Learning From Data
Lecture 5: Support Vector Machines

Yang Li yangli@sz.tsinghua.edu.cn

October 16, 2020

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

- 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

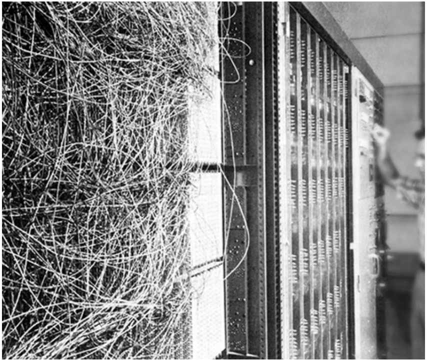
Supervised Learning (Part IV)

- Review: Perceptron Algorithm
- Support Vector Machines (SVM) $\leftarrow$ *another discriminative algorithm for learning linear classifiers*
- Kernel SVM $\leftarrow$ *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

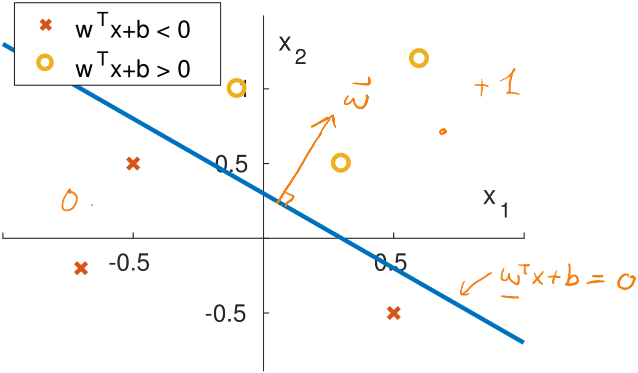


Hardware implementation: Mark I Perceptron

The perceptron learning algorithm

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

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



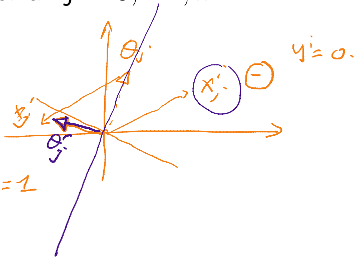
The perceptron learning algorithm

Perceptron hypothesis function:

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

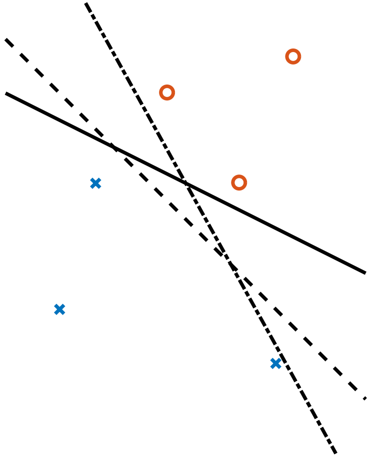
Parameter update rule:

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



- When prediction is correct: $\theta_j = \theta_j$
- When prediction is incorrect:
  - $y^i = 0$ predicted "1": $\theta_j = \theta_j - \alpha x_j$ $\alpha = 1$
  - $y^i = 1$ predicted "0": $\theta_j = \theta_j + \alpha x_j$

Issues with linear hyperplane perceptron:



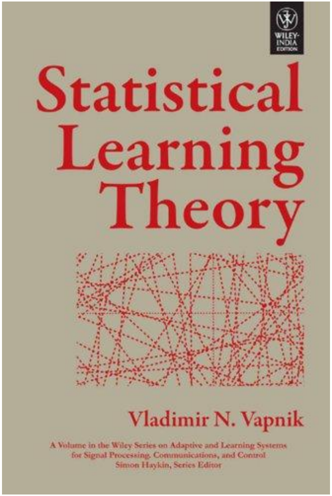
- Infinitely many solutions if data are separable
- Can not express “confidence” of the prediction

Support Vector Machines

- Optimal margin classifier
- Lagrange Duality
- Soft margin SVM

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



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

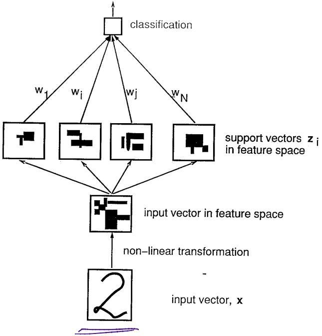
CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

Editor: Lorenza Saitta

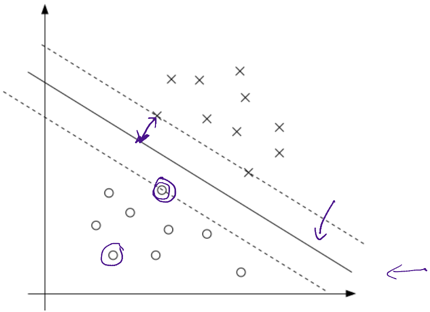
Abstract. The *support-vector network* is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. 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.

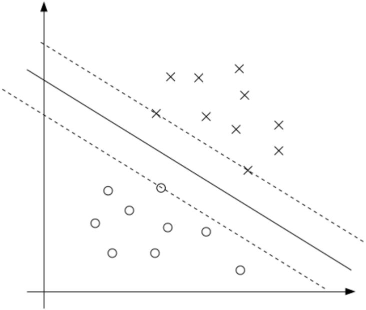


Support Vector Machine: Overview



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? *+ a detour on Lagrange Duality*

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 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 $(x^{(i)}, y^{(i)})$

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

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

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

$y^{(i)}$	≥ 0	< 0
1	++	--
-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 (w, b) with respect to training data S:

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

Function Margins

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

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

$wx + b = 0$

Issue: $\hat{\gamma}$ depends on $||w||$ and b $\boxed{\underbrace{(2w)}_{w'}x + \underbrace{(2b)}_{b'} = 0}$

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

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

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

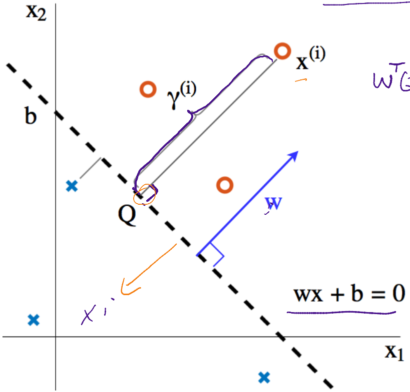
Geometric Margins

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

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

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

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



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

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:

geometric

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

Geometric Margins

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

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

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

Geometric Margins

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

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

- $\hat{\gamma} = \gamma$ 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

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

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

func marg

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 $\hat{\gamma} = 1$

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

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

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 $\hat{\gamma} = 1$

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

min $\underline{||w||}$

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

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

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

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 $\hat{\gamma} = 1$

$$\max_{\gamma,w,b} \quad \frac{1}{||w||}$$
$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1\ i = 1, \ldots, m$$
$$\iff \min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1\ i = 1, \ldots, m$$

can be solved using QP software

Review: Lagrange Duality

The **primal** optimization problem:

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

Review: Lagrange Duality

The **primal** optimization problem:

$$\begin{aligned} \min_{w} \quad & f(w) \\ s.t. \quad & g_i(w) \leq 0, i, \dots, k \\ & h_i(w) = 0, i = 1, \dots, l \end{aligned}$$

Define the **generalized Lagrange function** :

$\alpha_i \geq 0$

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

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

For a given w,

primal function

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

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

For a given w,

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

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

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

For a given w,

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

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

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

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

The primal problem (alternative form)

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

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)

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

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)

$\max_{\alpha,\beta} \min_w L(w,\alpha,\beta) \leq \min_w \max_{\alpha,\beta} L(w,\alpha,\beta)$

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) ← weak duality

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 $\boxed{g_i(w) < 0}$ 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

← strong .

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

Karush-Kuhn-Tucker (KKT) conditions

Under previous conditions, w^*, α^*, β^* are solutions of P and D **if and only if** they statisty the following conditions:

$$\frac{\delta}{\delta w_i} L(w^*, \alpha^*, \beta^*) = 0, \; i = 1, \dots n \tag{1}$$

$$\frac{\delta}{\delta \beta_i} L(w^*, \alpha^*, \beta^*) = 0, \; i = 1, \dots l \tag{2}$$

$$\alpha_i^* g_i(w^*) = 0, \; i = 1, \dots, k \tag{3}$$

$$g_i(w^*) \leq 0, \; i = 1, \dots, k \tag{4}$$

$$\alpha^* \geq 0, \; i = 1, \dots, k \tag{5}$$

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

Optimal Margin Classifier

$\min_{w} f(w)$ $\to b$

s.t. $g_i(w) \le 0$. for $i=1,..k$

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

Optimal margin classifier

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

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

$-(y^i(w^Tx^i+b)-1) \le 0.$

- $f(w) = \frac{1}{2}||w||^2$
- $g_i(w) = -\left(y^{(i)}(w^T x^{(i)} + b) - 1\right)$ $= y^i(w^Tx^i+b)-1.$

Generalized Lagrangian function: $\sum_i^m \alpha_i(g_i(w))$

$\max_{\alpha,\beta,w} \min L(w,\alpha,\beta)$

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

① $\frac{\partial L}{\partial w} = 0.$ $\quad \frac{\partial L(w,b,\alpha)}{\partial w_i} = w - \sum_{i=1}^m \alpha_i y^i x^i = 0.$ $\quad w = \sum_{i=1}^m \alpha_i y^i x^i$

② $\frac{\partial L}{\partial b} = 0$ $\quad \frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^m \alpha_i y^i = 0$

By the complementary slackness condition in KKT:

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

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

By the complementary slackness condition in KKT:

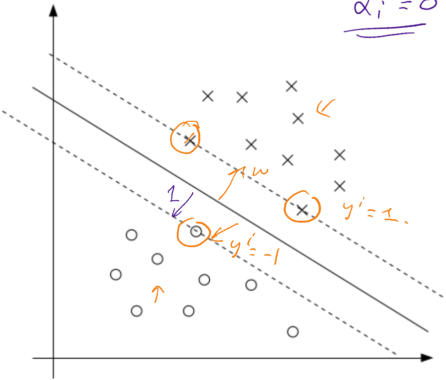
$$\alpha_i^* g_i(w^*) = 0, \;\; i = 1, \ldots, k$$

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

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

$g_i(w^*) = 0$.

$\alpha_i^* = 0 \leftarrow g_i(w^*) < 0$.



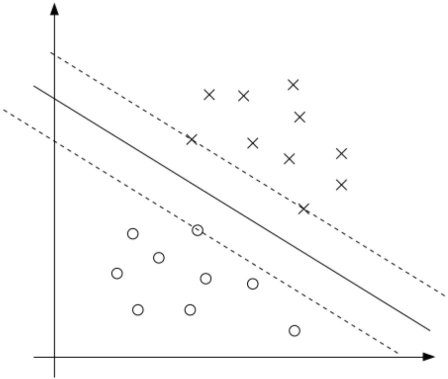
Support vectors lie on hyperplane ${w^*}^T x + b = 1$ when $y^{(i)} = 1$, or ${w^*}^T x + b = -1$ when $y^{(i)} = -1$

By the complementary slackness condition in KKT:

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

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

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 $g_i(w) \leq 0$ is only **active** on support vectors

Dual optimization problem: *(Check derivation)*

dual.

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

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

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

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

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

$\|w\|^2$

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

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

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

$$= -\frac{1}{2} \left(\sum_{i=1}^{m} \alpha_i y^i x^i \right)^T \left(\sum_{i=1}^{m} \alpha_i y^i x^i \right) + \sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1} \sum_{j=1} y^i y^j \alpha_i \alpha_j \underbrace{x^{i^T} x^j}_{\langle x^i, x^j \rangle}$$

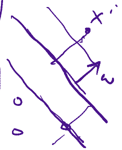
$$= \underline{W(\alpha)}$$

Dual optimization problem: *(Check derivation)*

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

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

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



If x^i is positive

$\min_{i, y^i=1} w^{*T} x^i + b = 1.$ ← worst margin for pos x^i

Solution to the primal problem:

$\max_{i, y^i=-1} w^{*T} x^i + b = -1.$ ← worst margin for neg x.

$$w^* = \sum_i \alpha_i^* y^{(i)} x^{(i)}$$

$\min_{y^i=-1} w^{*T} x^i + b + \min_{y^i=1} w^{*T} x^i + b = 0.$

$\max_{y^i=-1} w^{*T} x + \min_{y^i=1} w^{*T} x = -2b$

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

Dual optimization problem: *(Check derivation)*

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

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

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

Solution to the primal problem:

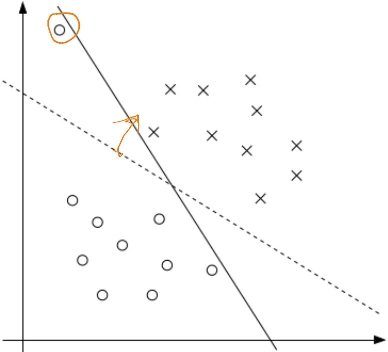
$$w^* = \sum_i \alpha_i^* y^{(i)} x^{(i)}$$

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

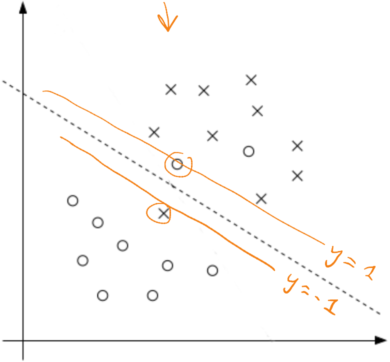
For a new sample z, the SVM prediction is sign $\left[{w^*}^T z + b^* \right]$

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

Limitations of the basic SVM



Outliers



Non-linearly separable cases

Soft Margin SVM

$0 \le \xi_i$

slackness

Functional margin $\boxed{1-\xi_i} \le 1$:

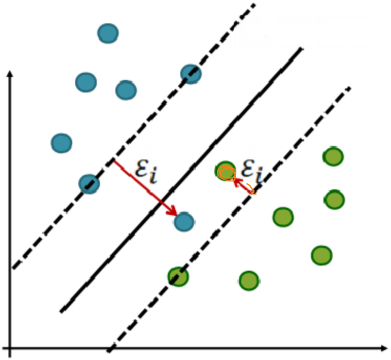
$m \times 1$

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \quad |\xi|$$

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

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

- C: relative weight on the regularizer
- $\underline{L_1}$ regularization let most $\underline{\xi_i} = 0$, such that their functional margins $1 - \xi_i = 1$



Soft Margin SVM

The generalized Lagrangian function:

$g_i(w)$.

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

$$-\sum_{i=1}^{m} r_i\xi_i \leftarrow \xi_i \geq 0$$

Soft Margin SVM

The generalized Lagrangian function:

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

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

By the KKT condition,

$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$

$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0.$ $\quad \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - r_i = 0$ for all i

Dual problem:

$$L(w, b, \xi, \alpha, r) = \frac{1}{2}||w||_2^2 - \sum_{i=1}^{m} \alpha_i [y^i(w^T x^i + b) - 1] - \underbrace{\sum_{i=1}^{m} \xi_i (C - \alpha_i - r_i)}_{=0.}$$

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

Since $C - \alpha_i - r_i = 0$

① $r_i = C - \alpha_i$

② $\alpha_i \geq 0, \; r_i \geq 0$

$\Rightarrow \alpha_i \leq C$

$\left.\right\} \; 0 \leq \alpha_i \leq C$

Soft Margin SVM

The generalized Lagrangian function:

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

Dual problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$
$$s.t.\ \boxed{0 \le \alpha_i \le C,}\ i = 1, \ldots, m$$
$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

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

Soft Margin SVM

Dual problem:

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

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

$r_i \varepsilon_i = 0 \leftarrow 2$

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

$\alpha_i (y_i (w^T x_i + b^*) - 1 + \varepsilon_i) = 0 \leftarrow 1$

if $\alpha_i > 0$, then $y_i (w^T x_i + b) - 1 + \varepsilon_i = 0$

if $\alpha_i \neq 0$, $\begin{cases} \varepsilon_i = 0 & y^{(i)}(w^T x^i + b) = 1 \\ \varepsilon_i > 0 & y^i (w^T x^i + b) \leq 1 \end{cases}$

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

$\alpha_i = 0 \quad \Longleftrightarrow$ ← see supplementary notes

$\alpha_i = C \quad \Longleftrightarrow$

$0 < \alpha_i < C \quad \Longleftrightarrow$

Soft Margin SVM

Dual problem:

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

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

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

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

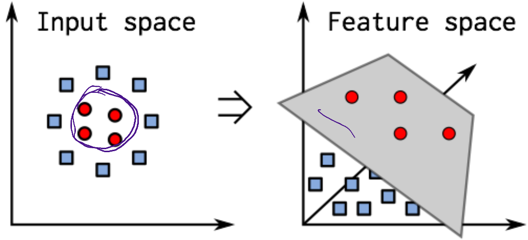
$$\begin{array}{llll}
\alpha_i = 0 & \iff & y^{(i)}(w^T x^{(i)} + b) \geq 1 & \text{correct side of margin} \\
\alpha_i = C & \iff & y^{(i)}(w^T x^{(i)} + b) \leq 1 & \text{wrong side of margin} \\
0 < \alpha_i < C & \iff & y^{(i)}(w^T x^{(i)} + b) = 1 & \text{at margin}
\end{array}$$

Kernel SVM

Non-linear SVM

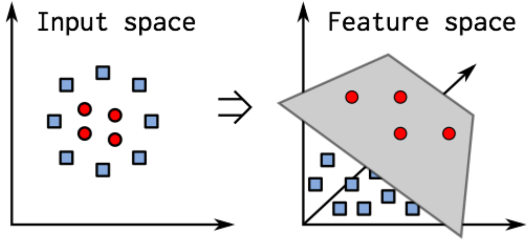
For non-separable data, we can use the **kernel trick**: Map input values $x \in \mathbb{R}^d$ to a higher dimension $\phi(x) \in \mathbb{R}^D$, such that the data becomes separable.

$\phi : \mathbb{R}^d \to \mathbb{R}^D$



Non-linear SVM

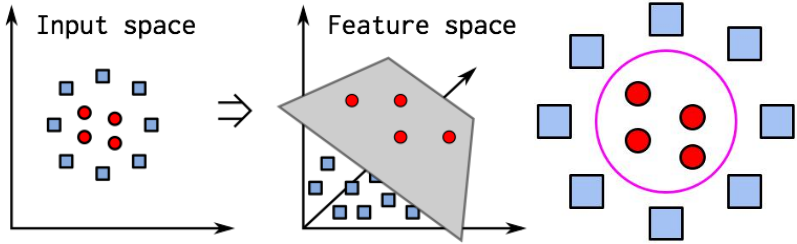
For non-separable data, we can use the **kernel trick**: Map input values $x \in \mathbb{R}^d$ to a higher dimension $\phi(x) \in \mathbb{R}^D$, such that the data becomes separable.



- 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.

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

Kernel Function

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

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

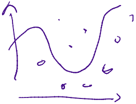
Some kernel functions are easier to compute than $\phi(x)$, e.g.

a polynomial kernel

$$K(x, z) = (x^T z)^2 = \sum_{i=1}^{n} x_i, z_i \sum_{j=1}^{n} x_j, z_j = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i, x_j, z_i, z_j$$

$O(n)$

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



n^2

$\phi(x)^T \phi(z)$

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 $\langle \phi(x_i), \phi(y_i) \rangle = K(x_i, x_j)$

$\langle x_i, x_j \rangle$. $\langle \phi(x_i), \phi(x_j) \rangle$

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

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

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

Kernel SVM

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

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

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

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

Big-O Notation:

Let $T(n)$ be the number of arithmitic operations an algorithm takes on input of size n, then we write $T(n) = O(f(n))$ if there exits constant $C > 0$ such that $T(n) \leq C f(n)$ for all n.

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

$$w^T x + b = \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \phi(x^{(i)}) \right)^T \phi(x) + b$$

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

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

Kernel Matrix

$\underbrace{K(x^i, x^j)}_{(m\times m)}$

$$K = \begin{array}{c} K(x^1,x^1) \\ \begin{bmatrix} \phi(x^1)^T\phi(x^1) & \phi(x^1)^T\phi(x^2) & \cdots & \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ \phi(x^m)^T\phi(x^1) & \cdots & \cdots & \phi(x^m)^T\phi(x^m) \end{bmatrix} \end{array}$$

Intuitively, kernel functions measures the similarity between samples x and z. $\phi(x) = x$ $\phi(z) = z$

Examples:

- Linear kernel: $K(x, z) = (x^T z + c)^n$
- Gaussian or radial basis function (RBF) kernel:
  $K(x, z) = \underbrace{\exp\left(-\frac{||x - z||^2}{2\sigma^2}\right)}$

Kernel Matrix

Intuitively, kernel functions measures the similarity between samples x and z.

Examples:

- Linear kernel: $K(x, z) = (x^T z + c)^n$
- Gaussian or 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 $K \in \mathbb{R}^{n \times n}$ where $K_{i,j} = K(x_i, x_j) = \phi(x_i)\phi(x_j)$.

Kernel Matrix

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

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 $\{x^{(i)}, \dots, x^{(m)}\}$, K is symmetric positive semi-definite.

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

PSD

$K = K^T$

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.

libsvm

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.)