Learning From Data
Lecture 5: Support Vector Machines

Yang Li yangli@sz.tsinghua.edu.cn

October 21, 2022

Ask me a question

What do $x^{(i)}, y^{(i)}$ and $\boxed{x_j^{(i)}}$ mean in the Multivariate Bernoulli event model for text classification?

$x^{(i)}$ - ith sample (document)

n: dictionary size

$$\text{dictionary} = \{\overset{1}{\text{"a"}}, \overset{2}{\text{"free"}}, \overset{3}{\text{"gift"}}\}$$

$n = 3.$

$x^{(1)}$: "a gift"

$x^{(2)}$: "free gift"

$x^{(3)}$: "free, free gift"

$$x^{(1)} = \begin{bmatrix} 1 \\ \boxed{0} \\ 1. \end{bmatrix} \leftarrow \begin{cases} 1 \text{ if word 0 is in dictionary} \\ \text{otherwise} \end{cases}$$

$$x^{(2)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

$$x^{(3)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Multinomial model.

$$x^{(1)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad x^{(2)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$x^{(3)} = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}$$

Previously on Learning from Data

Algorithms we learned so far are mostly **probabilistic linear models**:

Type	Examples
Discrimative probablistic model	linear regression, logistic regression, softmax
Generative probablistic model	GDA, naive Bayes

LDA - shared Σ.

- Choice of model affects model performance; *may easily lead to model mismatch*
- Data are often sampled non-uniformly, forming a sparse distribution in high dimensional input space. *leading to ill-posed problems*

Possible solutions: regularization (more in later lectures), sparse kernel methods (today's lecture)

Today's Lecture

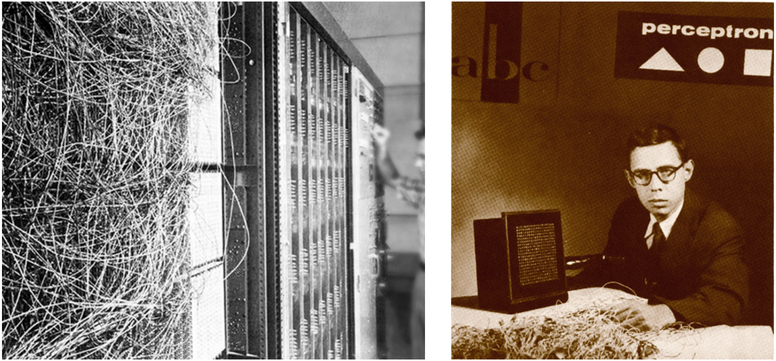
Supervised Learning (Part IV)

- Review: Perceptron Algorithm
- Support Vector Machines (SVM) ← *another discriminative algorithm for learning linear classifiers*
- Kernel SVM ← *non-linear extension of SVM*

Perceptron Learning Algorithm

The perceptron learning algorithm

- Invented in 1956 by Rosenblatt (Cornell University)
- One of the earliest learning algorithm, the first artificial neural network



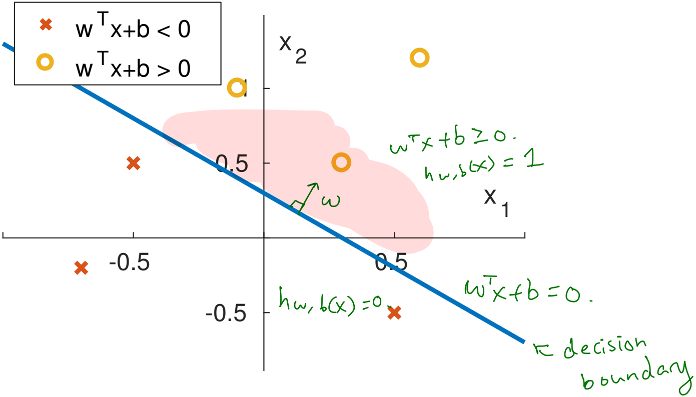
Hardware implementation: Mark I Perceptron

The perceptron learning algorithm

Given x, predict $\underline{y \in \{0,1\}}$

halfplane classifier

$$\underline{h_{w,b}(x)} = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



The perceptron learning algorithm

Perceptron hypothesis function:

$$\begin{bmatrix} w \\ b \end{bmatrix} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix}$$

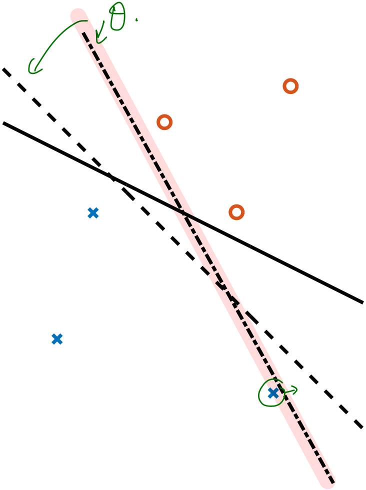
$$h_\theta(x) = \begin{cases} 1 & \text{if } \theta^T x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Parameter update rule:

$$\underline{\theta_j} = \underline{\theta_j} + \alpha \left(\underline{y^{(i)} - h_\theta(x^{(i)})} \right) x_j^{(i)} \text{ for all } j = 0, \ldots, n$$

- When prediction is correct: $\underline{\theta_j = \theta_j}$
- When prediction is incorrect:
    - predicted "1": $\theta_j = \theta_j - \alpha x_j$
    - predicted "0": $\theta_j = \theta_j + \alpha x_j$

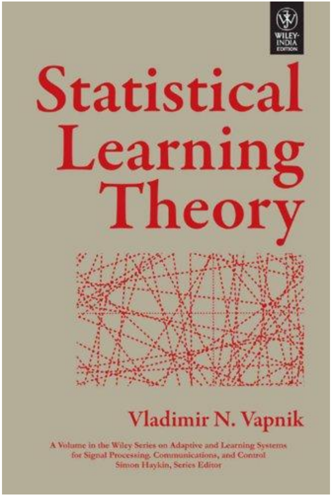
Issues with linear hyperplane perceptron:



- Infinitely many solutions if data are separable
- Can not express “confidence” of the prediction

Support Vector Machines in History

- Theoretical algorithm: developed from Statistical Learning Theory (Vapnik & Chervonenkis) since 60s
- Modern SVM was introduced in COLT 92 by Boser, Guyon & Vapnik



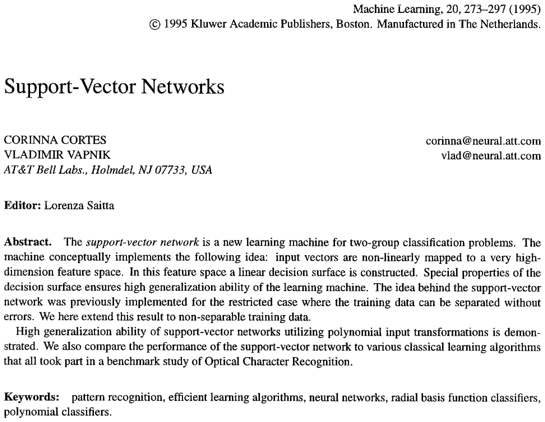
Statistical Learning Theory

Vladimir N. Vapnik

A Volume in the Wiley Series on Adaptive and Learning Systems
for Signal Processing, Communications, and Control
Simon Haykin, Series Editor

Support Vector Machines in History

- 1995 paper by Corte & Vapnik titled "Support-Vector Networks"
- Gained popularity in 90s for giving accuracy comparable to neural networks with elaborated features in a handwriting task



Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

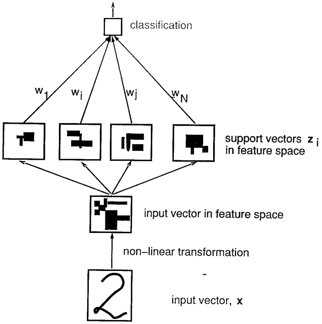
CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

Editor: Lorenza Saitta

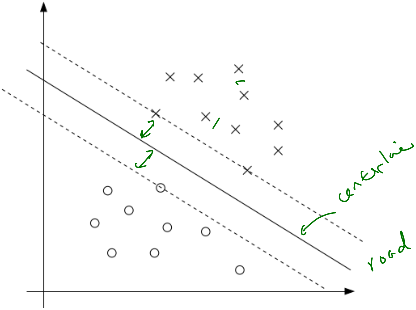
Abstract. The *support-vector network* is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data.

High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.

Keywords: pattern recognition, efficient learning algorithms, neural networks, radial basis function classifiers, polynomial classifiers.

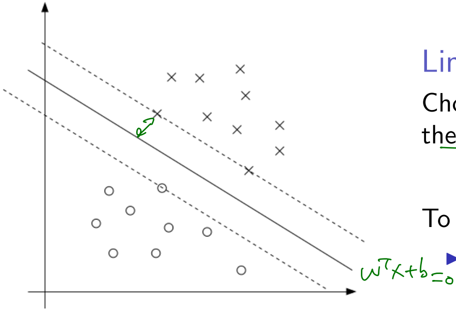


Support Vector Machine: Overview



Margin: smallest distance between the decision boundary to any samples *(Margin also represents classification confidence)*

Support Vector Machine: Overview



Margin: smallest distance between the decision boundary to any samples *(Margin also represents classification confidence)*

Linear SVM

Choose a linear classifier that maximizes the margin.

To be discussed:

- How to measure the margin? (functionally vs geometrically)
- How to find the decision boundary with optimal margin? w^*, b^*

+ a detour on Lagrange Duality

Functional margins

$w^Tx+b=0$

Class labels: $\underline{y \in \{-1, 1\}}$

$$h_{w,b}(x) = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ \underline{-1} & \text{otherwise} \end{cases}$$

Functional margins

Class labels: $y \in \{-1, 1\}$

$sign(x) \in \{-1, 1\}$

$$h_{w,b}(x) = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Given (w, b),

Functional Margin

Given training sample $(x^{(i)}, y^{(i)})$

$h_{w,b}(x)$ if $w^Tx^i+b \geq 0.$

$$\hat{\gamma}^{(i)} = y^{(i)} \left(w^T x^{(i)} + b \right)$$

$w^Tx^i+b >> 0$

if $y^i = 1$, $w^Tx^i+b > 0.$

$sign(\hat{y}^i) = +1$

$\hat{y}^i = 1 \cdot (w^Tx^i+b) > 0$

$w^Tx^i+b < 0$

$sign(\hat{y}^i) = -1$

$sign(\hat{\gamma}^{(i)})$: whether the hypothesis is correct

---

If $y^i = -1$, $w^Tx^i+b > 0$

$sign(\hat{y}^i) = -1$

If $w^Tx^i+b < 0$

$w^Tx^i+b << 0$

$sign(\hat{y}^i) = +1$

Functional margins

Class labels: $y \in \{-1, 1\}$

$$h_{w,b}(x) = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Functional Margin

Given training sample $(x^{(i)}, y^{(i)})$

$$\hat{\gamma}^{(i)} = y^{(i)} \left(w^T x^{(i)} + b \right)$$

$sign(\hat{\gamma}^{(i)})$: whether the hypothesis is correct

- $\hat{\gamma}^{(i)} >> 0$: prediction is correct with high confidence

Functional margins

Class labels: $y \in \{-1, 1\}$

$$h_{w,b}(x) = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Functional Margin

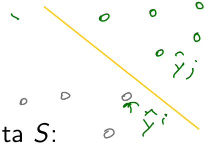
Given training sample $(x^{(i)}, y^{(i)})$

$$\hat{\gamma}^{(i)} = y^{(i)} \left(w^T x^{(i)} + b \right)$$

$sign(\hat{\gamma}^{(i)})$: whether the hypothesis is correct

- $\hat{\gamma}^{(i)} >> 0$: prediction is correct with high confidence
- $\hat{\gamma}^{(i)} << 0$: prediction is incorrect with high confidence

Function Margins



Functional margin of $\underline{(w, b)}$ with respect to training data S:

$$\hat{\gamma} = \min_{i=1,..,m} \hat{\gamma}^{(i)} = \min_{\underline{i=1,..,m}} y^{(i)} \left(\underset{\sim}{w}^T x^{(i)} + \underline{b} \right)$$

Function Margins

Functional margin of (w, b) with respect to training data S:

$$\hat{\gamma} = \min_{i=1,..,m} \hat{\gamma}^{(i)} = \min_{i=1,..,m} y^{(i)} \left(w^T x^{(i)} + b \right)$$

$\underline{w^T x + b} = 0 .$ $\quad c\,\underline{w^T x + cb} = 0.$

Issue: $\underline{\hat{\gamma}}$ depends on $\underline{||w||}$ and $\underline{b}$

e.g. Let $w' = \underline{2}w, b' = \underline{2}b$. The decision boundary parameterized by $(\underline{w', b'})$ and $(\underline{w, b})$ are the same. However,

$$\underline{\hat{\gamma}'^{(i)}} = y^{(i)} \left(\underline{2} w^T x^{(i)} + \underline{2} b \right) = 2y^{(i)}(w^T x^{(i)} + b) = \underline{2\hat{\gamma}^{(i)}}$$

Can we express the margin so that it is invariant to $||w||$ and b?

$w' = cw$, $b' = cb$. $\quad \hat{y} = c\hat{y}$.

Geometric Margins

The **geometric margin** $\gamma^{(i)}$ of a training example $(x^{(i)}, y^{(i)})$ is the distance from the hyperplane:

$$\gamma^{(i)} = y^{(i)} \left(\frac{w}{||w||}^{T} x^{(i)} + \frac{b}{||w||} \right)$$

write in terms of $x^{(i)}$, $\gamma^{(i)}$ and $\vec{\omega}$

$$Q = x^{(i)} - \boxed{\frac{w}{||w||}} \gamma^{(i)}$$

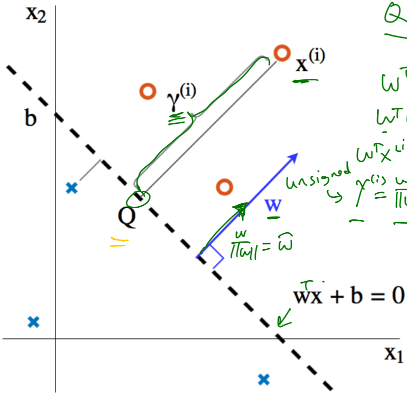
$$w^T Q + b = 0.$$

$$w^T \left(x^{(i)} - \frac{w}{||w||} \gamma^{(i)} \right) + b = 0.$$

$$w^T x^{(i)} - ||w|| \gamma^{(i)} + b = 0$$

unsigned ↳

$$\gamma^{(i)} = \frac{w^T}{||w||} x^i + \frac{b}{||w||}$$



- w is normal to hyperplane
  $w^T x + b = 0$
- We want $\gamma^{(i)} > 0$ when prediction is correct

signed margin:

$$\gamma^{(i)} = y^{(i)} \left(\frac{w^T}{||w||} x^i + \frac{b}{||w||} \right) \in \mathbb{R}.$$

Geometric Margins

The **geometric margin** of (w, b) with respect to training data S is the minimum distance from any point to the hyperplane:

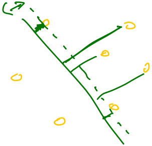
$$\gamma = \min_{i=1,..,m} \gamma^{(i)} = \min_{i=1,..,m} y^{(i)} \left(\frac{w}{||w||}^T x^{(i)} + \frac{b}{||w||} \right)$$

Geometric Margins

The **geometric margin** of (w, b) with respect to training data S is the minimum distance from any point to the hyperplane:

$$\begin{aligned}\underline{\gamma} = \min_{i=1,..,m} \gamma^{(i)} &= \min_{i=1,..,m} y^{(i)} \left(\frac{w}{\boxed{||w||}}^{T} x^{(i)} + \frac{b}{\boxed{||w||}} \right) \\ &= \frac{1}{||w||} \underbrace{\min_{i=1,..,m} y^{(i)} \left(w^{T} x^{(i)} + b \right)}_{\text{functional margin of } x^1, y^1 .. x^m y^m} \\ &= \frac{1}{||w||} \underline{\underline{\hat{\gamma}}}\end{aligned}$$

Geometric Margins



The **geometric margin** of (w, b) with respect to training data S is the minimum distance from any point to the hyperplane:

$$\begin{aligned}\underline{\gamma} = \min_{i=1,..,m} \gamma^{(i)} &= \min_{i=1,..,m} y^{(i)} \left(\frac{w}{||w||}^T x^{(i)} + \frac{b}{||w||} \right) \\ &= \frac{1}{||w||} \min_{i=1,..,m} y^{(i)} \left(w^T x^{(i)} + b \right) \\ &= \frac{1}{\underline{||w||}} \boxed{\hat{\gamma}}\end{aligned}$$

for some $\underline{cw}$, $\underline{cb}$.

$\hat{\gamma} = 1$

- $\hat{\gamma} = \gamma$ when $||w|| = 1$

Geometric Margins

The **geometric margin** of (w, b) with respect to training data S is the minimum distance from any point to the hyperplane:

$$\begin{aligned}
\underline{\gamma} = \min_{i=1,..,m} \gamma^{(i)} &= \min_{i=1,..,m} y^{(i)} \left(\frac{w}{||w||}^T x^{(i)} + \frac{b}{||w||} \right) \\
&= \frac{1}{||w||} \min_{i=1,..,m} y^{(i)} \left(w^T x^{(i)} + b \right) \\
&= \frac{1}{||w||} \hat{\gamma}
\end{aligned}$$

- $\hat{\gamma} = \gamma$ when $||w|| = 1$
- Geometric margins are invariant to parameter scaling

Optimal Margin Classifier

Assume data is linearly separable

Find (w, b) that maximize geometric margin $\gamma = \frac{\hat{\gamma}}{||w||}$ of the training data

$\max_{\gamma,w,b} \gamma$, - geometry

$$\max_{\gamma,w,b} \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \; i = 1, \ldots, m$$

$$\hat{\gamma} = \min_{y} y^{i}(w^{T} x^{i} + b)$$

Optimal Margin Classifier

Assume data is linearly separable

Find (w, b) that maximize geometric margin $\gamma = \frac{\hat{\gamma}}{||w||}$ of the training data

$$\underline{\max_{\gamma,w,b}} \frac{\hat{\gamma}}{||\underline{w}||}$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \ i = 1, \ldots, m$$

There exists some $\delta \in \mathbb{R}$ such that the functional margin of $\overset{(w,b) \ =}{(\underline{\delta} w, \underline{\delta} b)}$ is $\underline{\hat{\gamma} = 1}$

$$= \max_{\gamma,w,b} \quad \gamma \quad \boxed{\frac{1}{||\underline{w}||}} \text{ after scaling}$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 \ i = 1, \ldots, m$$

Optimal Margin Classifier

Assume data is linearly separable

Find (w, b) that maximize geometric margin $\gamma = \frac{\hat{\gamma}}{||w||}$ of the training data

$$\max_{\gamma, w, b} \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \; i = 1, \ldots, m$$

There exists some $\delta \in \mathbb{R}$ such that the functional margin of $(\delta w, \delta b)$ is $\hat{\gamma} = 1$

$$\max_{\underline{\gamma, w, b}} \quad \frac{1}{\boxed{||w||}}$$
$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 \; i = 1, \ldots, m$$
$$\left.\begin{aligned} \iff \min_{\gamma, w, b} \quad & \boxed{\frac{1}{2}}\underline{||w||}^2 \\ & \text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 \; i = 1, \ldots, m \end{aligned}\right\}$$

Optimal Margin Classifier

Assume data is linearly separable

Find (w, b) that maximize geometric margin $\gamma = \frac{\hat{\gamma}}{||w||}$ of the training data

$$\max_{\gamma,w,b} \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma},\ i = 1, \ldots, m$$

There exists some $\delta \in \mathbb{R}$ such that the functional margin of $(\delta w, \delta b)$ is $\hat{\gamma} = 1$

$$\max_{\gamma,w,b} \qquad \frac{1}{||w||}$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 \ i = 1, \ldots, m$$

$$\iff \min_{\underline{\gamma,w,b}} \qquad \frac{1}{2}\underline{||w||^2}$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 \ i = 1, \ldots, m$$

can be solved using QP software

Review: Lagrange Duality

The **primal** optimization problem:

$$\begin{aligned} \min_{w} \quad & \underline{f(w)} \\ s.t. \quad & g_i(w) \leq 0, i, \ldots, k \quad - \text{ k inequality constraints} \\ & h_i(w) = 0, i = 1, \ldots, l \quad - \text{ } \ell \text{ equality constraints} \end{aligned}$$

Review: Lagrange Duality

The **primal** optimization problem:

$$\left.\begin{aligned} \min_{w} \quad & f(w) \\ s.t. \quad & \underline{g_i(w) \leq 0}, i, \ldots, k \\ & h_i(w) = 0, i = 1, \ldots, l \end{aligned}\right\} \text{standard form constrained optimization}$$

Define the **generalized Lagrange function** :

$$L(w, \underline{\underline{\alpha}}, \underline{\underline{\beta}}) = \underline{f(w)} + \sum_{i=1}^{k} \alpha_i \underline{g_i(w)} + \sum_{i=1}^{l} \beta_i \underline{h_i(w)}$$

α_i and β_i are called the **Lagrange multipliers**

For a given $\underline{\underline{w}}$,

$$\begin{aligned}\underline{\theta_P(w)} &= \max_{\alpha,\beta:\underline{\alpha_i\geq 0}} \underline{L(w,\alpha,\beta)} \\ &= \max_{\alpha,\beta:\alpha_i\geq 0} \underline{f(w)} + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)\end{aligned}$$

For a given w,

$$\theta_P(w) = \max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta)$$

$$= \max_{\alpha,\beta:\alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \underbrace{\alpha_i g_i(w)}_{\leq 0} + \sum_{i=1}^{l} \underbrace{\beta_i \underbrace{h_i(w)}_{=0}}_{=0}$$

for $i=1,\dots k$.

Recall the primal constraints: $g_i(w) \leq 0$ and $h_i(w) = 0$: $i=1,\dots,l$.

- $\theta_P(w) = f(w)$ if w satisfies primal constraints

$\theta_P(w) = \infty$ if $g_i(w) > 0$. or. $h_i(w) \neq 0$.

For a given w,

$$\begin{aligned}\theta_P(w) &= \max_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta) \\ &= \max_{\alpha,\beta:\alpha_i\geq 0} f(w) + \sum_{i=1}^{k}\alpha_i g_i(w) + \sum_{i=1}^{l}\beta_i h_i(w)\end{aligned}$$

Recall the primal constraints: $g_i(w) \leq 0$ and $h_i(w) = 0$:

- $\theta_P(w) = \underline{f(w)}$ if w satisfies primal constraints
- $\theta_P(w) = \underline{\infty}$ otherwise

$$\min_w f(w)$$

$$\text{s.t. } g_i(w) \leq 0 \ \forall i=1..k$$

$$h_i(w) = 0 \ \forall i=1,..,l$$

The primal problem (alternative form)

$$\underline{\min_w \theta_P(w)} = \min_w \left(\underline{\max_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta)} \right)$$

The primal problem (P)

$$\underline{p^*} = \min_w \theta_P(w) = \underline{\min}_w \underline{\max}_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta)$$

The dual problem (D)

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \underline{\theta_D(\alpha,\beta)} = \boxed{\max_{\alpha,\beta:\underline{\alpha_i\geq 0}}} \left(\min_{\underline{w}} L(w,\alpha,\beta) \right)$$

The primal problem (P)

$$p^* = \min_w \theta_P(w) = \min_w \max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta)$$

The dual problem (D)

$$d^* = \max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \min_w L(w,\alpha,\beta)$$

In general, $\underline{d^* \leq p^*}$ *(max-min inequality)*

The primal problem (P)

$$p^* = \min_w \theta_P(w) = \min_w \max_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta)$$

The dual problem (D)

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0} \min_w L(w,\alpha,\beta)$$

In general, $d^* \leq p^*$ (max-min inequality)

Theorem (Lagrange Duality)

Suppose f and all g_i's are convex, all h_i's are affine, and there exists some w such that $g_i(w) < 0$ for all i (strictly feasible) *.*
There must exists w^*, α^*, β^* so that w^* is the solution to P and α^*,β^* are the solution to D, and

$$p^* = d^* = L(w^*, \alpha^*, \beta^*)$$

Karush-Kuhn-Tucker (KKT) conditions

Under previous conditions, $\underline{w^*, \alpha^*, \beta^*}$ are solutions of P and D **if and only if** they statisty the following conditions:

$$\left.\begin{aligned} \frac{\delta}{\delta w_i} \underline{L(w^*, \alpha^*, \beta^*)} &= 0, \ i = 1, \dots n && (1) \\ \frac{\delta}{\delta \beta_i} L(w^*, \alpha^*, \beta^*) &= 0, \ i = 1, \dots l && (2) \end{aligned}\right\} \text{stationary}$$

$$\left.\underline{\alpha_i^* g_i(w^*) = 0}, \ i = 1, \dots, k \right\} \text{c.s.} \qquad (3)$$

$$\left.\begin{aligned} g_i(w^*) &\leq 0, \ i = 1, \dots, k && \text{- primal feasibility} \quad (4) \\ \alpha^* &\geq 0, \ i = 1, \dots, k && \text{- dual feasibility constraints} \quad (5) \end{aligned}\right.$$

Equation 3 is called the **complementary slackness condition**.

$$\left.\begin{aligned} &\alpha_i^* \geq 0. \\ &g_i(w^*) \leq 0 \end{aligned}\right\} \alpha_i > 0. \Rightarrow g_i(w^*) = 0.^{(\text{C.S.})}$$

Optimal Margin Classifier

$$\min_{w} f(w) \quad \text{s.t. } g_i(w) \le 0$$

Optimal margin classifier

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2$$

$$\text{s.t. } \underline{y^{(i)}(w^T x^{(i)} + b) \geq 1} \; \underline{i = 1, \ldots, m}$$

$$y^i(w^T x^i + b) - 1 \ge 0 \Rightarrow \underbrace{-y^i(w^T x^i + b) + 1}_{g_i(w)} \le 0.$$

- $\underline{f(w)} = \underline{\frac{1}{2}||w||^2}$
- $\underline{g_i(w)} = \underline{-\big(y^{(i)}(w^T x^{(i)} + b) - 1}\big)$

Generalized Lagrangian function:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 \underline{- \sum_{i=1}^{m} \underline{\alpha_i} \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]}$$

By the complementary slackness condition in KKT:

$$\underline{\alpha_i^*} g_i(\underline{w^*}) = 0, \ i = 1, \ldots, k$$

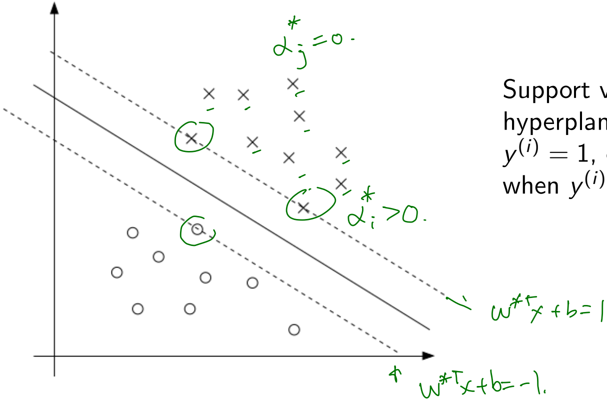
$$\underline{\alpha_i^* > 0} \iff \underline{g_i(w^*)} = -y^{(i)}({w^*}^T x^{(i)} + b) + 1 = \underline{0}$$

By the complementary slackness condition in KKT:

$$\alpha_i^* g_i(w^*) = 0,\ \ i = 1, \ldots, k$$

$$\underline{\alpha_i^* > 0} \iff g_i(w^*) = \underline{-y^{(i)}(w^{*T}x^{(i)} + b) + 1} = 0$$

Training examples $\underline{(x^{(i)}, y^{(i)})}$ such that $\underline{y^{(i)}(w^{*T}x^{(i)} + b)} = 1$ are called **support vectors**



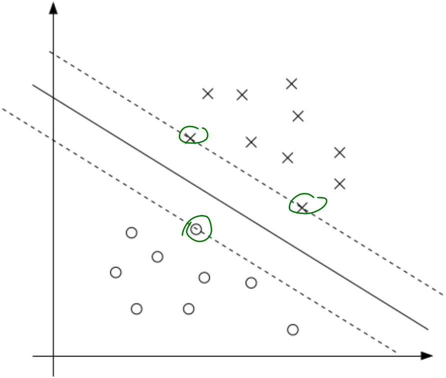
Support vectors lie on hyperplane $w^{*T}x + b = 1$ when $y^{(i)} = 1$, or $w^{*T}x + b = -1$ when $y^{(i)} = -1$

By the complementary slackness condition in KKT:

$$\alpha_i^* g_i(w^*) = 0, \; i = 1, \ldots, k$$

$$\alpha_i^* > 0 \iff g_i(w^*) = -y^{(i)}({w^*}^T x^{(i)} + b) + 1 = 0$$

Training examples $(x^{(i)}, y^{(i)})$ such that $y^{(i)}({w^*}^T x^{(i)} + b) = 1$ are called **support vectors**



Support vectors lie on hyperplane ${w^*}^T x + b = 1$ when $y^{(i)} = 1$, or ${w^*}^T x + b = -1$ when $y^{(i)} = -1$

Constraints $\underline{g_i(w) \leq 0}$ is only **active** on support vectors

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i\left(y^i(w^Tx^i+b)-1\right)$$

Dual optimization problem: *(Check derivation)*

✓

$$\max_{\alpha}\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\underline{\alpha_i}\alpha_j\langle x^{(i)},x^{(j)}\rangle$$

$$s.t.\ \alpha_i\geq 0, i=1,\ldots,m$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

By the KKT condition,

① $\frac{\partial L}{\partial w}=0. \Rightarrow w-\sum_{i=1}^{m}\alpha_i y^i x^i=0. \Rightarrow \underline{w}=\sum_{i=1}^{m}\alpha_i y^i x^i$

② $\frac{\partial L}{\partial b}=0 \Rightarrow \sum_{i=1}^{m}\alpha_i y^i=0.$

Dual optimization problem: *(Check derivation)*

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i\left(y^i(w^Tx^i+b)-1\right)$$

$$\max_{\alpha} \underline{W(\alpha)} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$

$$s.t.\ \underline{\alpha_i \geq 0}, i = 1, \ldots, m$$

by (2) : $$\sum_{i=1}^{m} \underline{\alpha_i y^{(i)} = 0}$$

$\rightarrow$ ① $w^* = \sum_{i=1}^{m} \alpha_i^* y^i x^i$

② $\sum_{i=1}^{m} \alpha^{*i} y^i = 0.$

Given optimal solutions of $\alpha_1, \ldots, \alpha_b$, how to find $\underline{w^*}$ and $\underline{b^*}$?

$$L(w,b,\alpha)=\frac{1}{2}w^Tw-\sum_{i=1}^{m}(\alpha_i y^i w^T x^i+\alpha_i y^i b-\alpha_i)$$

Subsitute w by (1).

$$=\underline{\frac{1}{2}w^T\left(\sum_{i=1}^{m}\alpha_i y^i x^i\right)}-\underline{\sum_{i=1}^{m}\alpha_i y^i \underline{w}^T x^i}-\underbrace{\sum_{i=1}^{m}\alpha_i y^i \underline{b}}_{=0.}+\sum_{i=1}^{m}\alpha_i$$

By (2), $\sum_{i=1}^{m}\alpha_i y^i = 0.$

$$=-\frac{1}{2}\underline{w}^T\left[\sum_{i=1}^{m}\alpha_i y^i x^i\right]+\sum_{i=1}^{m}\alpha_i$$

by (1) $$=-\frac{1}{2}\underbrace{\left(\sum_{i=1}^{m}\alpha_i y^i x^i\right)^T\left(\sum_{i=1}^{m}\alpha_i y^i x^i\right)}+\sum_{i=1}^{m}\alpha_i$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y^i y^j \alpha_i \alpha_j \underbrace{x^{iT}x^j}_{\Rightarrow\langle x^i,x^j\rangle} \quad \triangleq W(\alpha)$$

Solution to the primal problem:

$$w^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} x^{(i)}$$

Find b^*:

① Suppose $y^i = 1$.

since $g_i(w) \le 0 \Rightarrow$ $\underline{w^T x^i + b \ge 1}$

if $\alpha_i > 0$, by the C.S. $w^T x^i + b = 1$

This implies $\min_{x^i, y^i = 1} w^T x^i + b = 1$ (1)

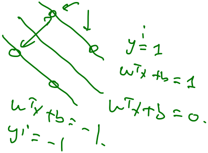
② Suppose $y^i = -1$.

$g_i(w) \le 0 \Rightarrow$ $w^T x^i + b \le -1$.

if $\alpha_i > 0$, $w^T x^i + b = -1$ for some x^i

$\left(\max_{x^i, y^i = -1} w^T x^i + b \right) = -1$. (2)

add (1), (2) $\min_{x^i, y^i = 1} w^T x^i + \underline{b} + \max_{x^i, y^i = -1} w^T x^i + \underline{b} = 0$.



$\underline{b} = -\frac{1}{2}\left(\min_{y^i = 1} w^T x^i + \max_{y^i = -1} w^T x^i \right)$

Solution to the primal problem:

$$\underline{w}^* = \sum_{i=1}^{m} \underline{\alpha_i^*} y^{(i)} x^{(i)}$$

$$\underline{b}^* = -\frac{1}{2}\left(\max_{i:y^{(i)}=-1} {w^*}^T x^{(i)} + \min_{i:y^{(i)}=1} {w^*}^T x^{(i)}\right)$$

$h_{w,b}$

For a new sample z, the SVM prediction is sign $\left[{w^*}^T z + b\right]$

$$w^T z + b = \sum_{i=1}^{m} \alpha_i y^{(i)} \boxed{\langle x^{(i)}, z\rangle} + b$$

$$w^* = \sum_{i=1}^{m} \alpha_i y^i x^i$$

Linear SVM Summary

- Input:: m training samples $(x^{(i)}, y^{(i)}), y^i \in \{-1, 1\}$
- Output: optimal parameters w^*, b^*
- Step 1: solve the dual optimization problem

$$\alpha^* = \max_{\alpha} W(\alpha)$$

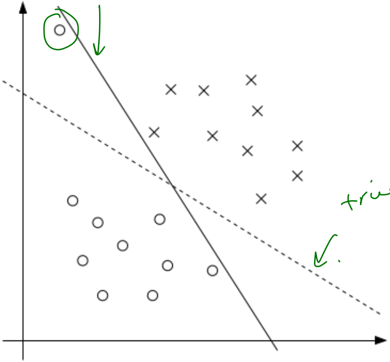
$$s.t.\ \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, i = 1, \ldots, m$$

- Step 2: compute the optimal parameters w^*, b^*

$$w^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} x^{(i)}$$

$$b^* = -\frac{1}{2} \left(\max_{i:y^{(i)}=-1} {w^*}^T x^{(i)} + \min_{i:y^{(i)}=1} {w^*}^T x^{(i)} \right)$$

Limitations of the basic SVM

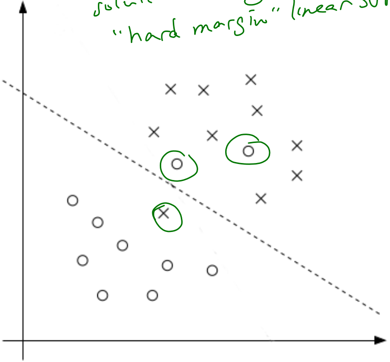


outlier

less robust.

true ✓.

Outliers



can not find feasible solution using "hard margin" linear SVM.

Non-linearly separable cases

Soft Margin SVM

L1 regularization

Functional margin $1-\xi_i \le 1$:

f(w)

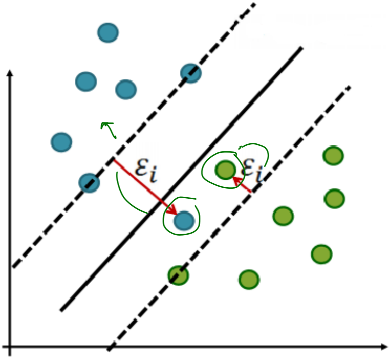
$$\min_{\underline{w,b,\xi}} \frac{1}{2}||w||^2 + \boxed{C}\underbrace{\sum_{i=1}^{m} \xi_i}_{} = \|\varepsilon\|_1$$

$$s.t.\ \underline{y^{(i)}(w^T x^{(i)} + b) \ge} \boxed{1-\xi_i}$$

g(w)

$$\underline{\xi_i \ge 0}, i = 1, \ldots, m$$

- C: <u>relative weight</u> on the regularizer
- L_1 regularization let most $\xi_i = 0$, such that their functional margins $1-\xi_i = 1$



Soft Margin SVM

The generalized Lagrangian function:

$$L(w, b, \xi, \alpha, r) = \underbrace{\frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i}_{f(\omega)} - \sum_{i}^{m}\alpha_i\left[y^{(i)}(w^T x^{(i)} + b) - \underline{1 + \xi_i}\right]$$

$$\underline{-\sum_{i=1}^{m} r_i\xi_i}$$

Soft Margin SVM

The generalized Lagrangian function:

$$\underline{L(w,b,\xi,\alpha,r)} = \frac{1}{2}||w||^2 + \underline{C}\sum_{i=1}^{m} \underline{\xi_i} - \sum_{i}^{m} \underline{\alpha_i}\left[y^{(i)}(w^Tx^{(i)}+b) - 1 + \underline{\xi_i}\right] - \sum_{i=1}^{m} \underline{\gamma_i}\xi_i$$

By the KKT condition

(1) $\frac{\partial L}{\partial w} = 0 \Rightarrow w^* = \sum_{i=1}^{m} \alpha_i y^i x^i$

(2) $\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0.$

(3) $\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \gamma_i = 0$ for all i

Dual problem:

$$L(w,b,\xi,\alpha,\gamma) = \underbrace{\frac{1}{2}||w||_2^2 - \sum_{i=1}^{m}\alpha_i\left(y^i(w^Tx^i+b) - 1\right)}_{\text{by (1), (2), we have}} - \sum_{i=1}^{m}\xi_i\underbrace{(C-\alpha_i-\gamma_i)}_{\text{By (3)} = 0.}$$

$$L(w,b,\xi,\alpha,\gamma) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j x^{iT} x^j$$

since $C - \alpha_i - \gamma_i = 0$, $\gamma_i = C - \alpha_i$

since $\underbrace{\alpha_i \geq 0,\ \gamma_i \geq 0}_{\text{Lagrange multiples}} \Rightarrow \alpha_i \leq C.$

$$0 \leq \boxed{\alpha_i \leq C}$$

Soft Margin SVM

The generalized Lagrangian function:

$$L(w, b, \xi, \alpha, r) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i}^{m}\alpha_i\left[y^{(i)}(w^Tx^{(i)} + b) - 1 + \xi_i\right] - \sum_{i=1}^{m} r_i\xi_i$$

Dual problem:

$$\max_{\alpha}\ W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$

$$s.t.\ 0 \leq \alpha_i \leq C, i = 1, \dots, m$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

w^* is the same as the non-regularizing case, but b^* has changed.

Soft Margin SVM

Dual problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t.\ \underline{0 \le \alpha_i \le C}, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

By the KKT dual-complentary conditions, for all i, $\alpha_i^* g_i(w^*) = 0$

$$\begin{aligned} &\alpha_i = 0 && \Longleftrightarrow \\ &\alpha_i = C && \Longleftrightarrow \\ &0 < \alpha_i < C && \Longleftrightarrow \end{aligned}$$

Soft Margin SVM

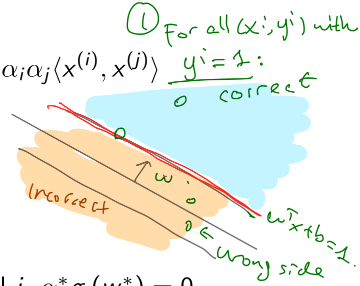
The value of $w^* = \sum_{i=1}^{m} \alpha_i x^i y^i$ only depends on support vectors $(\alpha_i \neq 0)$!

Dual problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$s.t.\ 0 \le \alpha_i \le C, i = 1, \dots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$



By the KKT dual-complentary conditions, for all i, $\alpha_i^* g_i(w^*) = 0$

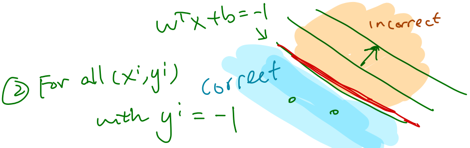
(x^i, y^i)

$\alpha_i = 0$ → not support vector $\iff$ $y^{(i)}(w^T x^{(i)} + b) \ge 1$ correct side of margin

$\alpha_i = C$ $\iff$ $y^{(i)}(w^T x^{(i)} + b) \le 1$ wrong side of margin

$0 < \alpha_i < C$ $\iff$ $y^{(i)}(w^T x^{(i)} + b) = 1$ at margin

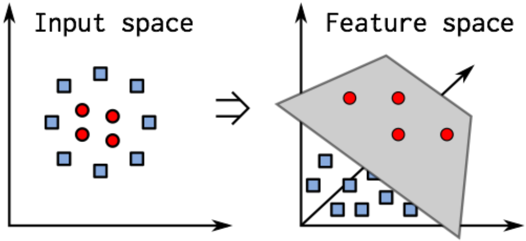
$\alpha_i \neq 0$ support vectors



Kernel SVM

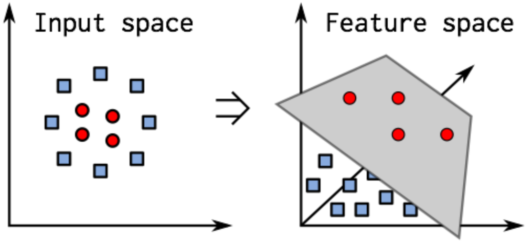
Non-linear SVM

For non-separable data, we can use the **kernel trick**: Map input values $x \in \mathbb{R}^d$ to a higher dimension $\phi(x) \in \mathbb{R}^D$, such that the data becomes separable.



Non-linear SVM

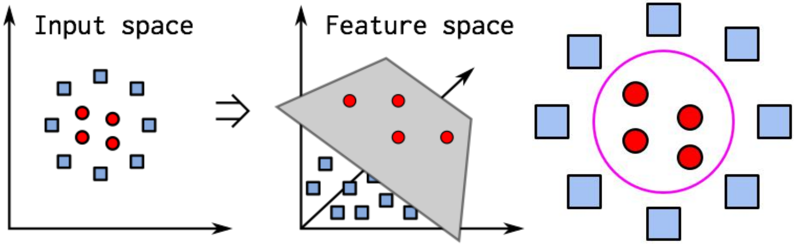
For non-separable data, we can use the **kernel trick**: Map input values $x \in \mathbb{R}^d$ to a higher dimension $\phi(x) \in \mathbb{R}^D$, such that the data becomes separable.



- ϕ is called a **feature mapping**.

Non-linear SVM

For non-separable data, we can use the **kernel trick**: Map input values $x \in \mathbb{R}^d$ to a higher dimension $\phi(x) \in \mathbb{R}^D$, such that the data becomes separable.



- ϕ is called a **feature mapping**.
- The classification function $w^T x + b$ becomes nonlinear:$w^T \phi(x) + b$

Kernel Function

Given a feature mapping ϕ, we define the **kernel function** to be

$$K(x, z) = \phi(x)^T \phi(z)$$

Kernel Function

Given a feature mapping ϕ, we define the **kernel function** to be

$$K(x, z) = \phi(x)^T \phi(z)$$

Some kernel functions are easier to compute than $\phi(x)$, e.g.

$$K(x, z) = (x^T z)^2$$

Kernel Function

Given a feature mapping ϕ, we define the **kernel function** to be

$$K(x, z) = \phi(x)^T \phi(z)$$

Some kernel functions are easier to compute than $\phi(x)$, e.g.

$$\begin{aligned} K(x, z) = (x^T z)^2 &= \left(\sum_{i=1}^{n} x_i z_i \right) \left(\sum_{j=1}^{n} x_j z_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j z_i z_j \\ &= \phi(x)^T \phi(z) \end{aligned}$$

Kernel Function

Given a feature mapping ϕ, we define the **kernel function** to be

$$K(x, z) = \phi(x)^T \phi(z)$$

Some kernel functions are easier to compute than $\phi(x)$, e.g.

$$\begin{aligned} K(x, z) = (x^T z)^2 &= \left(\sum_{i=1}^{n} x_i z_i \right) \left(\sum_{j=1}^{n} x_j z_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j z_i z_j \\ &= \phi(x)^T \phi(z) \end{aligned}$$

where $\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ \vdots \\ x_n x_{n-1} \\ x_n x_n \end{bmatrix}$ takes $O(n^2)$ operations to compute, while $(x^T z)^2$ only takes $O(n)$

Kernel SVM

In the dual problem, replace $\langle x_i, y_j \rangle$ with $\langle \phi(x_i), \phi(y_i) \rangle = K(x_i, x_j)$

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x_i, x_j)$$

$$s.t.\ 0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Kernel SVM

In the dual problem, replace $\langle x_i, y_j \rangle$ with $\langle \phi(x_i), \phi(y_i) \rangle = K(x_i, x_j)$

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x_i, x_j)$$

$$s.t.\ 0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

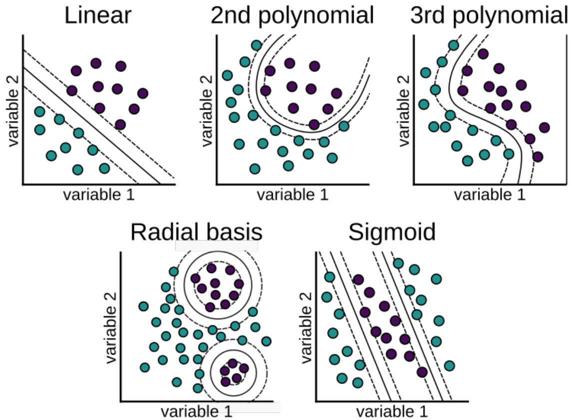
No need to compute $w^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} \phi(x^{(i)})$ explicitly since

$$\begin{aligned} f(x) = w^T \phi(x) + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \phi(x^{(i)}) \right)^T \phi(x) + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle \phi(x^{(i)}), \phi(x) \rangle + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b \end{aligned}$$

Kernel Matrix

kernel functions measure the similarity between samples x, z, e.g.

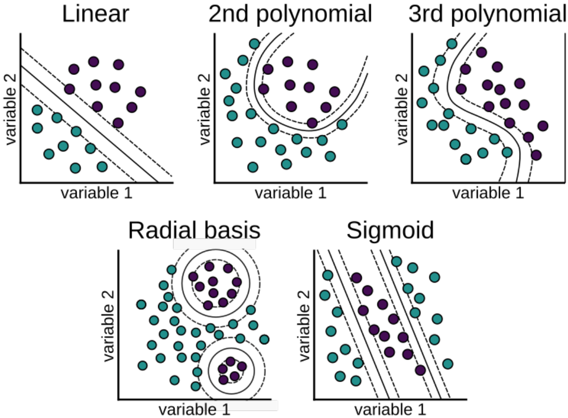
- Linear kernel: $K(x,z) = (x^T z)$
- Polynomial kernel: $K(x,z) = (x^T z + 1)^p$
- Gaussian / radial basis function (RBF) kernel: $K(x,z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$



Kernel Matrix

kernel functions measure the similarity between samples x, z, e.g.

- Linear kernel: $K(x, z) = (x^T z)$
- Polynomial kernel: $K(x, z) = (x^T z + 1)^p$
- Gaussian / radial basis function (RBF) kernel: $K(x, z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$



Can any function $K(x, y)$ be a kernel function?

Kernel Matrix

Represent kernel function as a matrix $K \in \mathbb{R}^{m \times m}$ where $K_{i,j} = K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.

Kernel Matrix

Represent kernel function as a matrix $K \in \mathbb{R}^{m \times m}$ where $K_{i,j} = K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$.

Theorem (Mercer)

Let $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ Then K is a valid (Mercer) kernel if and only if for any finite training set $\{x^{(i)}, \ldots, x^{(m)}\}$, K is symmetric positive semi-definite.

i.e. $K_{i,j} = K_{j,i}$ and $x^T K x \geq 0$ for all $x \in \mathbb{R}^n$

Kernel SVM Summary

- Input: m training samples $(x^{(i)}, y^{(i)}), y^i \in \{-1, 1\}$, kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, constant $C > 0$
- Output: non-linear decision function $f(x)$
- Step 1: solve the dual optimization problem for α^*

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)})$$

$$s.t.\ 0 \leq \alpha_i \leq C, \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, i = 1, \ldots, m$$

- Step 2: compute the optimal decision function

$$b^* = y^{(j)} - \sum_{i=1}^{m} \alpha_i^* y^{(i)} K(x^{(i)}, x^{(j)}) \text{ for some } 0 \leq \alpha_j \leq C$$

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b^*$$

In practice, it's more efficient to compute kernel matrix K in advance.

SVM in Practice

Sequential Minimal Optimization: a fast algorithm for training soft margin kernel SVM

- Break a large SVM problem into smaller chunks, update two α_i's at a time
- Implemented by most SVM libraries.

SVM in Practice

Sequential Minimal Optimization: a fast algorithm for training soft margin kernel SVM

- Break a large SVM problem into smaller chunks, update two α_i's at a time
- Implemented by most SVM libraries.

Other related algorithms

- Support Vector Regression (SVR)
- Multi-class SVM (Koby Crammer and Yoram Singer. 2002. *On the algorithmic implementation of multiclass kernel-based vector machines*. J. Mach. Learn. Res. 2 (March 2002), 265-292.)