) - 1\right]$$

$$= \frac{1}{2}\left(\beta_1^2 + \cdots + \beta_p^2\right) - \sum_{i=1}^{N} \left[\alpha_i y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) - \alpha_i\right]$$

- Set partial derivatives to zero

$$\frac{\partial L_p}{\partial \beta_0} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial L_p}{\partial \beta_{j\neq 0}} = 0 \rightarrow \boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

- Get L_D by substitute these into L_P

COLUMBIA UNIVERSITY

Optimization (2)

- Wolfe dual

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k \boldsymbol{x}_i^T \boldsymbol{x}_k$$

$$\text{subject to} \quad \alpha_i \geq 0, \qquad 0 = \sum_{i=1}^{N} \alpha_i y_i, \qquad \forall i$$

- Solutions satisfy the Karush-Kuhn-Tucker (KKT) conditions

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

$$\alpha_i \geq 0, \forall i$$

$$\alpha_i \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0 \right) - 1 \right] = 0, \forall i$$

COLUMBIA UNIVERSITY

Support Vectors

- The margin around the linear decision boundary has thickness $M = \frac{1}{\|\boldsymbol{\beta}\|}$
- For any $\boldsymbol{x}_i$ more than M away from the boundary, $y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \beta_0\right) > 1$

$$\alpha_i\left[\, y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \beta_0\right) - 1\right] = 0 \rightarrow \alpha_i = 0$$

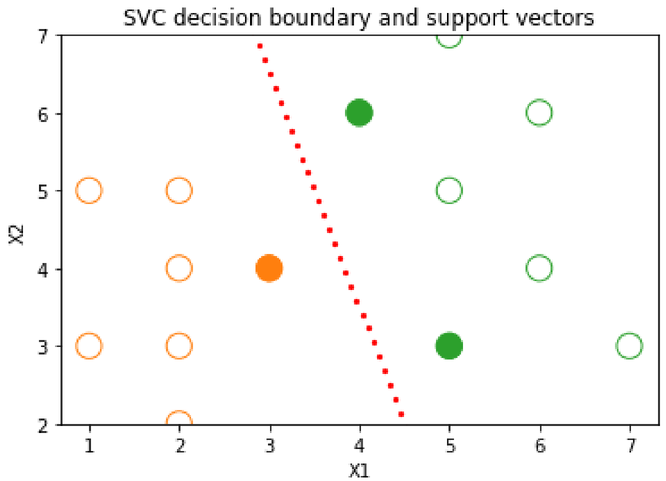
- The support vectors, those on the margin and $\alpha_i > 0$, define the decision boundary
- For SVs, $y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \beta_0\right) - 1 = 0$
- $\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$ is a linear combination of support vectors

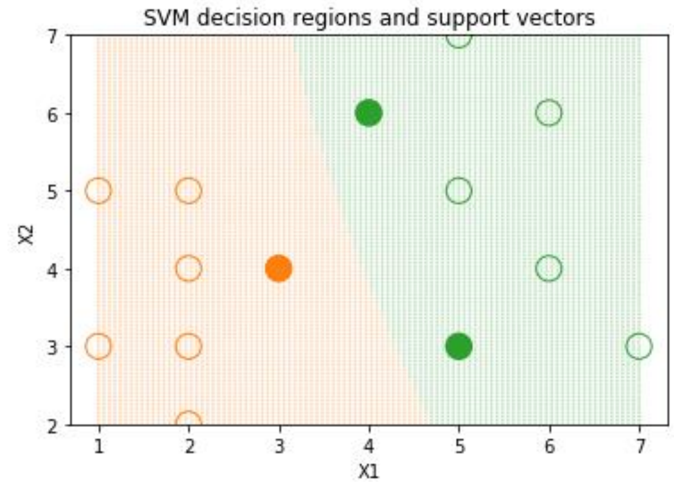
COLUMBIA UNIVERSITY

Maximal Margin Classifier

$$\text{clf}(\boldsymbol{x}) = \text{sign}\left[\boldsymbol{x}^T \widehat{\boldsymbol{\beta}} + \hat{\beta}_0\right]$$

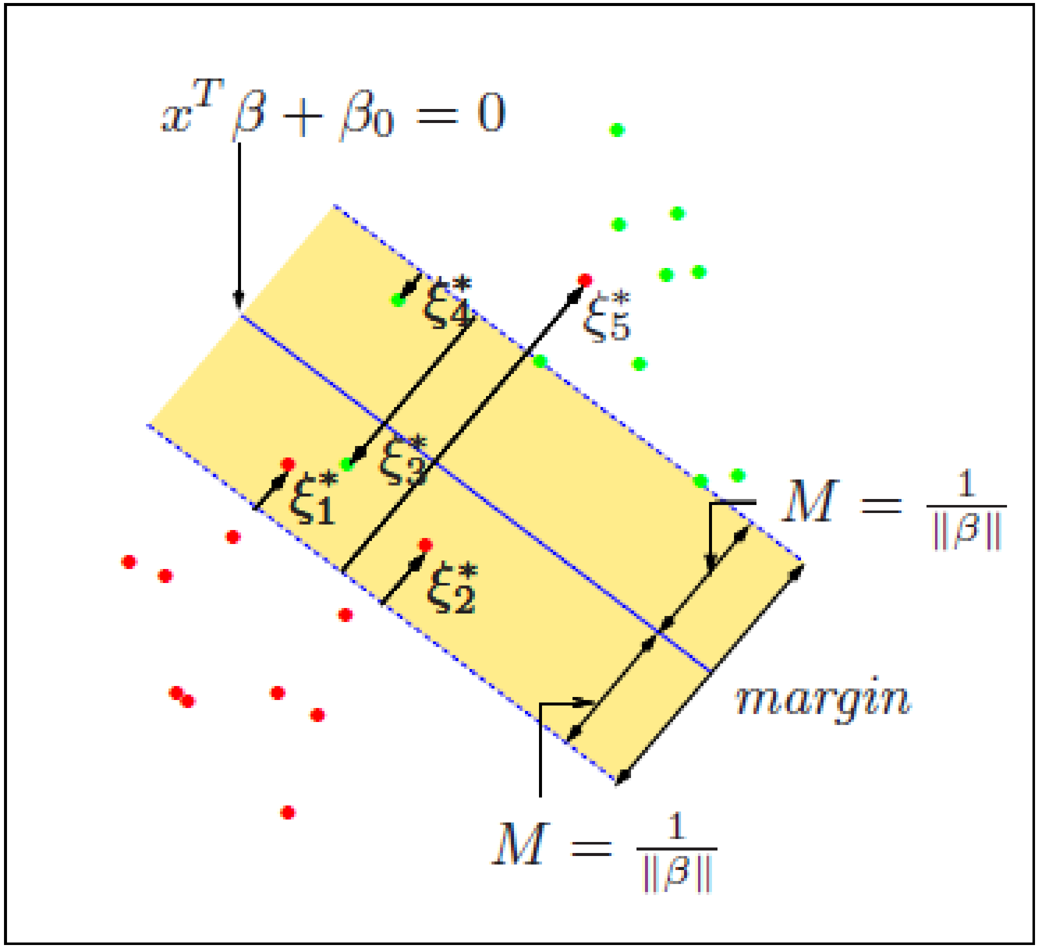
Decision Boundary
separable case





Columbia University

Non-Separable Case



The Support Vector Classifier is the generalization of the Maximal Margin Classifier for the non-separable case.

Optimal Separating Hyperplane

- Label y_i indicates where $\boldsymbol{x}_i$ is in relation to L

$$y_i = \begin{cases} +1, & \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} > 0 \\ -1, & \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} < 0 \end{cases}$$

- $\boldsymbol{x}_i$ is at least a distance M from L, with allowance for some margin violation

$$y_i \frac{1}{\|\boldsymbol{\beta}\|}\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq M(1 - \epsilon_i) \rightarrow$$

$$y_i\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq \|\boldsymbol{\beta}\| M(1 - \epsilon_i)$$

- The slack variable $\epsilon_i \geq 0$, is the proportional amount by which $\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0$ is on the wrong side of its margin
- The decision boundary is one that maximizes M

COLUMBIA UNIVERSITY

Slack Variable

- The slack variable $\epsilon_i \geq 0$, is the proportional amount by which $\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0$ is on the wrong side of its margin

$$\epsilon_i = \begin{cases} = 0, & OK\ wrt\ margin\ and\ L \\ > 0, & violates\ margin, OK\ wrt\ L \\ > 1, & misclassifiction \end{cases}$$

COLUMBIA UNIVERSITY

Optimization Problem

- Maximizing M is equivalent to minimizing $\|\boldsymbol{\beta}\|$. WLOG, set $\|\boldsymbol{\beta}\| = \frac{1}{M}$

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^{N} \epsilon_i$$

subject to

$$y_i(\boldsymbol{x}^T \boldsymbol{\beta} + \beta_0) \geq 1 - \epsilon_i, \qquad \forall i$$
$$\epsilon_i \geq 0$$

- cost parameter C
  - Replaces the constant in the constraint $\sum_{i=1}^{N} \epsilon_i \leq \text{constant}$
  - Penalty for margin violation
  - $C = \infty$ is for the separable case

Columbia University

Optimization

- Lagrange primal

$$L_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N}\epsilon_i - \sum_{i=1}^{N}\alpha_i\left[\, y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \beta_0\right) - (1-\epsilon_i)\right] - \sum_{i=1}^{N}\mu_i\epsilon_i$$

- Wolfe dual

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_k y_i y_k \boldsymbol{x}_i^T\boldsymbol{x}_k$$

- Solutions satisfy

$$\sum_{i=1}^{N}\alpha_i y_i = 0, \qquad \boldsymbol{\beta} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i$$

$$\alpha_i = C - \mu_i,$$

$$\alpha_i\left[y_i\left(\boldsymbol{x}_i^T\beta + \beta_0\right) - (1-\epsilon_i)\right] = 0$$

$$\mu_i\epsilon_i = 0$$

$$y_i\left(\boldsymbol{x}_i^T\beta + \beta_0\right) - (1-\epsilon_i) \geq 0$$

$$\alpha_i, \mu_i, \epsilon_i \geq 0$$

Columbia University

Support Vectors

- Support vectors alone define the decision boundary
- Support vectors are those on or violate the margin
- Support vectors satisfy

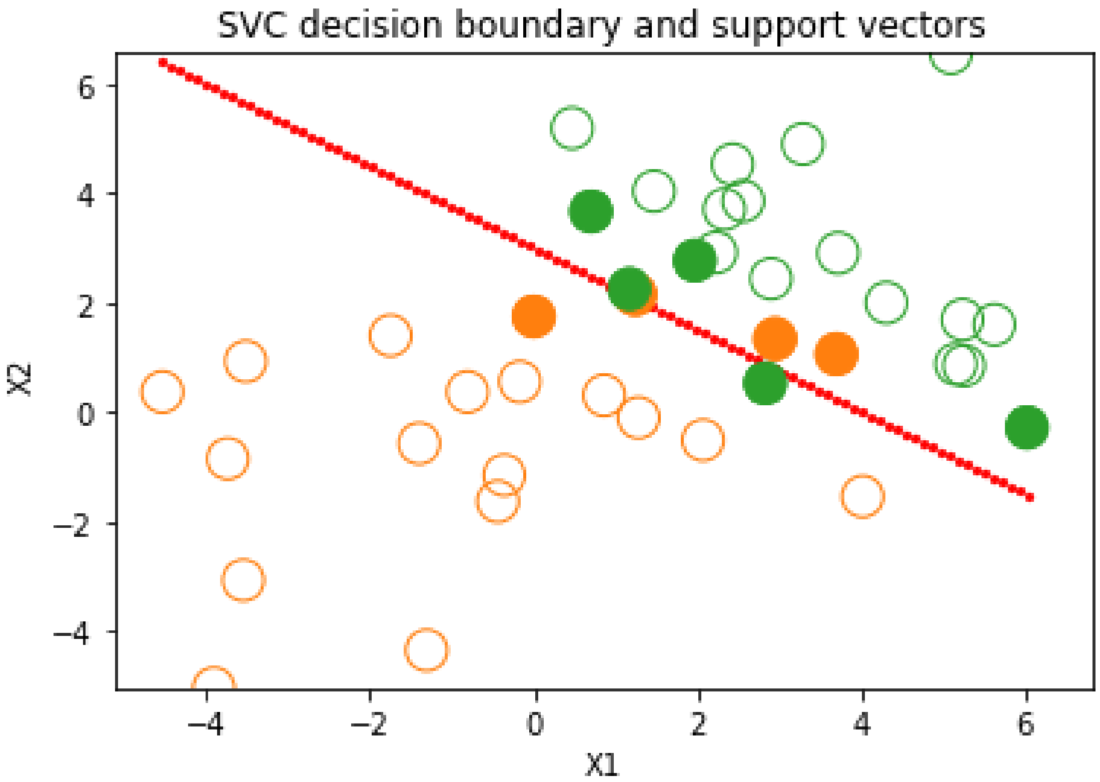
$$y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \beta_0\right) - (1 - \epsilon_i) = 0$$

COLUMBIA UNIVERSITY

Support Vector Classifier

$$\mathrm{clf}(\mathbf{x}) = \mathrm{sign}\left[\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \hat{\beta}_0\right]$$

Decision Boundary (SVC) non-separable case



SVC

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k \boldsymbol{x}_i^T \boldsymbol{x}_k$$

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}} + \hat{\beta}_0 = \sum_{i=1}^{N} \hat{\alpha}_i \, y_i \boldsymbol{x}^T \boldsymbol{x}_i + \hat{\beta}_0$$

$$\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

COLUMBIA UNIVERSITY

SVC with Kernel Function (1)

- Generalize the inner products to kernel functions

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k \boldsymbol{x}_i^T \boldsymbol{x}_k$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k < \boldsymbol{x}_i, \boldsymbol{x}_k >$$

$$\rightarrow L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k < h(\boldsymbol{x}_i), h(\boldsymbol{x}_k) >$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k K(\boldsymbol{x}_i, \boldsymbol{x}_k)$$

- For the linear kernel function, $K(\boldsymbol{x}, \boldsymbol{x}') =< h(\boldsymbol{x}), h(\boldsymbol{x}') > = \boldsymbol{x}^T \boldsymbol{x}'$

COLUMBIA UNIVERSITY

SVC with Kernel Function (2)

- Generalize the inner products to kernel functions

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} + \hat{\beta}_0 = \sum_{i=1}^{N} \hat{\alpha}_i\, y_i \boldsymbol{x}^T \boldsymbol{x}_i + \hat{\beta}_0 = \sum_{i=1}^{N} \hat{\alpha}_i\, y_i < \boldsymbol{x}, \boldsymbol{x}_i > + \hat{\beta}_0$$

$$\rightarrow \hat{f}(\boldsymbol{x}) = \sum_{i=1}^{N} \hat{\alpha}_i\, y_i < h(\boldsymbol{x}), h(\boldsymbol{x}_i) > + \hat{\beta}_0$$

$$= \sum_{i=1}^{N} \hat{\alpha}_i\, y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{\beta}_0$$

- For the linear kernel function, $K(\boldsymbol{x}, \boldsymbol{x}') =< h(\boldsymbol{x}), h(\boldsymbol{x}') > = \boldsymbol{x}^T \boldsymbol{x}'$

COLUMBIA UNIVERSITY

Feature Space Expansion

- A kernel function can expand the feature space.
- Example – from 2D to 6D

$$\boldsymbol{x} = (x_1, x_2)^T \rightarrow h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \dots, h_6(\boldsymbol{x}))^T$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}_k) = (1 + < \boldsymbol{x}_i, \boldsymbol{x}_k >)^2$$

COLUMBIA UNIVERSITY

Example – from 2D to 6D (1)

$$\boldsymbol{x} = (x_1, x_2)^T \rightarrow h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \dots, h_6(\boldsymbol{x}))^T$$

$$\begin{gathered} K(\boldsymbol{x}_i, \boldsymbol{x}_k) = (1 + <\boldsymbol{x}_i, \boldsymbol{x}_k>)^2 \\ = (1 + x_{i1}x_{k1} + x_{i2}x_{k2})^2 \\ = 1 + (x_{i1}x_{k1})^2 + (x_{i2}x_{k2})^2 \\ +2x_{i1}x_{k1} + 2x_{i2}x_{k2} + 2x_{i1}x_{k1}x_{i2}x_{k2} \end{gathered}$$

Columbia University

Example – from 2D to 6D (2)

$$K(\boldsymbol{x}_i, \boldsymbol{x}_k) = 1 + (x_{i1}x_{k1})^2 + (x_{i2}x_{k2})^2$$
$$+2x_{i1}x_{k1} + 2x_{i2}x_{k2} + 2x_{i1}x_{k1}x_{i2}x_{k2} = < h(\boldsymbol{x}_i), h(\boldsymbol{x}_k) >$$

- $h_1(\boldsymbol{x}) = 1 \rightarrow h_1(\boldsymbol{x}_i)h_1(\boldsymbol{x}_k) = 1$
- $h_2(\boldsymbol{x}) = x_1^2 \rightarrow h_2(\boldsymbol{x}_i)h_2(\boldsymbol{x}_k) = (x_{i1}x_{k1})^2$
- $h_3(\boldsymbol{x}) = x_2^2 \rightarrow h_3(\boldsymbol{x}_i)h_3(\boldsymbol{x}_k) = (x_{i2}x_{k2})^2$
- $h_4(\boldsymbol{x}) = \sqrt{2}x_1 \rightarrow h_4(\boldsymbol{x}_i)h_4(\boldsymbol{x}_k) = 2x_{i1}x_{k1}$
- $h_5(\boldsymbol{x}) = \sqrt{2}x_2 \rightarrow h_5(\boldsymbol{x}_i)h_5(\boldsymbol{x}_k) = 2x_{i2}x_{k2}$
- $h_6(\boldsymbol{x}) = \sqrt{2}x_1x_2 \rightarrow h_6(\boldsymbol{x}_i)h_6(\boldsymbol{x}_k) = 2x_{i1}x_{k1}x_{i2}x_{k2}$

COLUMBIA UNIVERSITY

Support Vector Machine

- The support vector machine is an extension of the support vector classifier, expanding the feature space using kernels
- Linear $K(\boldsymbol{x}, \boldsymbol{x}') =< \boldsymbol{x}, \boldsymbol{x}' > = \boldsymbol{x}^T \boldsymbol{x}'$
- Polynomial $K(\boldsymbol{x}, \boldsymbol{x}') = (1 +< \boldsymbol{x}, \boldsymbol{x}' >)^d$
- Radial basis $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$
- Neural Network

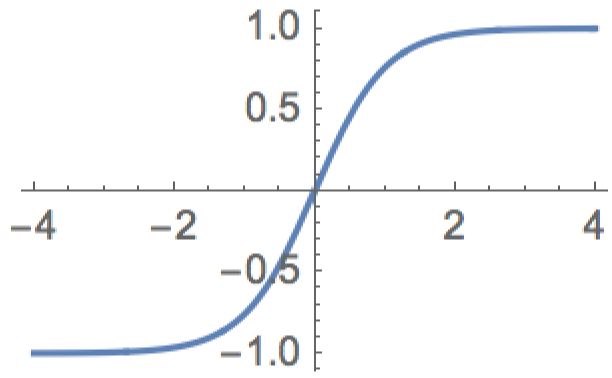
$$K(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\kappa_1 < \boldsymbol{x}, \boldsymbol{x}' > +\kappa_2)$$

COLUMBIA UNIVERSITY

tanh

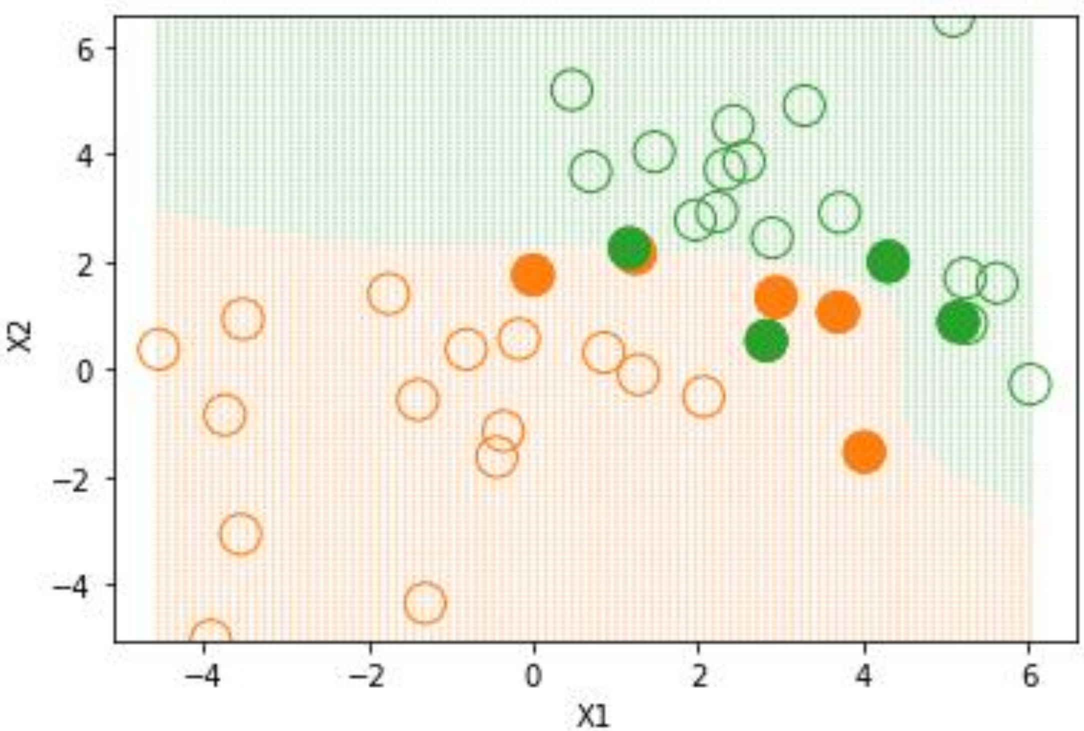
- Neural Network

$$K(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\kappa_1 < \boldsymbol{x}, \boldsymbol{x}' > +\kappa_2)$$



Decision Boundary (SVM) non-separable case

- Poly3, C=1



That was

COLUMBIA UNIVERSITY