50.007 Machine Learning

Support Vector Machines

Berrak Sisman

Assistant Professor, ISTD Pillar, SUTD

Introduction & Content

- **Support Vector Machines (week 4)**
- Logistic Regression (week 4/5)

Instructor: Prof. Berrak Sisman

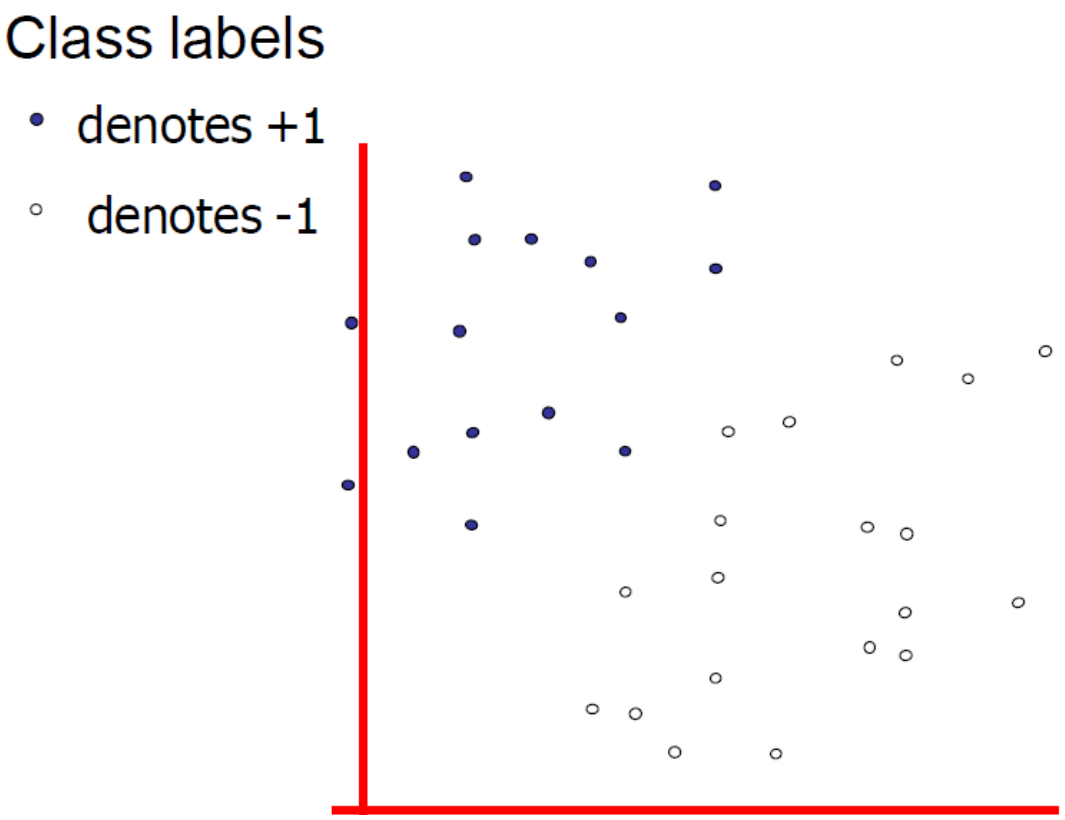
Email: berrak_sisman@sutd.edu.sg

Feel free to contact me!

BIG PICTURE OF SVM

Support Vector Machines

- State-of-the-art classifier

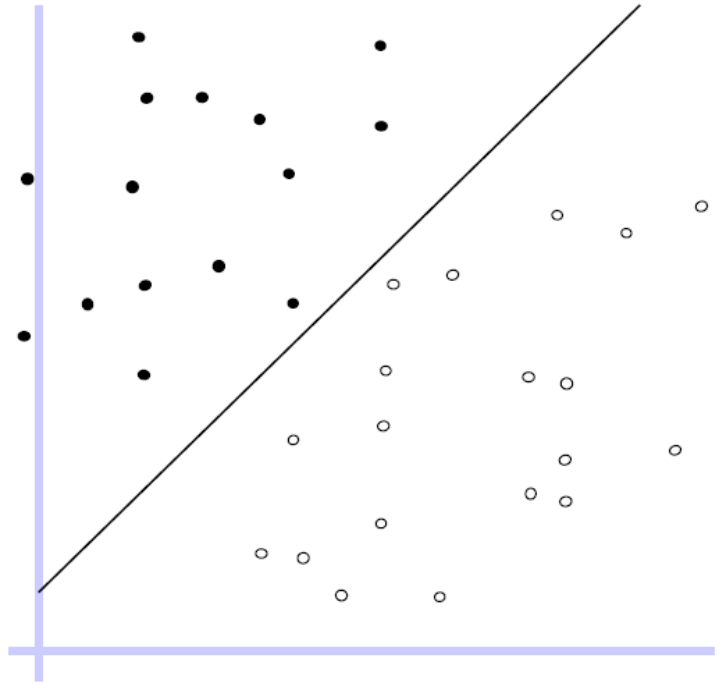


How do you classify this data?

Linear classifier

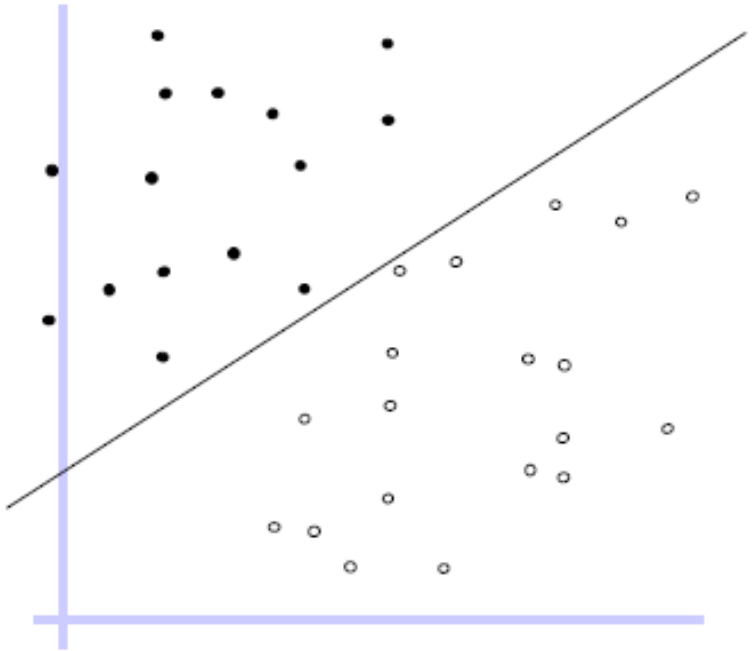
If all you can do is to draw a straight line:

‘linear decision boundary’

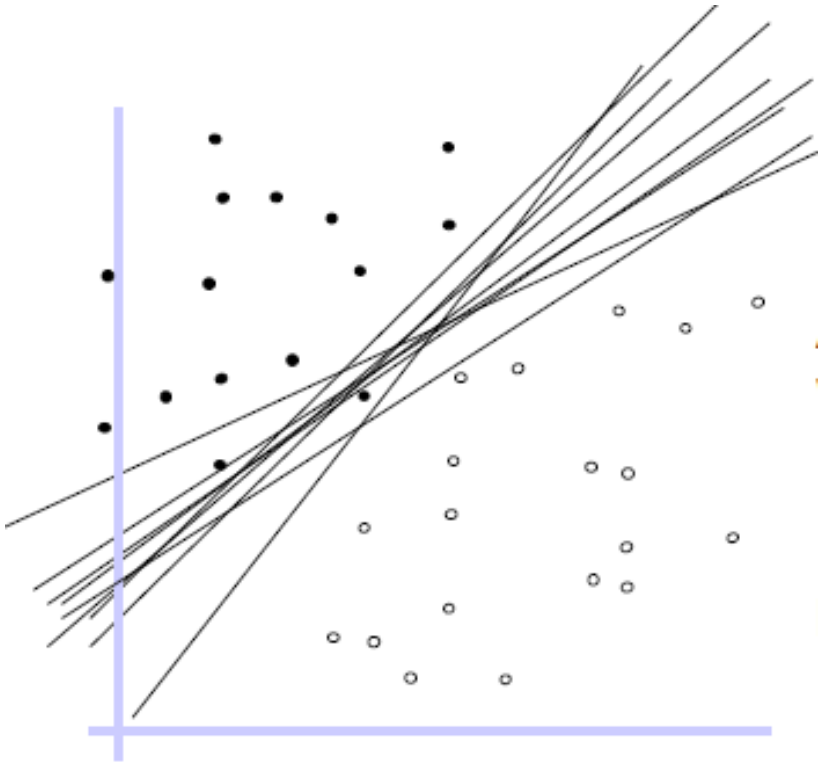


Linear classifier

Another OK ‘decision boundary’



Any of these would be fine…

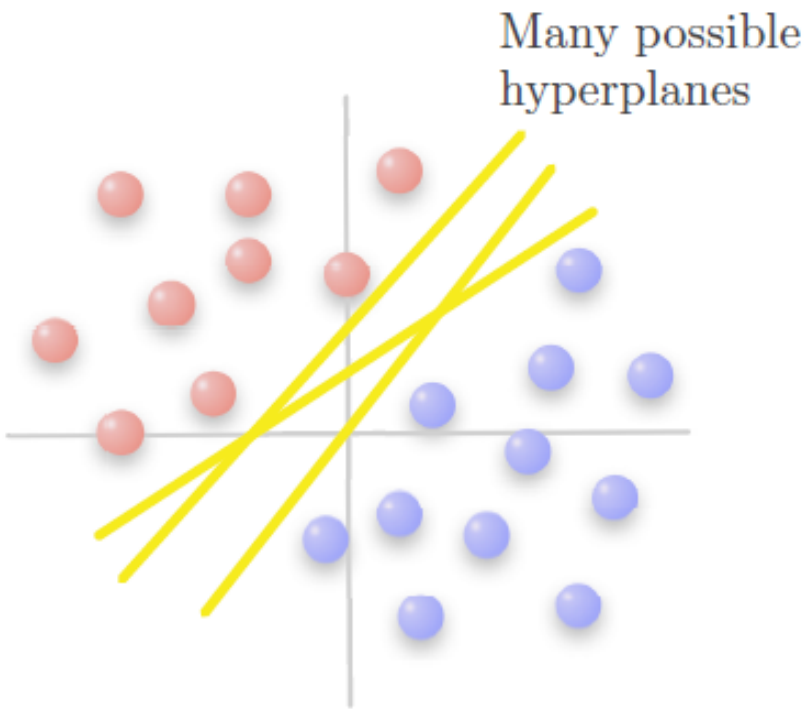


But which one is the best?

Is there an optimal way to separate the data?

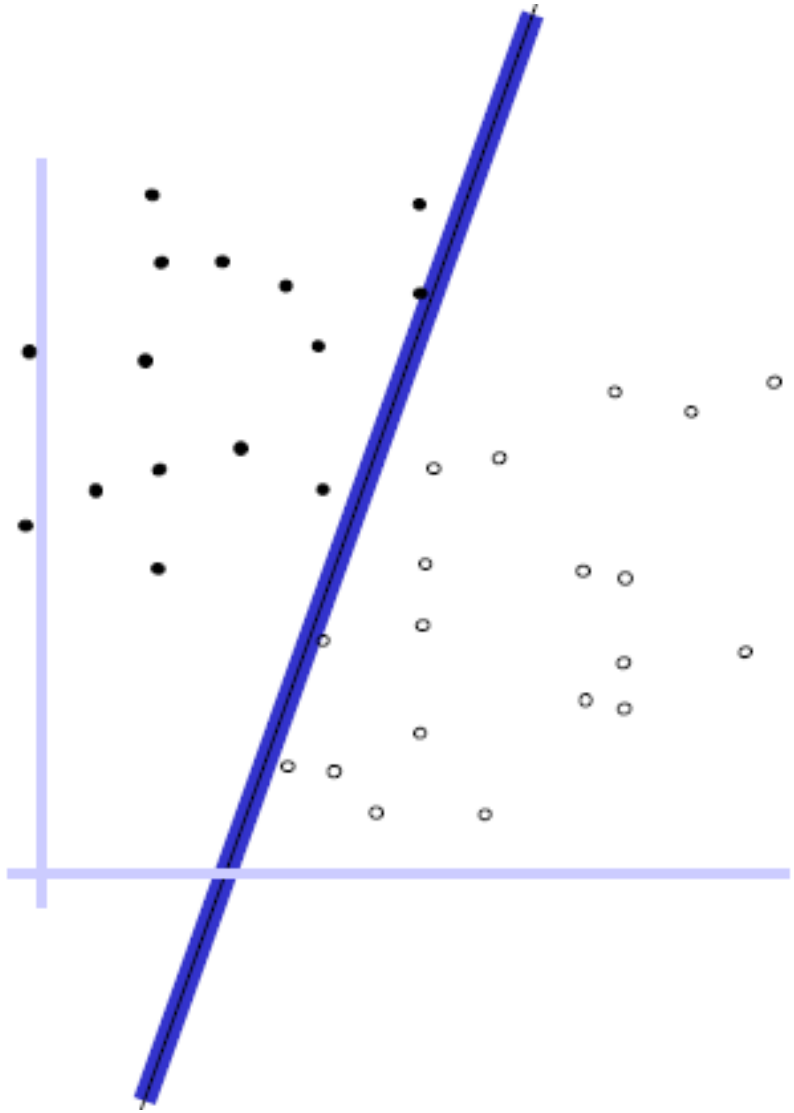
The theory of support vector machines provides a systematic method for separating the data optimally.

It involves a key concept called **margin.**



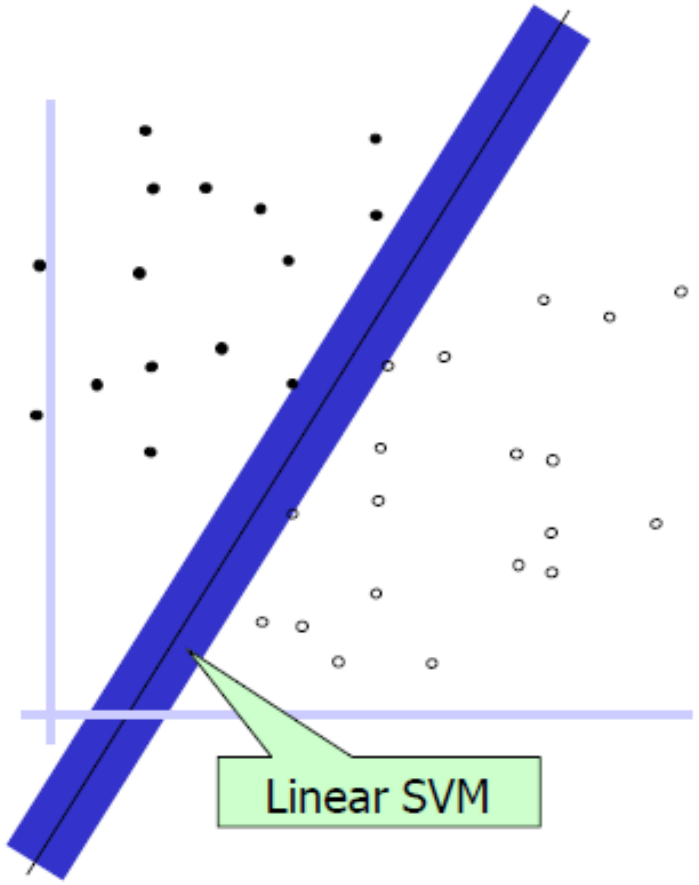
The Margin

Margin: The width that the boundary can be increased before hitting a data point.



SVM: maximize margin

The simplest SVM (linear SVM) is the linear classifier with the maximum margin.



SVM: not linearly separable data

What if the data is not linearly separable?

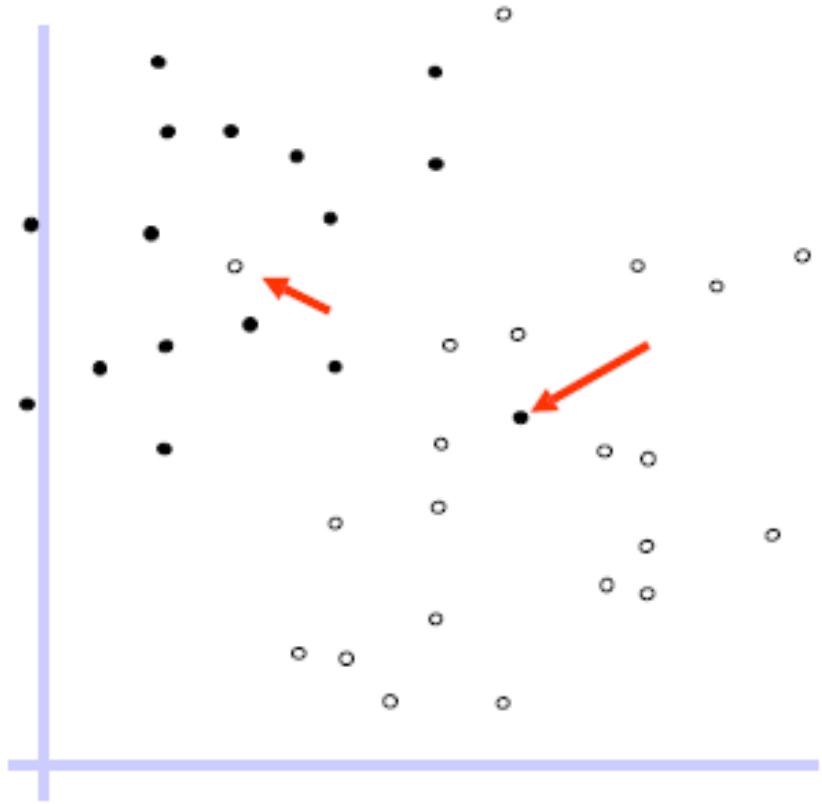
Two solutions:

1) Soft margin:

Allow a few points on the wrong side (slack variables), and/or

2) Kernel:

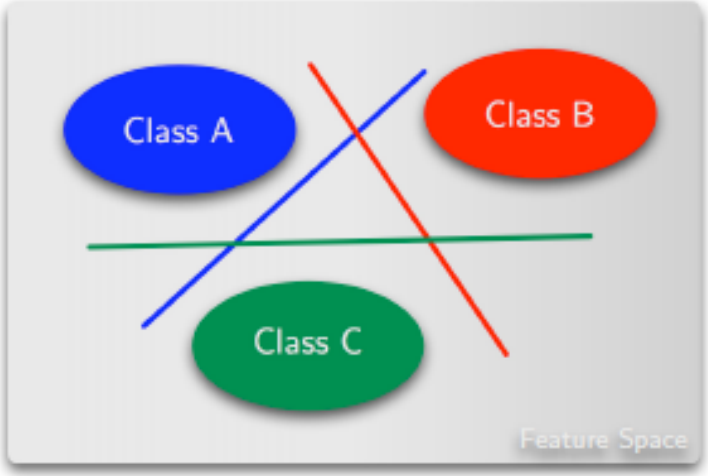
Map data to a higher dimensional space, do linear classification there.



SVM: more than two classes

N class problem: Split the task into N binary tasks:

- Class A vs. the rest (class B,…N)
- Class B vs. the rest (class A, B,.. N)
- ….
- Class N vs. the rest



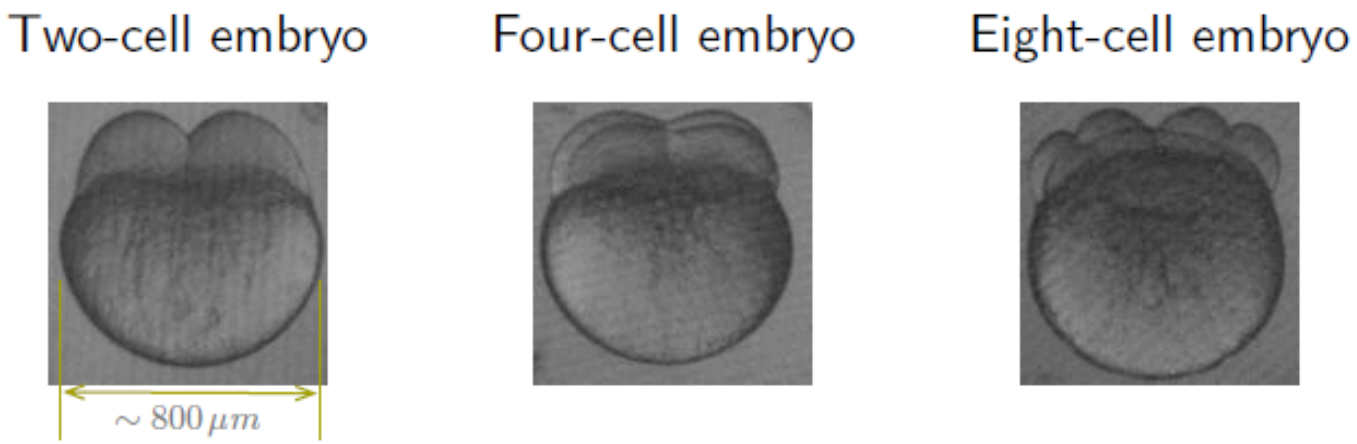
Finally, pick the class that put the point furthest into the positive region.

In this lecture, we'll work with 2 classes.

SVM Applications - 1

SVM for classification of zebra fish embryos

- Injecting DNA material into embryos to study development
- Injection done at the right development stage of embryo (before 8-cell embryo)
- Important to classify stage of embryos before injection, because the DNA that is used can be different.

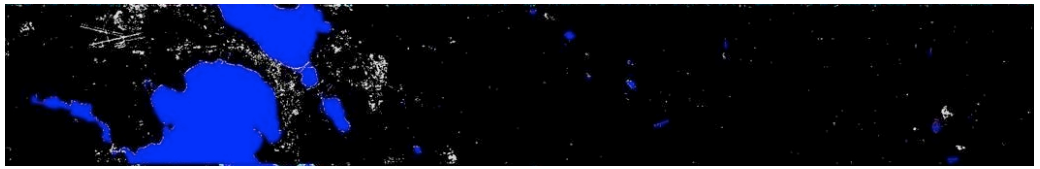


SVM Applications - 2

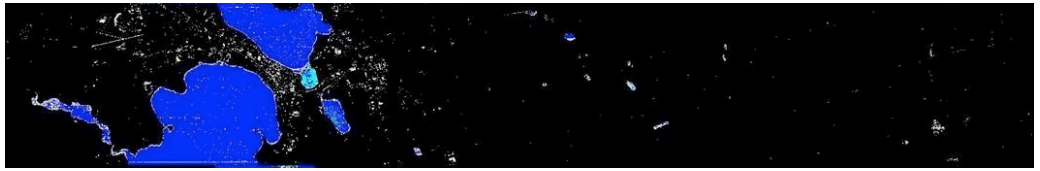
SVM for identifying areas of land cover (land, ice, water, snow) in a scene.

Lake Mendota, Wisconsin

SVM



Expert Labelled



Visible Image

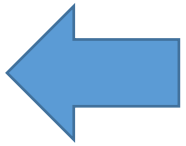
And many more applications where classification is required…

SVM: get your hands on it

You will have an implementation homework ☺

There are many implementations available:

- http://svmlight.joachims.org/
- http://www.support vector.net/software.html
- https://towardsdatascience.com/svm-implementation-from-scratch-python-2db2fc52e5c2
- https://github.com/cjlin1/libsvm

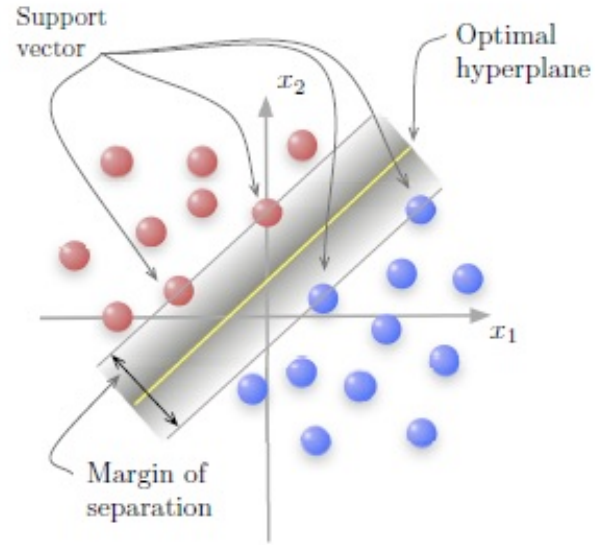


You will use this link for your homework.

You don't have to know the rest of the class in order to use SVM.

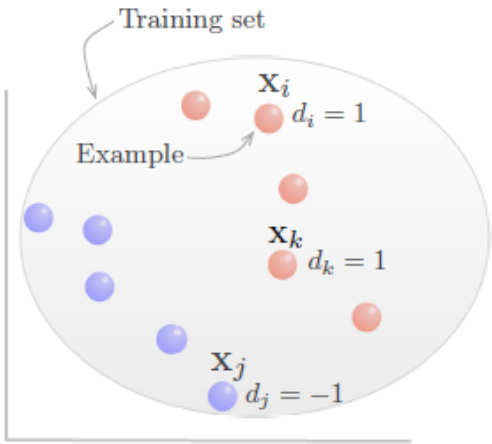
You want to learn more, don’t you? ☺

The math version:
Support Vector Machines

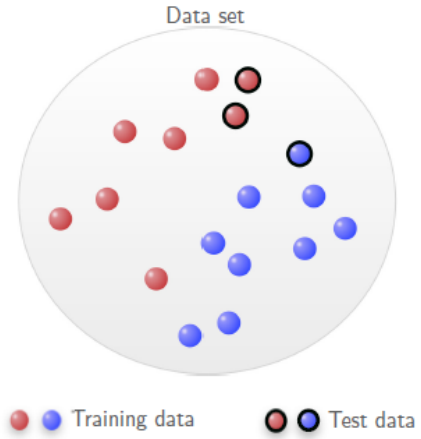


SVM Notation

- Let's assume we have N-dimensional input vector x, where x is equal to $[x_1, \dots, x_N]^T$, with a label $d = \pm 1$.
- SVM parameters: w and b



- The training set can be written as:



$$S = \{(x_1, d_1), (x_2, d_2), \dots, (x_N, d_N)\}$$

Training data is for constructing SVM, and test data is for evaluation.

In this slide, SVM parameters are denoted as w and b. In your lecture notes, they are denoted as θ and θ_0. It is the same thing! ☺

Classification with a hyperplane

- A hyperplane, denoted by (w, b), can be expressed as

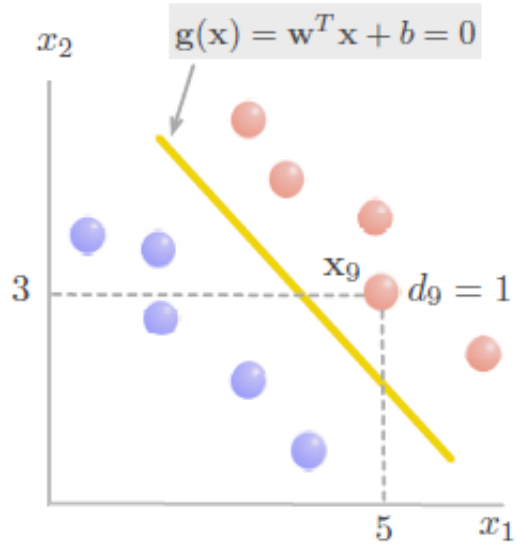
$$g(x) = w^T x + b = 0$$

Hyperplane classifies a given x_i with

$$sgn\,[g(x_i)] = \begin{cases} +1\ if\ g(x_i) > 0 \\ -1\ if\ g(x_i) < 0 \end{cases}$$

Hyperplane classifies an example (x_i, d_i) correctly if

$$sgn\,[g(x_i)] = d_i$$



Problem: Let $\mathbf{w}^T = [\ 7 \quad 6\ ]$ and $b = -42$. For $\mathbf{x}_9^T = [\ 5 \quad 3\ ]$

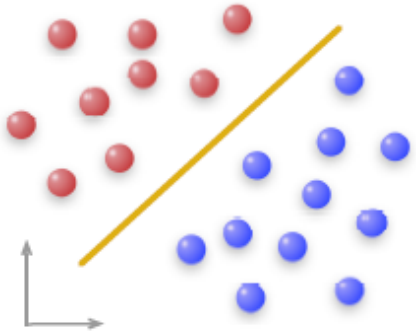
$$\mathbf{g(x_9)} = \mathbf{w}^T\mathbf{x}_9 + b = [\ 7 \quad 6\ ] \begin{bmatrix} 5 \\ 3 \end{bmatrix} - 42 = 11 > 0$$

$$\mathbf{sgn[g(x_9)]} = \mathbf{sgn[11]} = 1 = d_9$$

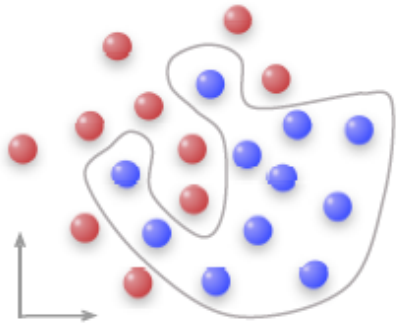
Linearly separable, or not?

Two classes of data are linearly separable if and only if there exists a hyperplane $w^T x + b = 0$ that separates the two classes.

Example in 2D:



Linearly separable



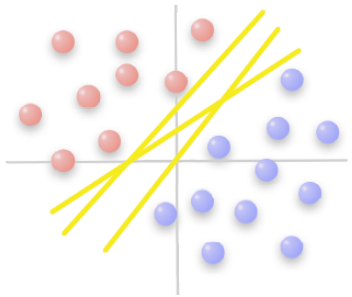
Non-linearly separable

We'll first study the case of "linearly separable".

Margins

The theory of support vector machines provides a systematic method for separating the data optimally, and it involves a key concept called **margin**.

Many possible hyperplanes



γ_i^f = Functional margin of an example (x_i, d_i)

γ_i^g = Geometric margin of an example (x_i, d_i)

γ^f = Functional margin of a training set

γ^g = Geometric margin of a training set

γ = Margin of a training set

defined w.r.t. a given hyperplane (w, b)

A training set S can have different geometric margins depending on how hyperplane is defined.

The optimal hyperplane for a given S is the one that gives **maximum γ^g over all possible hyperplanes**.

Functional Margin of an example

The functional margin of an example $(\mathbf{x}_i, d_i)$ with respect to a hyperplane $(\mathbf{w}, b)$ is defined as

$$\gamma_i^f = d_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)$$

Problem:

Training set with only 1 example

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \; d_1 = +1$$

Hyperplane defined by

$$\mathbf{w} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \; b = 6$$

Determine γ_1^f with respect to given hyperplane

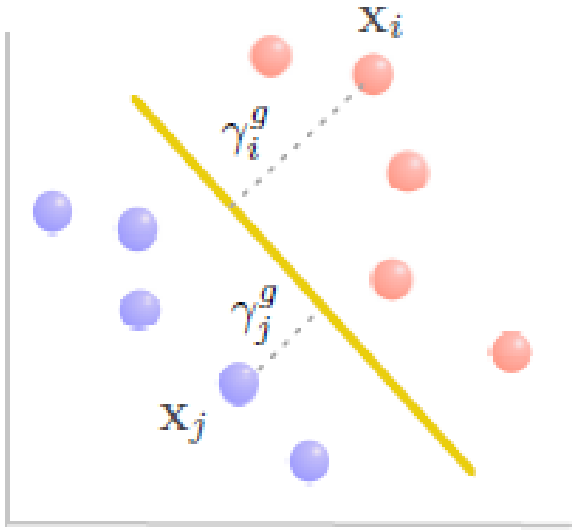
Solution:

$$\gamma_1^f = d_1 \left(\mathbf{w}^T \mathbf{x}_1 + b\right) = 1 \times \left(\begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} + 6 \right) = 1 \times 23 = 23$$

Geometric Margin of an example

The geometric margin of an example $(\mathbf{x}_i, d_i)$ with respect to a hyperplane $(\mathbf{w}, b)$ is defined as

$$\gamma_i^g = d_i \left(\frac{1}{\|\mathbf{w}\|} \mathbf{w}^T \mathbf{x}_i + \frac{1}{\|\mathbf{w}\|} b \right)$$



γ_i^g is the Euclidean distance from the point defined by $\mathbf{x}_i$ to the hyperplane $(\mathbf{w}, b)$.

Problem:

Training set with only 1 example

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \; d_1 = +1$$

Hyperplane defined by

$$\mathbf{w} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \; b = 6$$

Determine γ_1^g with respect to given hyperplane

Solution:

$$\gamma_1^g = d_1 \left(\frac{\mathbf{w}^T \mathbf{x}_1}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|} \right) = \frac{\begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} + 6}{\sqrt{5^2 + 3^2}} = \frac{23}{\sqrt{34}}$$

Remember: the norm of a vector $\|w\|$ is its length

Functional Margin vs Geometric Margin

Functional margin of an example

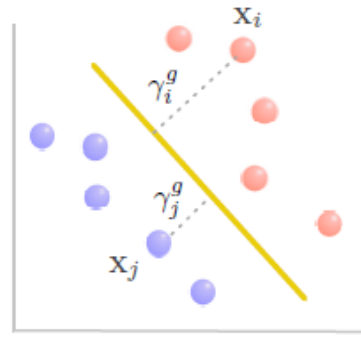
$$\gamma_i^f = d_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)$$

Geometric margin of an example

$$\gamma_i^g = d_i \left(\frac{1}{\|\mathbf{w}\|}\mathbf{w}^T \mathbf{x}_i + \frac{1}{\|\mathbf{w}\|}b\right)$$

Therefore, by definition:

$$\gamma_i^g = \frac{\gamma_i^f}{\|\mathbf{w}\|}$$



γ_i^g is the Euclidean distance from the point defined by $\mathbf{x}_i$ to the hyperplane $(\mathbf{w}, b)$.

Let's solve

(With a scaling factor, what happens to margin?)

Problem: (Note that from earlier slide, $\gamma_1^f = 23$ and $\gamma_1^g = \frac{23}{\sqrt{34}}$)

Training set with only 1 example

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \; d_1 = +1$$

Given hyperplane

$$\mathbf{w} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \; b = 6$$

Find scaling constant c such that $\gamma_1^f = 5$ with respect to the given hyperplane, and calculate γ_1^g associated with $c\mathbf{w}$ and cb.

Solution:

$$d_1 \left(c\,\mathbf{w}^T\mathbf{x}_1 + c\,b\right) = 1 \left(c \begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} + 6c\right) = 23\,c = 5 \;\Rightarrow\; c = \frac{5}{23}$$

Thus

$$c\mathbf{w} = \frac{5}{23} \begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 25/23 \\ 15/23 \end{bmatrix}, \quad cb = \frac{30}{23}$$

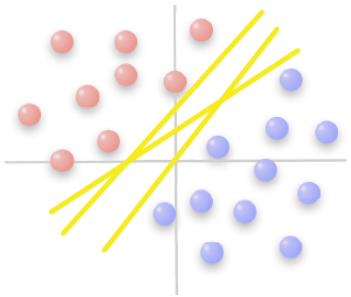
and the geometric margin of the example with respect to $c\mathbf{w}$ and cb is

$$\gamma_1^g = \frac{\gamma_1^f}{\|c\mathbf{w}\|} = \frac{5}{\sqrt{\left(\frac{25}{23}\right)^2 + \left(\frac{15}{23}\right)^2}} = \frac{5}{\frac{5\sqrt{34}}{23}} = \frac{23}{\sqrt{34}}$$

Margins

The theory of support vector machines provides a systematic method for separating the data optimally, and it involves a key concept called **margin**.

Many possible hyperplanes



γ_i^f = Functional margin of an example (x_i, d_i) ✓

γ_i^g = Geometric margin of an example (x_i, d_i) ✓

defined w.r.t. a given hyperplane (w, b)

γ^f = Functional margin of a training set

γ^g = Geometric margin of a training set

γ = Margin of a training set

Let's study these now!

A training set S can have different geometric margins depending on how hyperplane is defined.

The optimal hyperplane for a given S is one that gives **maximum γ^g over all possible hyperplanes**.

Functional margin of a training set

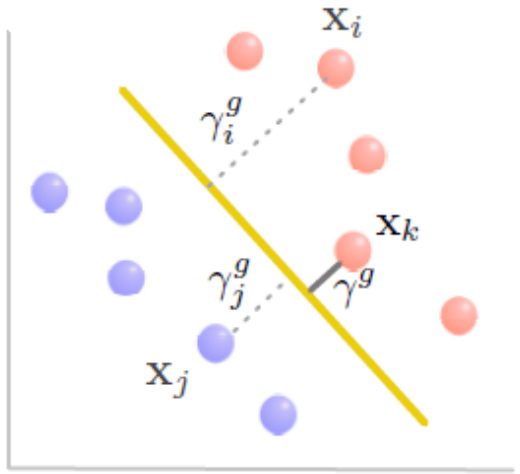
The functional margin of a training set S, with respect to a hyperplane (w, b), is the minimum of all the functional margins of the individual examples in S.

$$\gamma_i^f = d_i \left(\mathbf{w}^T \mathbf{x}_i + b \right)$$

$$\gamma^f = \min_{1 \le i \le N} \left\{ \gamma_i^f \right\}$$

Geometric margin of a training set

The geometric margin of a training set S, with respect to a hyperplane (w, b), is the minimum of all the geometric margins of the individual examples in S.



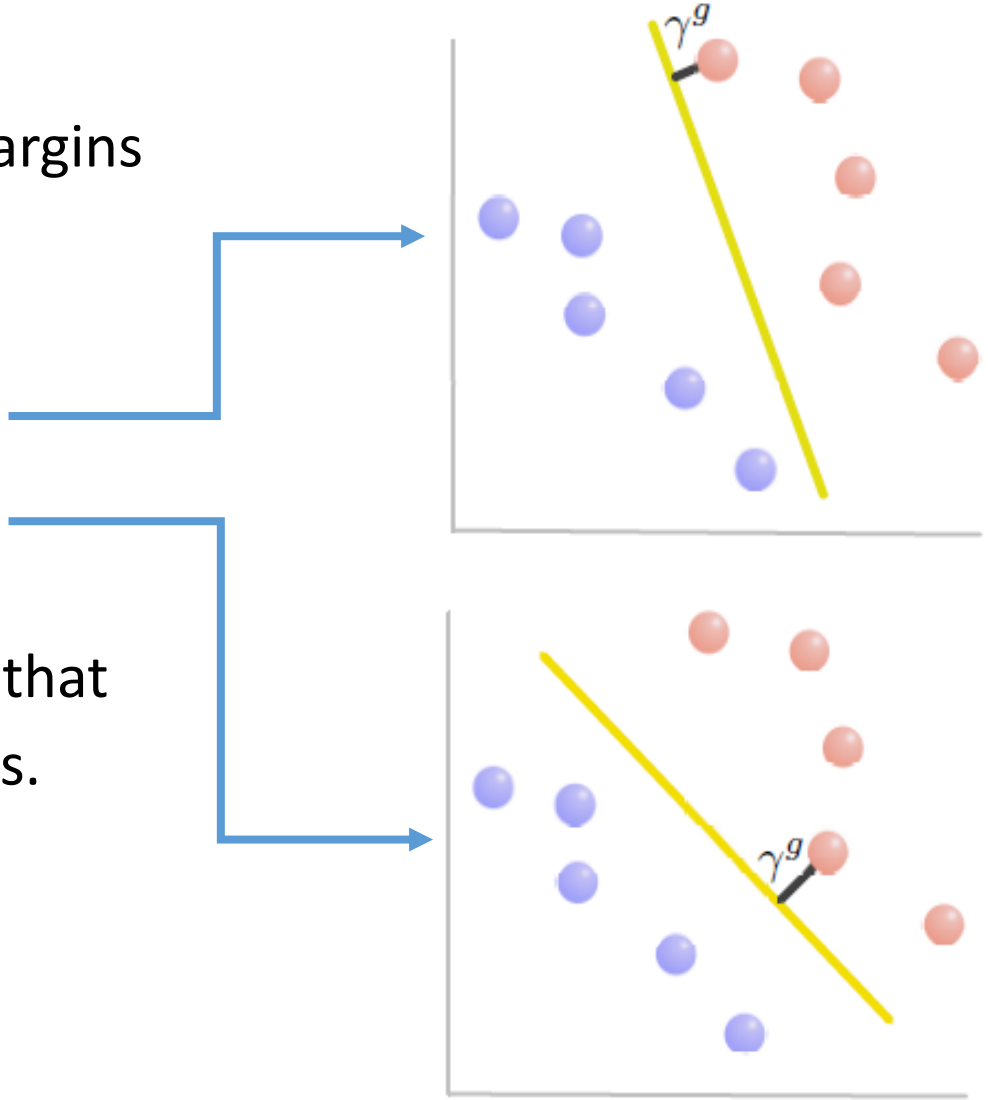
$$\gamma_i^g = d_i \left(\frac{1}{\|\mathbf{w}\|} \mathbf{w}^T \mathbf{x}_i + \frac{1}{\|\mathbf{w}\|} b \right)$$

$$\gamma^g = \min_{1 \le i \le N} \{ \gamma_i^g \}$$

How to find the optimal hyperplane?

A training set S can have different geometric margins depending on how hyperplane is defined.

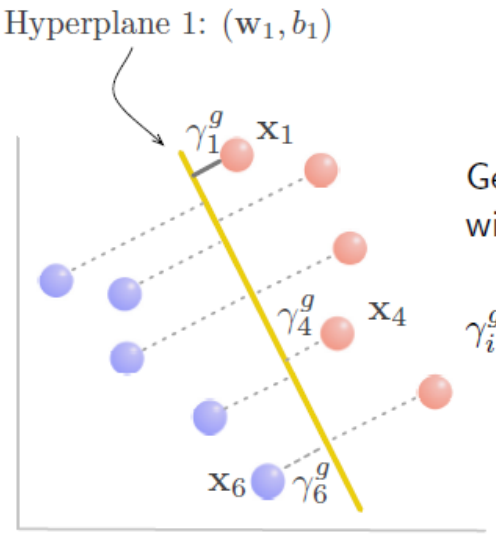
Which one?



The optimal hyperplane for a given S is the one that gives maximum $\boldsymbol{\gamma^g}$ over all possible hyperplanes.

This maximum $\boldsymbol{\gamma^g}$ is called the margin of $\boldsymbol{S}$.

Step 1: Geometric margin of training set

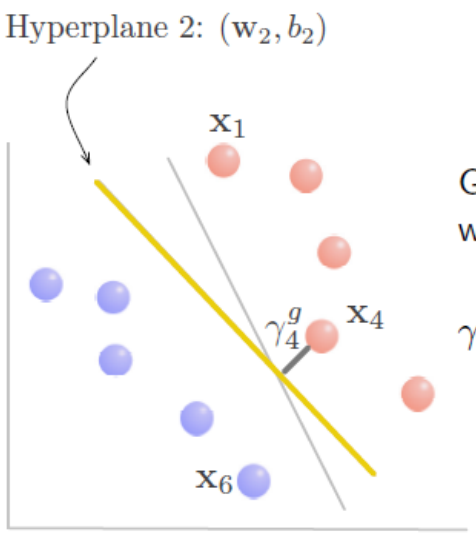


Geometric margin of an example i with respect to **hyerplane 1** is

$$\gamma_i^g = d_i \left(\left\langle \frac{1}{\|\mathbf{w}_1\|} \mathbf{w}_1 \cdot \mathbf{x}_i \right\rangle + \frac{1}{\|\mathbf{w}_1\|} b_1 \right)$$

Geometric margin of training set with respect to hyperplane 1 is:

$$\min\{\gamma_1^g, \ldots, \gamma_{10}^g\} = \gamma_1^g \equiv \gamma_{(\mathbf{w}_1, b_1)}^g$$



Geometric margin of an example i with respect to **hyerplane 2** is

$$\gamma_i^g = d_i \left(\left\langle \frac{1}{\|\mathbf{w}_2\|} \mathbf{w}_2 \cdot \mathbf{x}_i \right\rangle + \frac{1}{\|\mathbf{w}_2\|} b_2 \right)$$

Geometric margin of training set with respect to hyperplane 2 is:

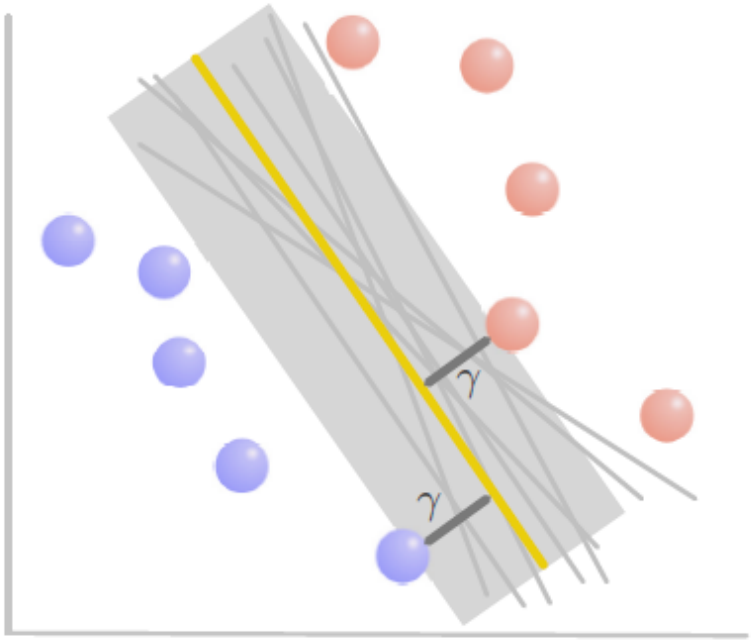
$$\min\{\gamma_1^g, \ldots, \gamma_{10}^g\} = \gamma_4^g \equiv \gamma_{(\mathbf{w}_2, b_2)}^g$$

Repeat this for all the possible hyperplanes…

Step 2: Margin of training set

The optimal hyperplane for a given S (data) is one that gives maximum $\boldsymbol{\gamma^g}$ over all possible hyperplanes

$$\gamma = \max \underbrace{\left\{\gamma^g_{(\mathbf{w}_1,b_1)}, \gamma^g_{(\mathbf{w}_2,b_2)}, \gamma^g_{(\mathbf{w}_3,b_3)}, \cdots\right\}}_{\text{All possible hyperplanes}}$$



Why is it important to find the margin of S?

Larger margin leads to lower probability of misclassification.

How to find the margin of $\mathbf{S}$?
(mathematically)

- Let's recall the relationship: $$\gamma_i^g = \frac{\gamma_i^f}{\|\mathbf{w}\|}$$

- For a given w, the example k that yields $\boldsymbol{\gamma}_i^f = \boldsymbol{\gamma}^f$ also yields $\boldsymbol{\gamma}_i^g = \boldsymbol{\gamma}^g$. So we can write

$$\gamma^g = \frac{\gamma^f}{\|\mathbf{w}\|}$$

We can maximize γ^g by fixing γ^f then minimizing $\|\mathbf{w}\|$

How to find the margin of S?

Let's fix γ^f

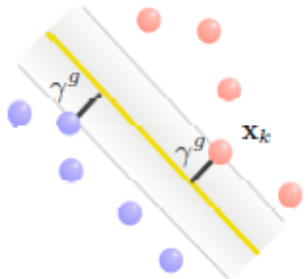
Since a hyperplane is invariant under scaling by a constant c, for any hyperplane $(\mathbf{w}, b)$, we can find c such that

$$\gamma^f \equiv \underbrace{\gamma_k^f = d_k \left(c\,\mathbf{w}^T \mathbf{x}_k + c\,b\right)}_{\text{functional margin of example } k} = 1$$

That is, we fix functional margin γ^f of training set to be 1.

Problem: Find c such that $\gamma_k^f = 1$, with

$$\mathbf{x}_k = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \quad d_k = +1\,, \quad \mathbf{w} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}, \quad b = 6$$



$$d_k \left(c\,\mathbf{w}^T \mathbf{x}_k + c\,b\right) = 1 \cdot c \left(\begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} + 6 \right) = 23\,c = 1 \;\Rightarrow\; c = \frac{1}{23}$$

How to find the margin of S?

Let's fix $\boldsymbol{\gamma}^f$

Recall:

$$\gamma^f = \min_{1 \le i \le N} \left\{ \gamma_i^f \right\}$$

$$\gamma_i^f = d_i \left(\mathbf{w}^T \mathbf{x}_i + b \right)$$

- With $\boldsymbol{\gamma}^f$ fixed at 1, for any example x_i, we have the following condition

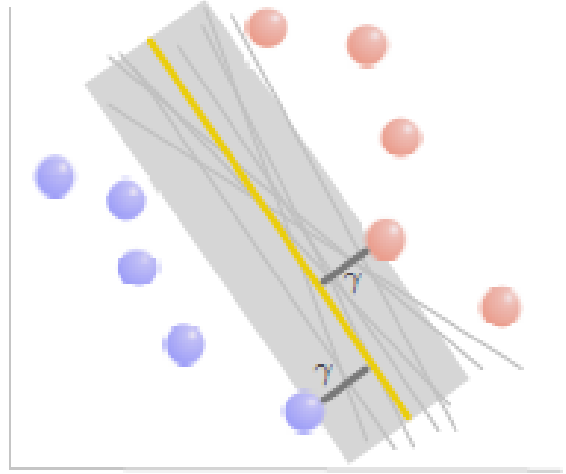
$$\gamma_i^f = \gamma_i^g \|\mathbf{w}\| = d_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) \geq 1$$

- The problem of minimizing $\|w\|$ can now be expressed as a

Constrained optimization problem

$$\text{Minimizing:} \quad f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{Subject to:} \quad d_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) \geq 1$$

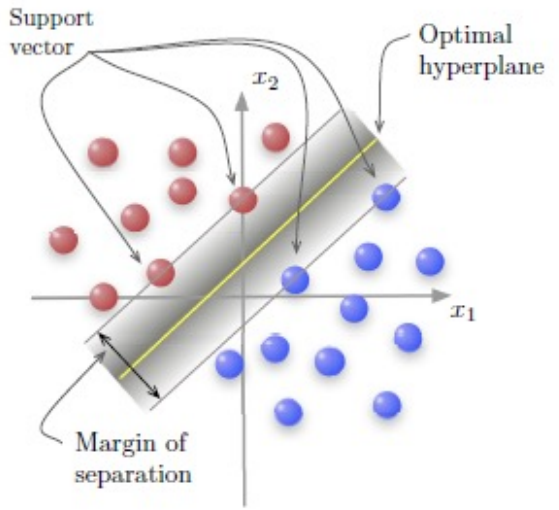


By solving this problem, we can obtain the optimal hyperplane (w_0, b_0)

Discriminant function and support vector

After obtaining w_0, b_0, for training data x_i, classification will be as follows:

$$g(x_i) = (w_0^T x_i + b_0) \geq 1 \text{ for } d_i = +1$$
$$g(x_i) = (w_0^T x_i + b_0) < -1 \text{ for } d_i = -1$$



Discriminant function:

$$g(x) = w_0^T x + b_0$$

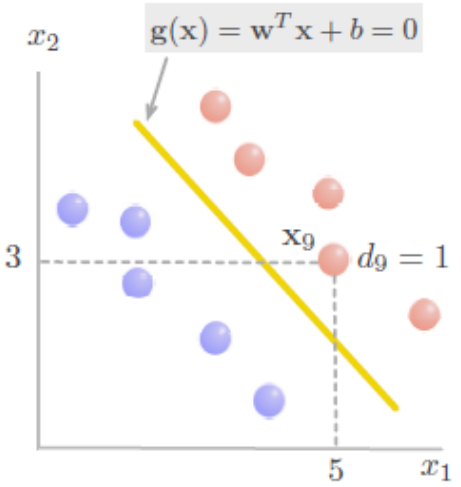
Support vector: x_i that satisfies

$$g(x_i) = \pm 1$$

Minimizing $\|\boldsymbol{w}\|$: Constrained optimization

Primal problem

$$\begin{aligned} \text{Given data set :} \quad & S = \{(\mathbf{x}_i, d_i)\}, \quad i = 1, 2, \ldots, N \\ \text{Find :} \quad & \mathbf{w} \text{ and } b \\ \text{Minimizing :} \quad & f(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{Subject to :} \quad & d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \end{aligned}$$



We will solve this problem with Lagrange Theorem. Solving this problem yields the optimal hyperplane (w_0, b_0)!

Joseph-Louis Lagrange
French mathematician
1736–1813

Recall

Background (pages 35-44)

Minimizing $\|w\|$: Constrained optimization

Lagrange Theorem:

Minimize: $f(w)$

Subject to: $h_i(w) = 0,\ i = 1,2,\dots,m$

Joseph-Louis Lagrange
French mathematician
1736–1813

Let's define the **Lagrangian Function**:

$$L(\mathbf{w}, \beta) = f(\mathbf{w}) + \sum_{i=1}^{m} \beta_i\, h_i(\mathbf{w})$$

The constant β_i are called Lagrange multipliers

The necessary conditions for the existence of an optimal solution $\boldsymbol{w_0}$ corresponding to a minimum of $f(w)$ subject to $h_i(w) = 0,\ i = 1,2,\dots,m$ are:

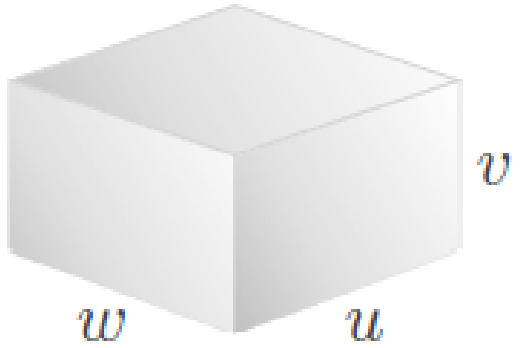
$$\frac{\partial L(\mathbf{w}_\circ, \beta_\circ)}{\partial \mathbf{w}} = \mathbf{0}$$

$$\frac{\partial L(\mathbf{w}_\circ, \beta_\circ)}{\partial \beta_i} = 0$$

This theorem can be used to solve problems with equality constraints!

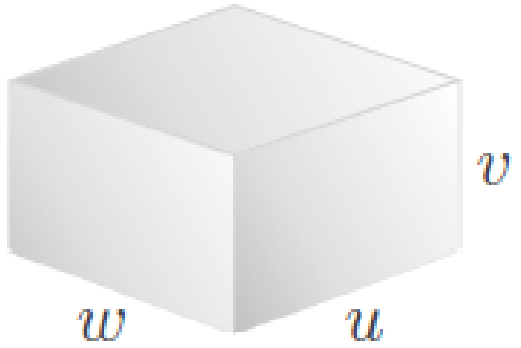
An example

Consider a box whose surface area is C. Formulate the optimization problem for maximizing the volume of the box.



An example

Consider a box whose surface area is C. Formulate the optimization problem for maximizing the volume of the box.



<u>Solution:</u>

Surface area: $2wu + 2uv + 2vw = C$

Volume: wuv

Maximizing wuv is equivalent to minimizing $-wuv$.

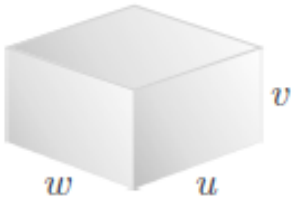
Optimization problem

$$\begin{aligned} \text{Minimize:} \quad & f(u,v,w) = -uvw \\ \text{Subject to:} \quad & h(u,v,w) = wu + uv + vw - \frac{C}{2} = 0 \end{aligned}$$

An example

Minimize: $f(u, v, w) = -uvw$

Subject to: $h(u, v, w) = wu + uv + vw - \frac{C}{2} = 0$

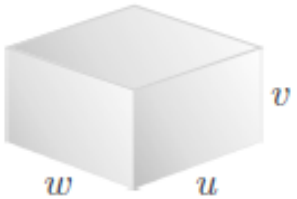


$$L(u, v, w) = -uvw + \beta\left(wu + uv + vw - \frac{C}{2}\right)$$

An example

Minimize: $f(u,v,w) = -uvw$

Subject to: $h(u,v,w) = wu + uv + vw - \frac{C}{2} = 0$



$$L(u,v,w) = -uvw + \beta\left(wu + uv + vw - \frac{C}{2}\right)$$

Lagrange's Theorem yields

$$\begin{aligned}
\frac{\partial L}{\partial u} = -vw + \beta(v+w) &= 0 \\
\frac{\partial L}{\partial v} = -wu + \beta(w+u) &= 0 \\
\frac{\partial L}{\partial w} = -uv + \beta(u+v) &= 0 \\
wu + uv + vw &= \frac{C}{2}
\end{aligned}$$

The solution is

$$u = v = w = \sqrt{\frac{C}{6}}$$

and the maximum volume is

$$\begin{aligned}
&\max(wuv) \\
&= -\min(-wuv) \\
&= -\left(-\left(\frac{C}{6}\right)^{\frac{3}{2}}\right) = \left(\frac{C}{6}\right)^{\frac{3}{2}}
\end{aligned}$$

Generalized Lagrangian Function & KKT conditions

Lagrange Theorem can be generalized to deal with problems having both **equality and inequality** constraints.

Minimize:

$$f(w)$$

Subject to:

$$q_i(w) \leq 0, \quad i = 1,2,\dots,k$$
$$h_i(w) = 0, \quad i = 1,2,\dots,m$$

Generalized Lagrangian Function can be written as:

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

The constants α_i, β_i are called Lagrange multipliers. The solution is characterized by **Karush-Kuhn-Tucker (KKT)** conditions.

Generalized Lagrangian Function & KKT conditions

Kuhn-Tucker Theorem

Consider the primal problem with the Lagrangian as defined earlier. The sufficient and necessary condition for a point $\mathbf{w}_\circ$ to be an optimal solution is the existence of $\alpha_\circ$ and $\beta_\circ$ such that

$$\begin{aligned}
\frac{\partial L\left(\mathbf{w}_\circ, \alpha_\circ, \beta_\circ\right)}{\partial \mathbf{w}} &= \mathbf{0} \\
\frac{\partial L\left(\mathbf{w}_\circ, \alpha_\circ, \beta_\circ\right)}{\partial \beta_i} &= 0 \\
\alpha_{\circ,i}\, q_i\left(\mathbf{w}_\circ\right) &= 0\,, \quad i = 1, \ldots, k \\
q_i\left(\mathbf{w}_\circ\right) &\leq 0\,, \quad i = 1, \ldots, k \\
\alpha_{\circ,i} &\geq 0\,, \quad i = 1, \ldots, k
\end{aligned}$$

These equations are the KKT conditions

WHY KKT? Because this will let us solve the problem by computing the just the inner products of x_i, x_j (which will be very important later on when we want to solve non-linearly separable classification problems)

Primal problem vs dual problem

The KKT conditions enable us to transform the primal problem into an alternative, simpler form called the dual problem.

Primal problem

$$\begin{aligned} \text{Minimize:} \quad & f(\mathbf{w}) \\ \text{Subject to:} \quad & q_i(\mathbf{w}) \leq 0 \\ & h_i(\mathbf{w}) = 0 \end{aligned}$$

Dual problem

$$\begin{aligned} \text{Maximize:} \quad & L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{Subject to:} \quad & \frac{\partial L(\mathbf{w}_\circ, \boldsymbol{\alpha}_\circ, \boldsymbol{\beta}_\circ)}{\partial \mathbf{w}} = 0 \\ & \boldsymbol{\alpha} \geq 0 \end{aligned}$$

The primal and dual problems are equivalent because they have the same value of the optimization problem:

$$f(\mathbf{w}_\circ) = L(\mathbf{w}_\circ, \boldsymbol{\alpha}_\circ, \boldsymbol{\beta}_\circ)$$

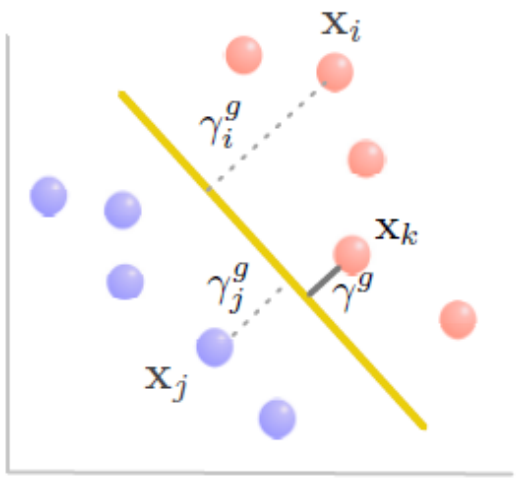
Please check: Practical Methods of Optimization by Fletcher, Wiley, 1987

Let’s go back to SVM…

The KKT conditions enable us to transform the primal problem into an alternative, simpler form called the dual problem

Let's recall our primal problem:

$$
\begin{aligned}
\text{Given data set}: &\quad S = \{(\mathbf{x}_i, d_i)\} \\
\text{Find}: &\quad \mathbf{w} \text{ and } b \\
\text{Minimizing}: &\quad f(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \\
\text{Subject to}: &\quad d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1
\end{aligned}
$$



In primal problem, known parameters are x_i, d_i and unknown variables are w, b.

Apply KKT conditions to reduce unknowns to just α_i, and to form a dual problem.

Transforming primal problem into dual problem by KKT conditions

Step 1: Write constraint $d_i(w^T x_i + b) \geq 1$ in a standard form as:

$$-d_i(w^T x_i + b) + 1 \leq 0$$

Step 2: Write the Lagrangian function

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i \left(d_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i \end{aligned}$$

Step 3: KKT conditions are:

$$\begin{aligned} &\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{0} \\ &\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \\ &d_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \\ &\alpha_i \left(d_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right) = 0 \\ &\alpha_i \geq 0 \end{aligned}$$

Let's focus on each condition one by one

Transforming primal problem into dual problem by KKT conditions

The first condition:

$$
\begin{aligned}
\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(\frac{\mathbf{w}^T \mathbf{w}}{2} - \sum_{i=1}^{N} \alpha_i \, d_i \, \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^{N} \alpha_i \, d_i + \sum_{i=1}^{N} \alpha_i \right) \\
&= \frac{1}{2} \frac{\partial \left(\mathbf{w}^T \mathbf{w} \right)}{\partial \mathbf{w}} - \sum_{i=1}^{N} \alpha_i d_i \left(\frac{\partial \left(\mathbf{w}^T \mathbf{x}_i \right)}{\partial \mathbf{w}} \right) \\
&= \frac{2\mathbf{w}}{2} - \sum_{i=1}^{N} \alpha_i d_i \left(\frac{\partial \left(\mathbf{x}_i^T \mathbf{w} \right)}{\partial \mathbf{w}} \right) \quad (\text{Since } \mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u}) \\
&= \mathbf{w} - \sum_{i=1}^{N} \alpha_i d_i \mathbf{x}_i \qquad \left(\text{Since } \frac{\partial \left(\mathbf{u}^T \mathbf{v} \right)}{\partial \mathbf{v}} = \mathbf{u} \right) \\
&= \mathbf{0}
\end{aligned}
$$

Therefore:

$$
\mathbf{w} = \sum_{i=1}^{N} \alpha_i \, d_i \, \mathbf{x}_i
$$

Transforming primal problem into dual problem by KKT conditions

The second condition:

$$\begin{aligned}
&\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} \\
= \quad &\frac{\partial}{\partial b}\left(\frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i\, d_i\, \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N}\alpha_i\, d_i + \sum_{i=1}^{N}\alpha_i\right) \\
= \quad &\sum_{i=1}^{N}\alpha_i\, d_i = 0
\end{aligned}$$

Therefore:

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

Transforming primal problem into dual problem by KKT conditions

Let's recall the Lagrangian function

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i\, d_i\, \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N} \alpha_i\, d_i + \sum_{i=1}^{N} \alpha_i$$

And remember:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i\, d_i\, \mathbf{x}_i$$

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

So:

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}\left[\sum_{i=1}^{N} \alpha_i\, d_i\, \mathbf{x}_i^T\right]\left[\sum_{j=1}^{N} \alpha_j\, d_j\, \mathbf{x}_j\right] = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j\, d_i\, d_j\, \mathbf{x}_i^T\mathbf{x}_j$$

$$\sum_{i=1}^{N} \alpha_i\, d_i\, \mathbf{w}^T\mathbf{x}_i = \sum_{i=1}^{N} \alpha_i\, d_i\left[\sum_{j=1}^{N} \alpha_j\, d_j\, \mathbf{x}_j^T\right]\mathbf{x}_i = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\, \alpha_j\, d_i\, d_j\, \mathbf{x}_i^T\mathbf{x}_j$$

Transforming primal problem into dual problem by KKT conditions

- The Lagrangian function can be written as:

$$
\begin{aligned}
L(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i \left(d_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right) \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i \\
&= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j \\
&\equiv Q(\boldsymbol{\alpha})
\end{aligned}
$$

Finding optimal hyperplane (dual problem)

$$
\begin{aligned}
\text{Given :} \quad & S = \{(\mathbf{x}_i, d_i)\} \\
\text{Find :} \quad & \text{Lagrange multipliers } \{\alpha_i\} \\
\text{Maximizing :} \quad & Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j \\
\text{Subject to :} \quad & (1) \quad \sum_{i=1}^{N} \alpha_i d_i = 0 \\
& (2) \quad \alpha_i \geq 0
\end{aligned}
$$

Primal Problem

$$
\begin{aligned}
\text{Given data set :} \quad & S = \{(\mathbf{x}_i, d_i)\} \\
\text{Find :} \quad & \mathbf{w} \text{ and } b \\
\text{Minimizing :} \quad & f(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \\
\text{Subject to :} \quad & d_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1
\end{aligned}
$$

1. Algorithms and software available for solving this problem.
2. α_i are the only unknowns.
3. $x_i^T x_j^T$ is called a linear kernel

An example

Suppose that a support vector machine is to be constructed using the training set below:

a) Determine the Lagrangian function

b) Show the explicit form of the dual problem

i	$\mathbf{x}_i$	d_i
1	$[\ 1 \quad -1\ ]^T$	-1
2	$[\ 2 \quad 1\ ]^T$	1
3	$[\ 3 \quad 1\ ]^T$	1

An example

a) Lagrangian function

- Let's remember the general formulation of Lagrangian function.

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i \left(d_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right)$$

- From the given data, we have N = 3, and $w = [w_1 \ , w_2 \ ]^T$

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\begin{bmatrix} w_1 & w_2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\ &\quad - \alpha_1(-1)\begin{bmatrix} w_1 & w_2 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} - \alpha_1(-1)b + \alpha_1 \\ &\quad - \alpha_2(1)\begin{bmatrix} w_1 & w_2 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} - \alpha_2(1)b + \alpha_2 \\ &\quad - \alpha_3(1)\begin{bmatrix} w_1 & w_2 \end{bmatrix}\begin{bmatrix} 3 \\ 1 \end{bmatrix} - \alpha_3(1)b + \alpha_3 \\ &= \frac{1}{2}\left(w_1^2 + w_2^2\right) + \alpha_1\left(w_1 - w_2 + 1\right) - \alpha_2\left(2w_1 + w_2 - 1\right) \\ &\quad -\alpha_3\left(3w_1 + w_2 - 1\right) + \left(\alpha_1 - \alpha_2 - \alpha_3\right)b \end{aligned}$$

An example

b) Explicit form of Dual problem

Finding optimal hyperplane (dual problem)

Given : $S = \{(\mathbf{x}_i, d_i)\}$

Find : Lagrange multipliers $\{\alpha_i\}$

Maximizing : $Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$

Subject to : (1) $\sum_{i=1}^{N} \alpha_i d_i = 0$

(2) $\alpha_i \geq 0$

Maximize:

$$Q(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}(2\alpha_1^2 + 5\alpha_2^2 + 10\alpha_3^2 - 2\alpha_1\alpha_2 - 4\alpha_1\alpha_3 + 14\alpha_2\alpha_3)$$

Subject to:

$$\sum_{i=1}^{3} \alpha_i d_i = \alpha_1(-1) + \alpha_2(1) + \alpha_3(1) = -\alpha_1 + \alpha_2 + \alpha_3 = 0$$

$$\alpha_1 \geq 0$$

$$\alpha_2 \geq 0$$

$$\alpha_3 \geq 0$$

Lagrange multiplier and support vector

Support vectors are the data points that lie closest to the decision

Recall:

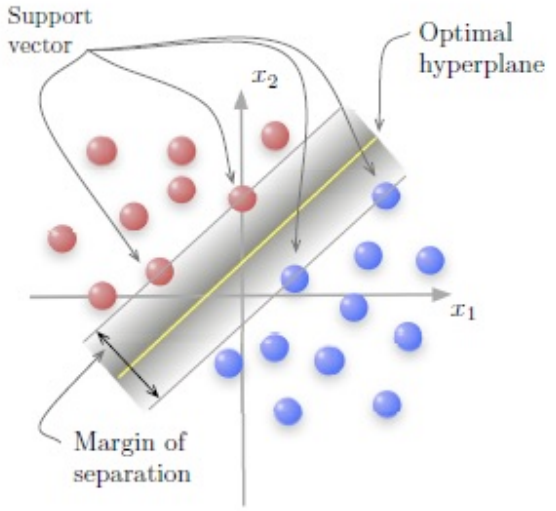
For support vector x_i: $g(x_i) = w_0^T x_i + b = \pm 1$

From the 3rd KKT condition : $d_i(w_0^T x_i + b_0) \geq 1$

From the 4th KKT condition: $\alpha_{0,i}\,(d_i(w_0^T x_i + b_0) - 1) = 0$

For data point x_i that is a support vector:

$$d_i(w_0^T x_i + b_0) = 1 \text{ AND } \alpha_{0,i} \neq 0$$



Most of the α_i values will turn out to have the value zero. The non-zero α_i will correspond to the support vectors. For a support vector x_i:

$$d_i(w_0^T x_i + b_0) = 1 \quad \Rightarrow \quad b_0 = \frac{1}{d_i} - w_0^T x_i$$

Optimal solution

	Primal	Dual
Find :	$\mathrm{w}\,, b$	α_i
Minimizing :	$f(\mathrm{w})$	$\mathbf{w}$
Maximizing :	$-$	$Q(\boldsymbol{\alpha})$
Subject to :	$d_i\,(\mathrm{w}^T\mathrm{x}_i + b) \geq 1$	$\sum_{i=1}^{N} \alpha_i d_i = 0$
		$\alpha_i \geq 0$

$$f\,(\mathrm{w}, \xi) = \frac{1}{2}\mathrm{w}^T\mathrm{w} \qquad Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j\,d_i\,d_j\,\mathrm{x}_i^T\mathrm{x}_j$$

After $\alpha_{\circ,i}$ is obtained, we can calculate $\mathbf{w}_\circ$ and $b_\circ$ as follows:

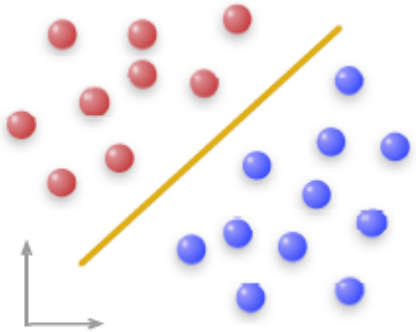
$$\mathbf{w}_\circ = \sum_{i=1}^{N}\alpha_{\circ,i}\,d_i\,\mathbf{x}_i\,, \quad b_\circ = \frac{1}{d^{(s)}} - \mathbf{w}_\circ^T\mathbf{x}^{(s)}$$

where $\mathbf{x}^{(s)}$ is a support vector with label $d^{(s)}$

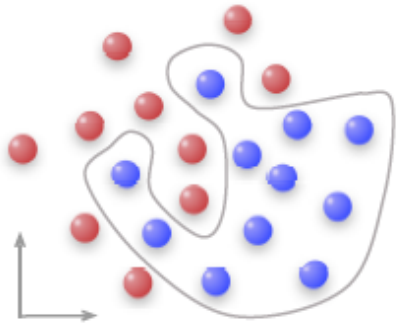
Linearly separable, or not?

Two classes of data are linearly separable if and only if there exists a hyperplane $w^T x + b = 0$ that separates the two classes.

Example in 2D:



Linearly separable



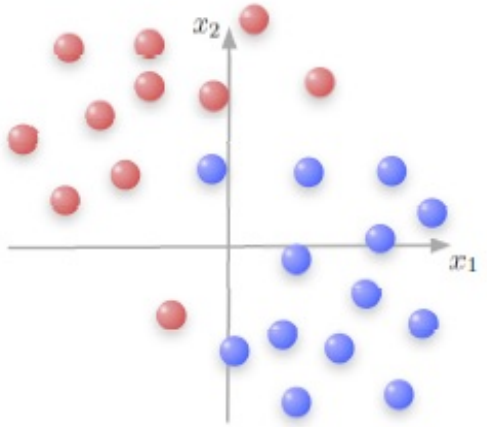
Non-linearly separable

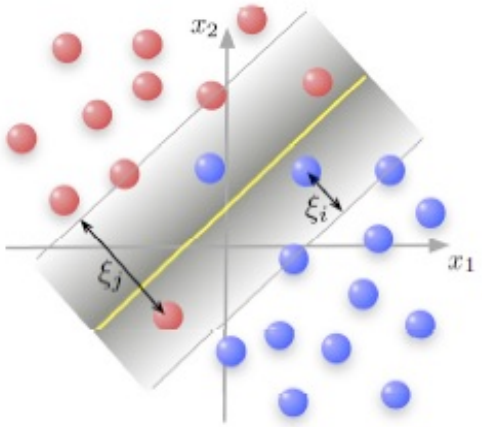
We'll now study the case of "non-linearly separable".

Not linearly separable case:

Option 1:

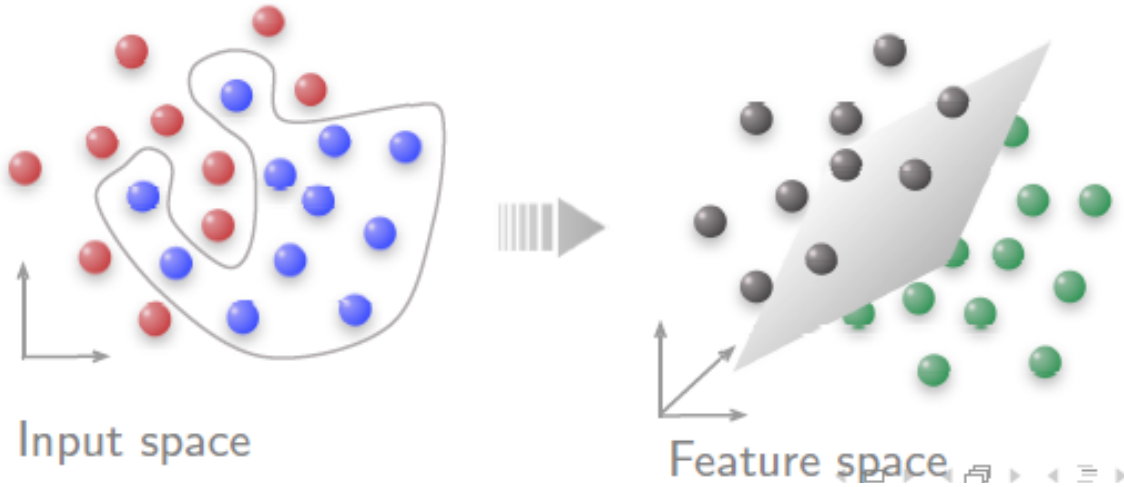
Find the optimal hyperplane to minimize classification error (soft margin)





Option 2:

Transform the data into higher dimensional space (kernel)

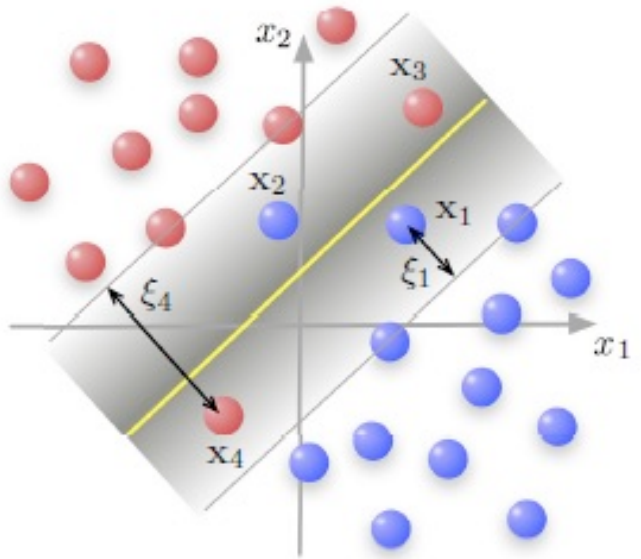


Option 1

Soft margin for minimizing classification error

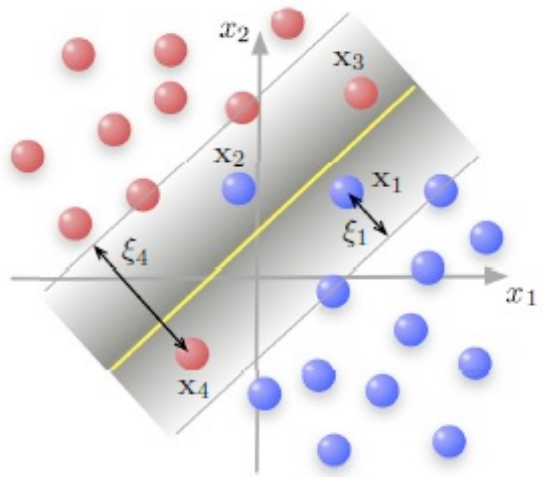
- Nonnegative slack variables: $\xi_i,\ \mathrm{i} = 1,2,3, \ldots \mathrm{N}$

  - Data point not in region of separation ξ_i=0
  - Data point in region of separation and on correct side of hyperplane $0 \leq \xi_i \leq 1$ (ξ_1, ξ_3)
  - Data point in region of separation but on wrong side of hyperplane $\xi_i > 1$ (ξ_2, ξ_4)



Soft margin for minimizing classification error

Optimal hyperplane must minimize the error penalty: $\sum_{i=1}^{N} \xi_i$



Minimize: $\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$

Subject to:

(i) $\quad d_i(w^T x_i + b) - 1 + \xi_i \geq 0$

(ii) $\quad \xi_i \geq 0$

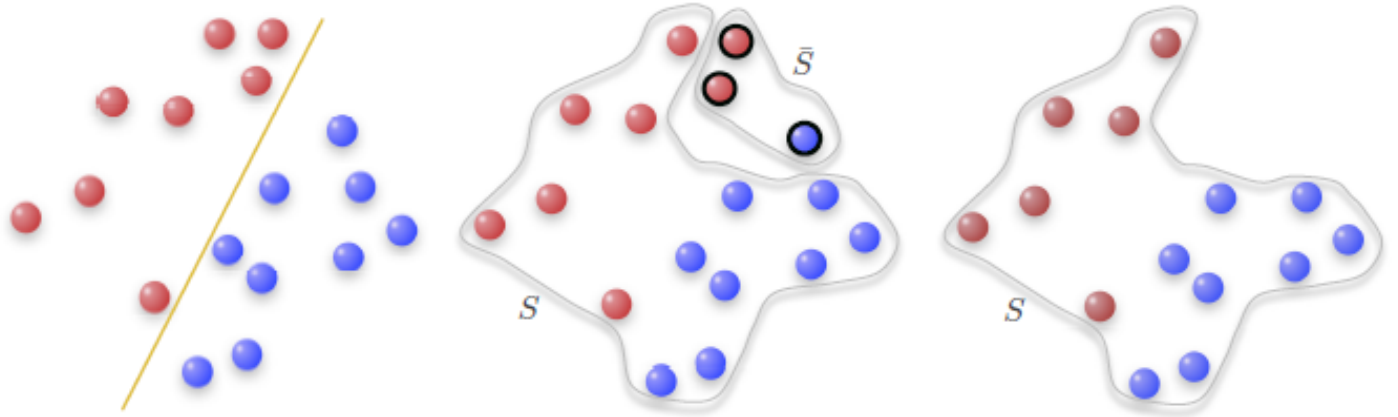
Primal Problem (after applying KKT conditions, Dual problem can be obtained)

Note 1: For $\xi_i = 0$, constraint is the same as that in basic formulation (which is also known as hard margin classification problem).

Note 2: Value of C > 0 represents the cost of violating constraints. A large C generally leads to smaller margin but also fewer misclassification of training data.

Soft Margin and Performance of SVM

- Performance of SVM should be evaluated based on the number of misclassifications for both training and testing data.
- SVM that classifies S $(training\ data)$ perfectly may not be the desired solution to the classification problem.



Data set Σ (assumed separable here for simple illustration)

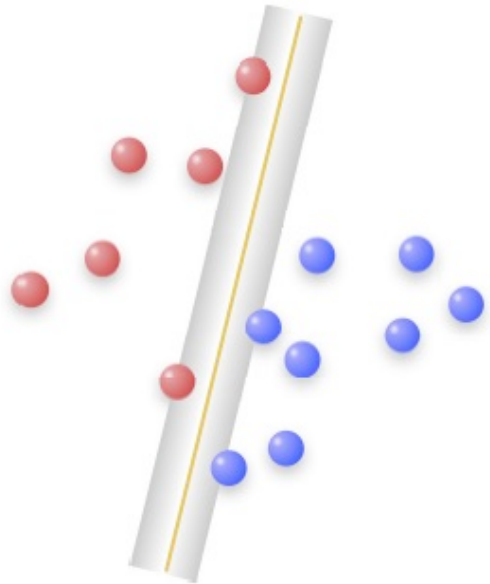
Divide Σ into training set S and test set $\bar{S}$

Use S to find $(\mathbf{w}_o, b_o)$ for SVM

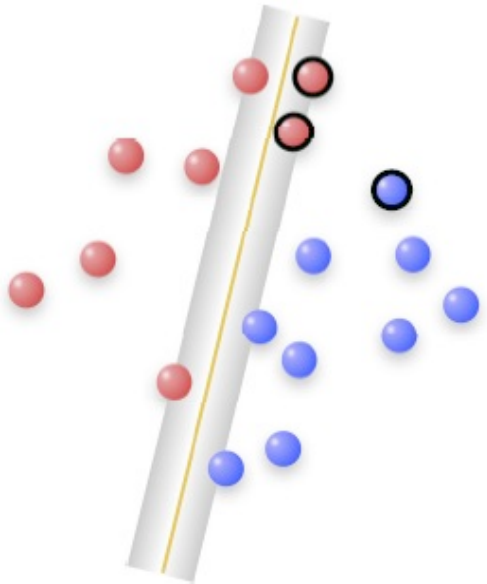
Soft Margin and Performance of SVM

- Optimal hyperplane with **hard margin** classifies all examples in training set S correctly, but misclassifies 2 examples in the test set.

- Total misclassification = 2



Training data

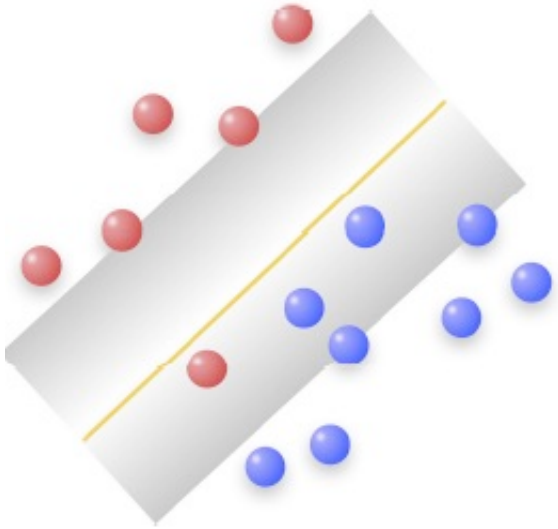


Training & test data

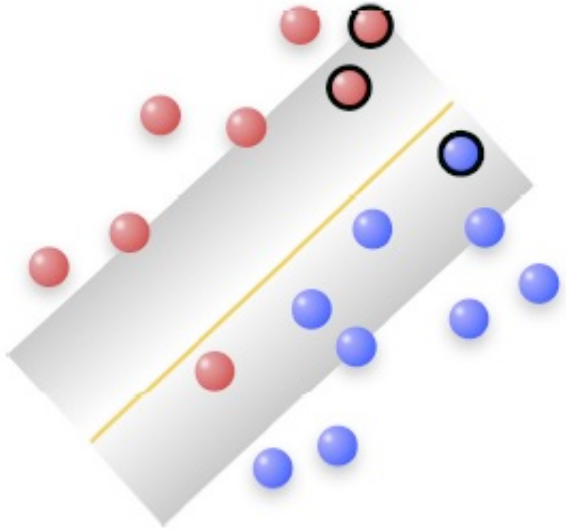
Soft Margin and Performance of SVM

- An optimal hyperplane with soft margin misclassifies 1 example in training set S, but it classifies correctly all examples in test set.

- Total misclassification = 1

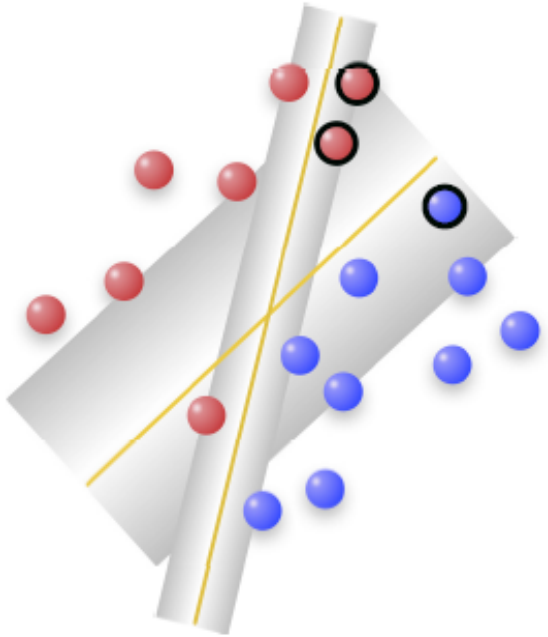


Not perfect for training data…



Good for test data!

SVM that classifies training set S perfectly may not be the desired solution to the classification problem…



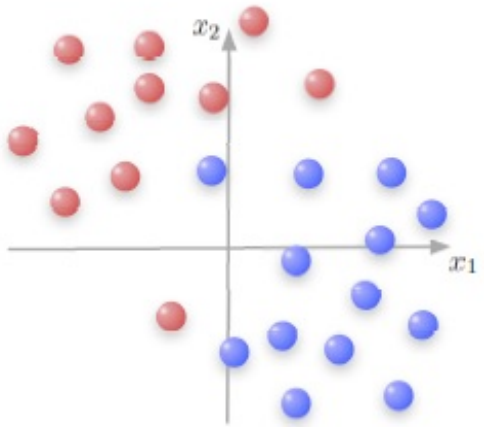
Summary of misclassifications:

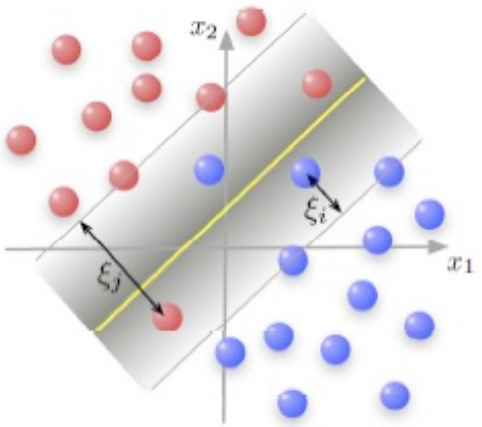
	Training	Test	
Type of SVM	S	$\bar{S}$	Σ
With hard margin	0	2	2
With soft margin	1	0	1

Recall:
Not linearly separable case

Option 1:

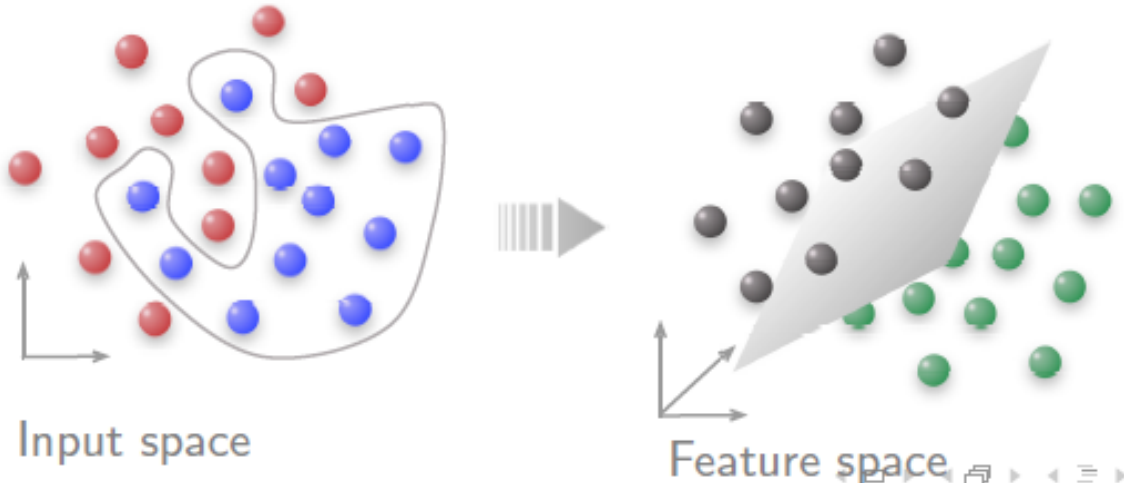
Find the optimal hyperplane to minimize classification error (soft margin)





Option 2:

Transform the data into higher dimensional space (kernel)

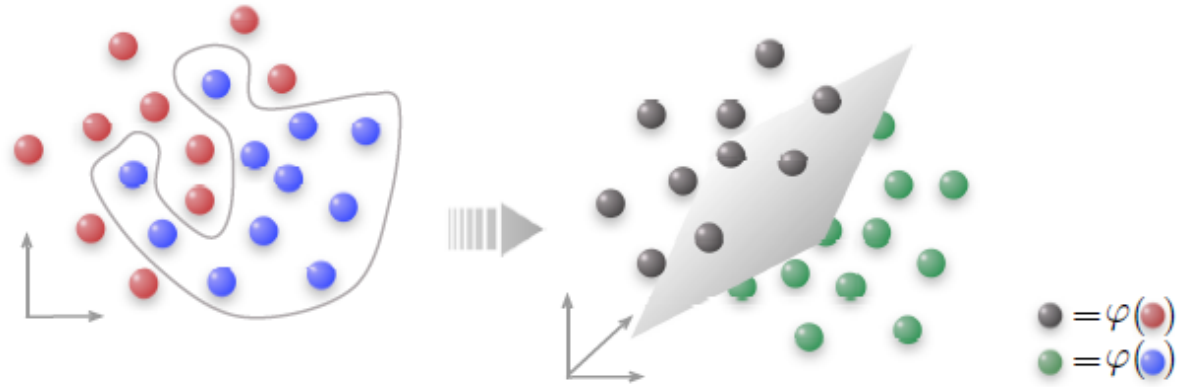


Option 2

Transformation of data into higher dimension space

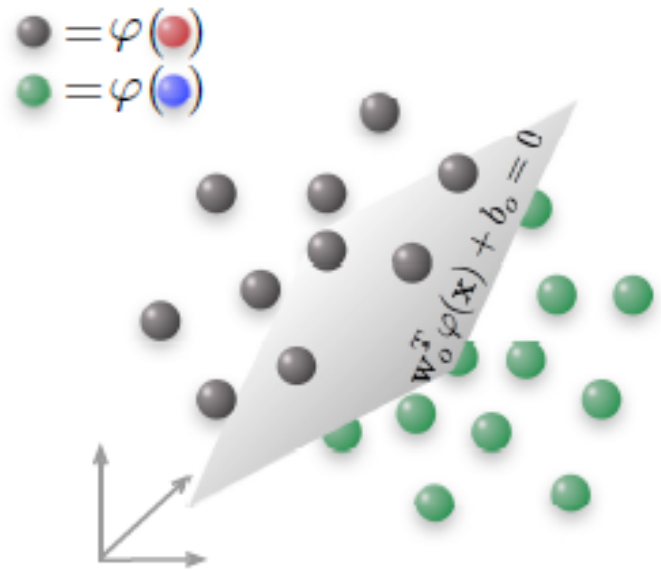
Cover's Theorem

Probability that classes are linearly separable increases when data points in input space are nonlinearly mapped to a higher dimensional feature space.



Transformation of data into higher dimension space

Optimal hyperplane in a feature space looks like:



$$g(x) = (w_0^T \varphi(x) + b_0) = 0$$

Transformation of data into higher dimension space

Example:

A non-separable dataset in 1-dimensional space:

We can map data x from 1 D to 2 D by (x, x^2)

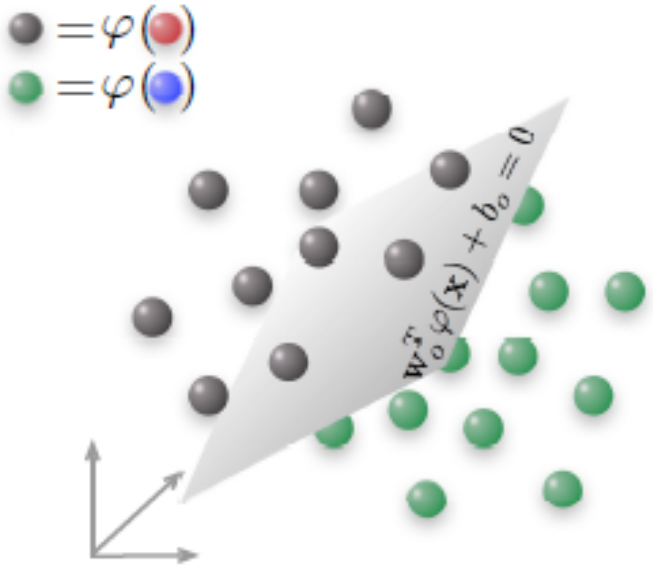
Kernel:
Transformation of data into higher dimension space

For training data x_i:

Classification will be as follows:

$g(x_i) = (w_0^T \varphi(x_i) + b_0) \geq 1$ for $d_i = +1$

$g(x_i) = (w_0^T \varphi(x_i) + b_0) \leq -1$ for $d_i = -1$



Kernel:
Transformation of data into higher dimension space

Lagrangian function can be written as

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i \left(d_i \left(\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) + b\right) - 1\right)$$

$$= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i d_i \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) - b\sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i$$

From KKT conditions: $\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \boldsymbol{\varphi}(\mathbf{x}_i)$ and $\sum_{i=1}^{N} \alpha_i d_i = 0$

So $\mathbf{w}^T\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_j)$

Let $Q(\boldsymbol{\alpha}) \equiv L(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \underbrace{\boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_j)}_{K(\mathbf{x}_i, \mathbf{x}_j)}$

Kernel:
Transformation of data into higher dimension space

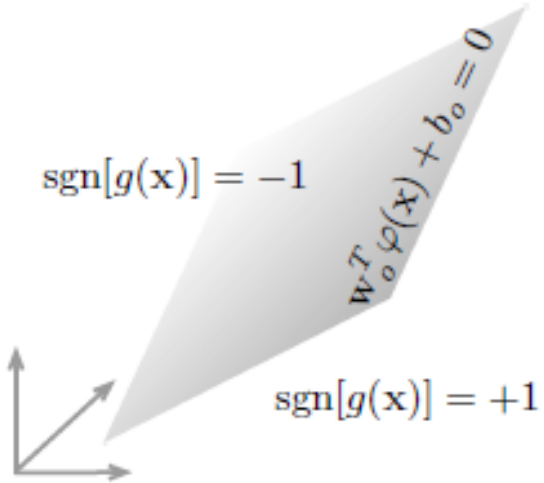
- Kernel can be written as:

$$\text{Kernel:}\quad \underbrace{K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)}_{\text{symmetric}} = \varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j) = \varphi^T(\mathbf{x}_j)\varphi(\mathbf{x}_i)$$

Given optimal value $\alpha_{o,i}$

$$\begin{cases} \mathbf{w}_o = \sum_{i=1}^{N} \alpha_{o,i}\, d_i \varphi(\mathbf{x}_i) \\ b_o = \frac{1}{d^{(s)}} - \mathbf{w}_o^T \varphi(\mathbf{x}^{(s)}) \\ (\mathbf{x}^{(s)} \text{ is a SV with label } d^{(s)}) \end{cases}$$

Almost identical to hard margin case!
$\boldsymbol{x}$ is now $\boldsymbol{\varphi(x)}$



How to find the kernel or $\varphi(x)$?

Dual formulation

Finding optimal hyperplane (dual problem)

Given : $S = \{(\mathbf{x}_i, d_i)\}$

Find : Lagrange multipliers $\{\alpha_i\}$

Maximizing : $Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$

Subject to : (1) $\sum_{i=1}^{N} \alpha_i d_i = 0$

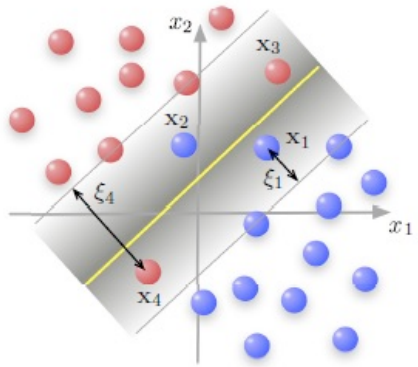
(2) $\alpha_i \geq 0$

Dual problem with soft margin

Find : α_i

Maximize : $Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$

Subject to : $\sum_{i=1}^{N} \alpha_i d_i = 0\,,\ 0 \leq \alpha_i \leq C$



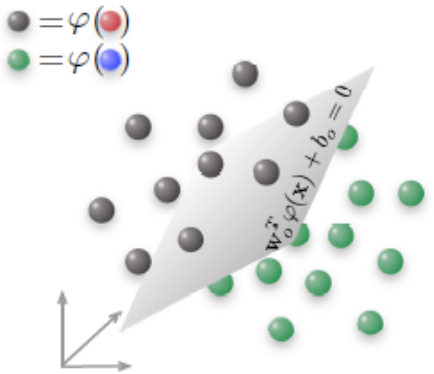
Dual problem with soft margin and transformation

Find : α_i

Maximize : $Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \boldsymbol{\varphi}^T(\mathbf{x}_i) \boldsymbol{\varphi}(\mathbf{x}_j)$

Subject to : $\sum_{i=1}^{N} \alpha_i d_i = 0\,,\ 0 \leq \alpha_i \leq C$

KERNEL



Kernel Trick

Solution requires finding $\boldsymbol{\varphi}(\quad)$ but finding the explicit form of $\boldsymbol{\varphi}(\quad)$ is difficult.

Kernel Trick: Find expression for K(·,·).

$$K(x, x') = \varphi(x)\varphi(x') = \exp(-||x - x'||/2)$$

An example: Radial basis function kernel (or RBF kernel)

- The radial basis function kernel is special in many ways. For example, running the SVM with such a kernel function will always be able to return you a separable solution provided the training examples are all distinct.
- *You will have an implementation homework to examine this.*

Properties of Kernels

- A kernel function is valid if and only if there exists some feature mapping $\varphi(x)$ such that $K(x, x')= \varphi(x)\varphi(x')$.
- We don't need to know what $\varphi(x)$ is (necessarily). We can build many common kernel functions based only on the following four rules:
  1. $\mathrm{K}(\mathrm{x}, \mathrm{x}')$=1 is a kernel function.
  2. If $K(x, x')$ is a kernel, then $\mathrm{f}(\mathrm{x})K(x, x')f(x')$ is also a kernel function (assuming that $\mathrm{f}(\mathrm{x})$ is a real valued function of x).
  3. If $K_1(x, x')$ and $K_2(x', x)$ are kernels, then their sum is also a kernel.
  4. If $K_1(x, x')$and $K_2(x', x)$ are kernels, then their product is also a kernel.

Common Kernels

- Polynomials of degree d

$$K(x, x') = (\mathrm{x}.\,\mathrm{x}')^d$$

- Gaussian Kernel

$$K(x, x') = \exp\left(\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

- And many others! (attractive research field)

True Risk and Empirical Risk

- Empirical risk minimization (ERM) is a principle in statistical learning theory which defines a family of learning algorithms and is used to give theoretical bounds on their performance.
- The core idea is that we cannot know exactly how well an algorithm will work in practice (the true "risk") because we don't know the true distribution of data that the algorithm will work on.
- What can we do?
  - We can instead measure its performance on a known set of training data (the "empirical" risk).

What you should know?

- The intuition, where to find software
- Why do we use SVM? What is margin?
- Primal vs Dual Problem
- Optimization with Lagrange Theorem / KKT conditions
- How to handle non separable data
  - Slack variables (soft margin)
  - Kernels – new feature space
  - Their primal and dual formulations

MUST: Please study the lecture notes on SVM.

Suggestion: If you're not very clear or want to learn more, please do read:

SVM part of "C. Bishop: Pattern Recognition and Machine Learning. Springer, 2006" *(recommended text book).* It will help you to understand the concept.

Thank you!

Acknowledgement

The following courses have been partially used:

- National University of Singapore – Pattern Recognition Module (EE5907R)
- National University of Singapore – Neural Networks Module (EE5904R)
- University of Edinburgh, United Kingdom – Machine Learning And Pattern Recognition (MLPR, INFR11130)
- Stanford University – Machine Learning (CS229)