The SVM approach

- We have briefly discussed Support Vector Machine (SVM) idea.

The SVM approach

- We have briefly discussed Support Vector Machine (SVM) idea.
- The idea is to map the feature vectors nonlinearly into another space and learn a linear classifier there.

The SVM approach

- We have briefly discussed Support Vector Machine (SVM) idea.
- The idea is to map the feature vectors nonlinearly into another space and learn a linear classifier there.
- The linear classifier in this new space would be an appropriate nonlinear classifier in the original space.

- Recall the simple example we saw earlier.

- Recall the simple example we saw earlier.
- Let $X = [x_1 \quad x_2]$

- Recall the simple example we saw earlier.
- Let $X = [x_1 \quad x_2]$ and let $\phi : \Re^2 \longrightarrow \Re^5$ given by $Z = \phi(X) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_2^2 \quad x_1 x_2]$

- Recall the simple example we saw earlier.
- Let $X = [x_1 \quad x_2]$ and let $\phi : \Re^2 \longrightarrow \Re^5$ given by
  $Z = \phi(X) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_2^2 \quad x_1 x_2]$
- Now,
  $$g(X) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_2^2 + a_5 x_1 x_2$$
  is a quadratic discriminant function in $\Re^2$;

- Recall the simple example we saw earlier.
- Let $X = [x_1 \quad x_2]$ and let $\phi : \Re^2 \longrightarrow \Re^5$ given by
  $Z = \phi(X) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_2^2 \quad x_1 x_2]$
- Now,
  $$g(X) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_2^2 + a_5 x_1 x_2$$
  is a quadratic discriminant function in $\Re^2$; but
  $$g(Z) = a_0 + a_1 z_1 + a_2 z_2 + a_3 z_3 + a_4 z_4 + a_5 z_5$$
  is a linear dscriminant function in the '$\phi(X)$' space.

- There are two major issues in naively using this idea.

- There are two major issues in naively using this idea.
- If we want, e.g., p^{th} degree polynomial discriminant function in the original feature space ($\Re^m$), then the transformed feature vector, Z, has dimension $O(m^p)$.

- There are two major issues in naively using this idea.
- If we want, e.g., p^{th} degree polynomial discriminant function in the original feature space ($\Re^m$), then the transformed feature vector, Z, has dimension $O(m^p)$.
- Results in huge computational cost both for learning and and final operation of the classifier.

- There are two major issues in naively using this idea.
- If we want, e.g., p^{th} degree polynomial discriminant function in the original feature space ($\Re^m$), then the transformed feature vector, Z, has dimension $O(m^p)$.
- Results in huge computational cost both for learning and and final operation of the classifier.
- We need to learn $O(m^p)$ parameters rather than $O(m)$ parameters. Hence may need much larger number of examples for achieving proper generalization.

- There are two major issues in naively using this idea.
- If we want, e.g., p^{th} degree polynomial discriminant function in the original feature space ($\Re^m$), then the transformed feature vector, Z, has dimension $O(m^p)$.
- Results in huge computational cost both for learning and and final operation of the classifier.
- We need to learn $O(m^p)$ parameters rather than $O(m)$ parameters. Hence may need much larger number of examples for achieving proper generalization.
- SVM offers an elegant solution to both.

Support Vector Machines

- Learning of **optimal** hyperplane.

Support Vector Machines

- Learning of **optimal** hyperplane.
  - Separating hyperplane that maximizes separation between Classes.

Support Vector Machines

- Learning of **optimal** hyperplane.
  - Separating hyperplane that maximizes separation between Classes.
- *Effectively* maps original feature vectors into a high dimensional space. Hence learns nonlinear discriminant functions.

Support Vector Machines

- Learning of **optimal** hyperplane.
  - Separating hyperplane that maximizes separation between Classes.
- *Effectively* maps original feature vectors into a high dimensional space. Hence learns nonlinear discriminant functions.
- By using **Kernel function** we never need to explicitly calculate the mapping.

Support Vector Machines

- Learning of **optimal** hyperplane.
  - Separating hyperplane that maximizes separation between Classes.
- *Effectively* maps original feature vectors into a high dimensional space. Hence learns nonlinear discriminant functions.
- By using **Kernel function** we never need to explicitly calculate the mapping.
- We need solve only a quadratic optimization problem.

Support Vector Machines

- Learning of **optimal** hyperplane.
  - Separating hyperplane that maximizes separation between Classes.
- *Effectively* maps original feature vectors into a high dimensional space. Hence learns nonlinear discriminant functions.
- By using **Kernel function** we never need to explicitly calculate the mapping.
- We need solve only a quadratic optimization problem.
- Now we formulate the SVM method, first for linearly separable case.

- Training set: $\{(X_i, y_i), \;\; i = 1, \ldots, n\}$, $X_i \in \Re^m$, $y_i \in \{+1, \;\; -1\}$.

- Training set:
  $\{(X_i, y_i), \;\; i = 1, \ldots, n\}$, $X_i \in \Re^m$, $y_i \in \{+1, \;\; -1\}$.
- To start with, assume training set is linearly separable. That is, exist $W \in \Re^m$ and $b \in \Re$ such that

$$
\begin{aligned}
W^T X_i + b &> 0, \quad \forall i \;\; s.t. \;\; y_i = +1 \\
W^T X_i + b &< 0, \quad \forall i \;\; s.t. \;\; y_i = -1
\end{aligned}
$$

(Note both inequalities are strict)

- Training set:
$\{(X_i, y_i), \;\; i = 1, \ldots, n\}$, $X_i \in \Re^m$, $y_i \in \{+1, \;\; -1\}$.
- To start with, assume training set is linearly separable. That is, exist $W \in \Re^m$ and $b \in \Re$ such that

$$\begin{aligned} W^T X_i + b &> 0, \quad \forall i \;\; s.t. \;\; y_i = +1 \\ W^T X_i + b &< 0, \quad \forall i \;\; s.t. \;\; y_i = -1 \end{aligned}$$

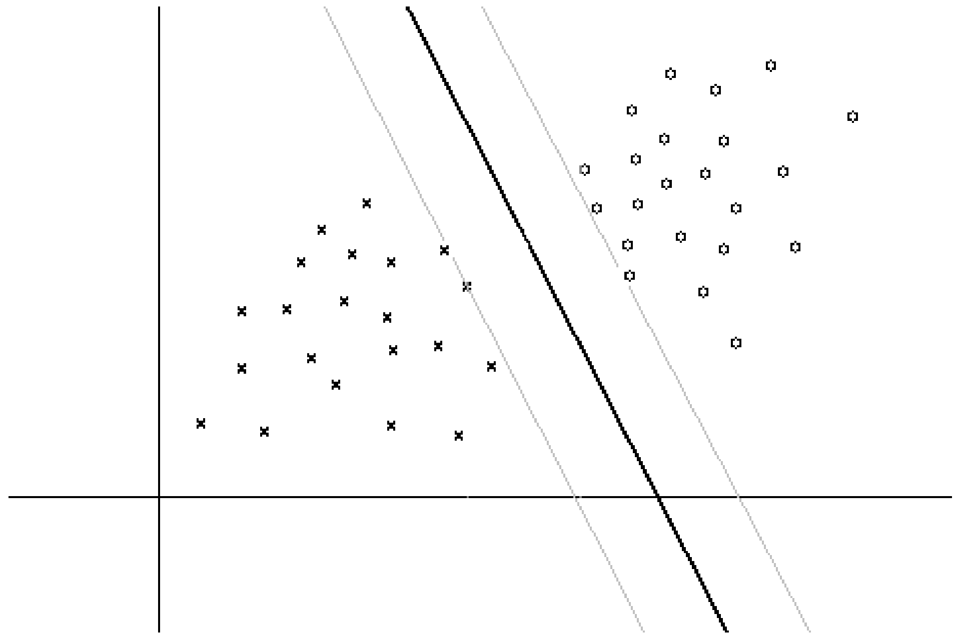
(Note both inequalities are strict)
- $W^T X + b = 0$ – A separating hyperplane.

- Training set:
  $\{(X_i, y_i), \;\; i = 1, \ldots, n\}$, $X_i \in \Re^m$, $y_i \in \{+1, \;\; -1\}$.
- To start with, assume training set is linearly separable. That is, exist $W \in \Re^m$ and $b \in \Re$ such that

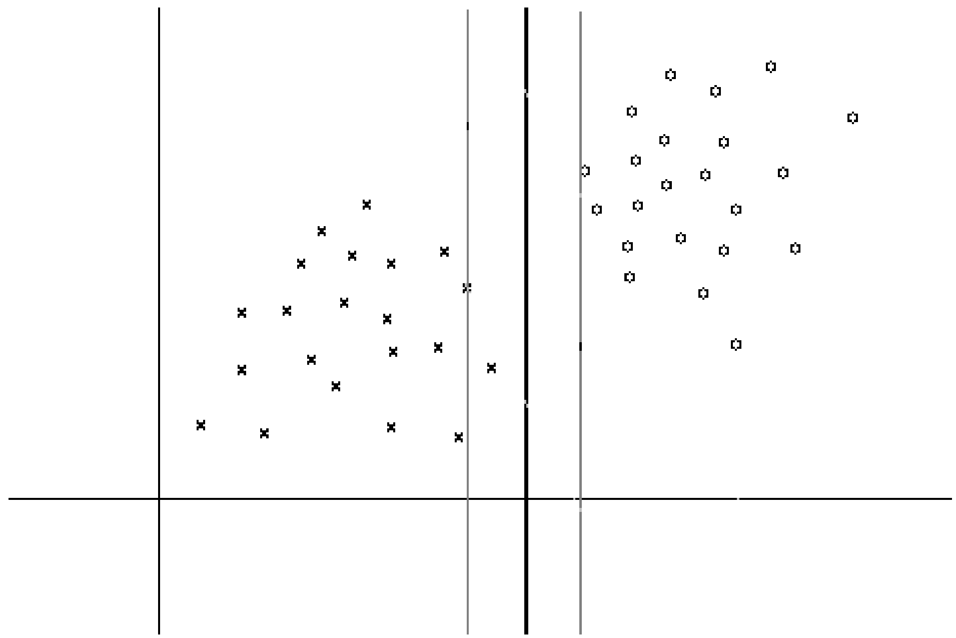
$$\begin{aligned} W^T X_i + b &> 0, \quad \forall i \;\; s.t. \;\; y_i = +1 \\ W^T X_i + b &< 0, \quad \forall i \;\; s.t. \;\; y_i = -1 \end{aligned}$$

  (Note both inequalities are strict)
- $W^T X + b = 0$ – A separating hyperplane.
- Infinitely many separating hyperplanes exist.

A good separating hyperplane



Another separating hyperplane



- Recall that we assume training set is linearly separable and hence

$$
\begin{aligned}
W^T X_i + b &> 0, \quad \forall i \;\; s.t. \;\; y_i = +1 \\
W^T X_i + b &< 0, \quad \forall i \;\; s.t. \;\; y_i = -1
\end{aligned}
$$

- Recall that we assume training set is linearly separable and hence

$$
\begin{aligned}
W^T X_i + b &> 0, \quad \forall i \;\; s.t. \;\; y_i = +1 \\
W^T X_i + b &< 0, \quad \forall i \;\; s.t. \;\; y_i = -1
\end{aligned}
$$

- Since the training set is finite, $\exists \epsilon > 0 \; s.t.$

$$
\begin{aligned}
W^T X_i + b &\geq \epsilon, \quad \forall i \;\; s.t. \;\; y_i = +1 \\
W^T X_i + b &\leq -\epsilon, \quad \forall i \;\; s.t. \;\; y_i = -1
\end{aligned}
$$

- Hence, we can scale $W,\ b$ such that

$$\begin{aligned} W^T X_i + b &\geq +1 \text{ if } y_i = +1 \\ W^T X_i + b &\leq -1 \text{ if } y_i = -1 \end{aligned}$$

- Hence, we can scale $W, \; b$ such that

$$\begin{aligned} W^T X_i + b \; & \geq +1 \;\text{ if }\; y_i = +1 \\ W^T X_i + b \; & \leq -1 \;\text{ if }\; y_i = -1 \end{aligned}$$

or, equivalently

$$y_i(W^T X_i + b) \geq 1, \;\; \forall i.$$

(Recall that $y_i \in \{+1, \quad -1\}$)

- When the training set is separable, any separating hyperplane, $W, \ b$, can be scaled to satisfy

$$y_i(W^T X_i + b) \geq 1, \quad \forall i.$$

- When the training set is separable, any separating hyperplane, $W, \ b$, can be scaled to satisfy

$$y_i(W^T X_i + b) \geq 1, \quad \forall i.$$

- Then there are no training patterns between the two parallel hyperplanes

$$W^T X + b = +1$$

and

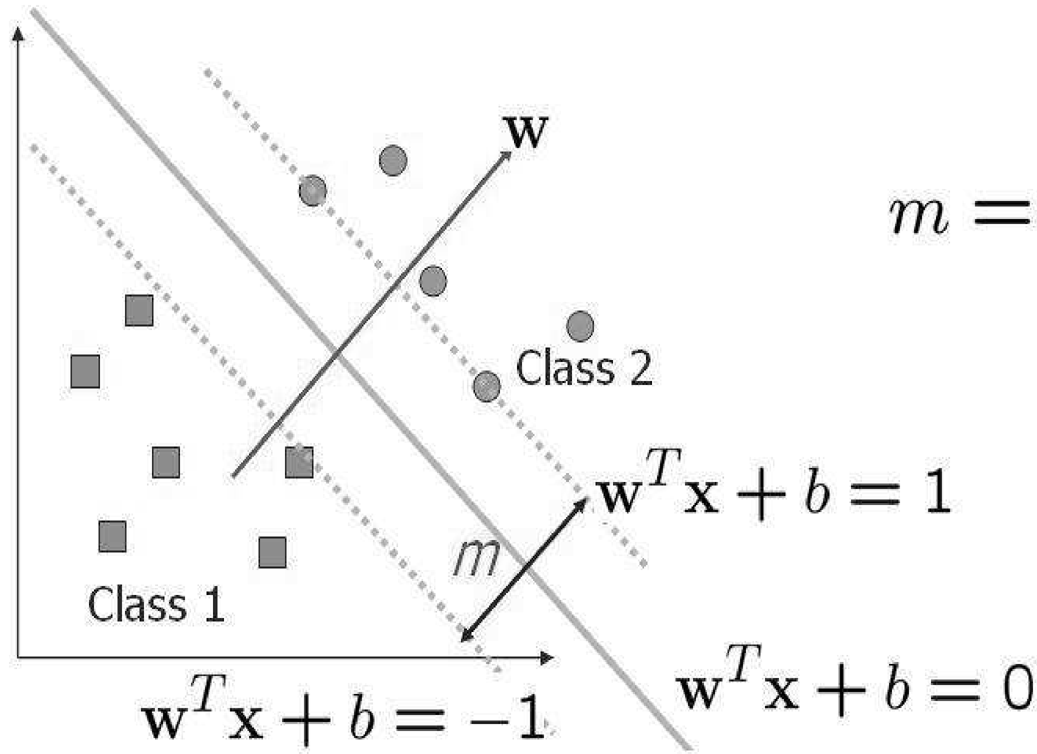
$$W^T X + b = -1$$

.

Optimal hyperplane

- Distance between these two hyperplanes is: $\frac{2}{||W||}$. Called **margin** of the separating hyperplane.

Margin of a hyperplane

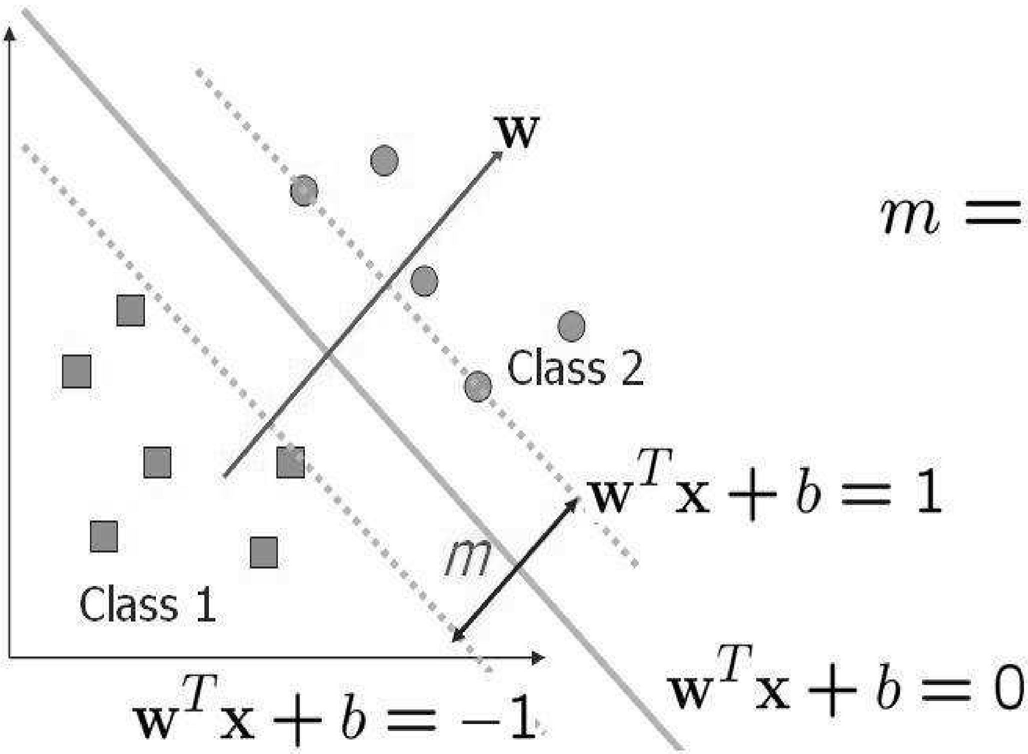


$$m = \frac{2}{||\mathbf{w}||}$$

Optimal hyperplane

- Distance between these two hyperplanes is: $\frac{2}{||W||}$. Called **margin** of the separating hyperplane.
- Hence distance between the hyperplane and the closest pattern is $\frac{1}{||W||}$.

Margin of a hyperplane



$$m = \frac{2}{||\mathbf{w}||}$$

Optimal hyperplane

- Distance between these two hyperplanes is: $\frac{2}{||W||}$. Called **margin** of the separating hyperplane.
- Hence distance between the hyperplane and the closest pattern is $\frac{1}{||W||}$.
- Intuitively, more the margin, better is the chance of correct classification of new patterns.

Optimal hyperplane

- Distance between these two hyperplanes is: $\frac{2}{||W||}$. Called **margin** of the separating hyperplane.
- Hence distance between the hyperplane and the closest pattern is $\frac{1}{||W||}$.
- Intuitively, more the margin, better is the chance of correct classification of new patterns.
- **Optimal Hyperplane** – separating hyperplane with maximum margin.

The optimization problem

- Among all separating hyperplanes, the one with largest margin is the optimal hyperplane.

The optimization problem

- Among all separating hyperplanes, the one with largest margin is the optimal hyperplane.
- So, the optimal hyperplane is a solution to the following optimization problem.

The optimization problem

- Among all separating hyperplanes, the one with largest margin is the optimal hyperplane.
- So, the optimal hyperplane is a solution to the following optimization problem.
- Find $W \in \Re^m$, $b \in \Re$ to

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} W^T W \\ \text{subject to} \quad & y_i(W^T X_i + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

The optimization problem

- Among all separating hyperplanes, the one with largest margin is the optimal hyperplane.
- So, the optimal hyperplane is a solution to the following optimization problem.
- Find $W \in \Re^m$, $b \in \Re$ to

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} W^T W \\ \text{subject to} \quad & y_i (W^T X_i + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

- This is a constrained optimization problem with quadratic cost function and linear inequality constraints.

Constrained Optimization

- Consider the following optimization problem

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{a}_j^T \mathbf{x} + b_j \leq 0, \quad j = 1, \ldots, r \end{array}$$

where $f : \Re^m \rightarrow \Re$ is a continuously differentiable function, and
$\mathbf{a}_j \in \Re^m$, $b_j \in \Re$, $j = 1, \cdots, r$.

Constrained Optimization

- Consider the following optimization problem

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{a}_j^T \mathbf{x} + b_j \leq 0, \quad j = 1, \ldots, r \end{array}$$

  where $f : \Re^m \rightarrow \Re$ is a continuously differentiable function, and
  $\mathbf{a}_j \in \Re^m$, $b_j \in \Re$, $j = 1, \cdots, r$.

- A point, $\mathbf{x} \in \Re^m$, is called a **feasible** point (for this problem) if $\mathbf{a}_j^T \mathbf{x} + b_j \leq 0$, $j = 1, \cdots, r$.

- Any $\mathbf{x}^* \in \Re^m$ is called a **local minimum** of the problem if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ that is feasible and is in a small neighbourhood of $\mathbf{x}^*$.

- Any $\mathbf{x}^* \in \Re^m$ is called a **local minimum** of the problem if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ that is feasible and is in a small neighbourhood of $\mathbf{x}^*$.
- If $f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \Re^m$ and $\mathbf{x}$ feasible, then $\mathbf{x}^*$ is a **global minimum.**

- Any $\mathbf{x}^* \in \Re^m$ is called a **local minimum** of the problem if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ that is feasible and is in a small neighbourhood of $\mathbf{x}^*$.
- If $f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \Re^m$ and $\mathbf{x}$ feasible, then $\mathbf{x}^*$ is a **global minimum.**
- Unlike in unconstrained optimization, here we need to minimize only over the feasible set.

- Here we would consider only the case where f is a convex function.

- Here we would consider only the case where f is a convex function.
- $f : \Re^m \longrightarrow \Re$ is said to be a convex function if for all $\mathbf{x}_1, \mathbf{x}_2 \in \Re^m$ and for all $\alpha \in (0,\ 1)$,

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$$

- Here we would consider only the case where f is a convex function.
- $f : \Re^m \rightarrow \Re$ is said to be a convex function if for all $\mathbf{x}_1, \mathbf{x}_2 \in \Re^m$ and for all $\alpha \in (0,\ 1)$,

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$$

- For example, $f(\mathbf{x}) = \mathbf{x}^T \mathbf{x}$ is a convex function.

- Here we would consider only the case where f is a convex function.
- $f : \Re^m \rightarrow \Re$ is said to be a convex function if for all $\mathbf{x}_1, \mathbf{x}_2 \in \Re^m$ and for all $\alpha \in (0,\ 1)$,

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$$

- For example, $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$ is a convex function.
- When f is convex, in our optimization problem, every local minimum is also a global minimum.

- We now look at one method of sloving the constrained optimization problem.

- We now look at one method of sloving the constrained optimization problem.
- Given our optimization problem, define

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{r} \mu_j (\mathbf{a}_j^T \mathbf{x} + b_j)$$

- We now look at one method of sloving the constrained optimization problem.
- Given our optimization problem, define

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{r} \mu_j (\mathbf{a}_j^T \mathbf{x} + b_j)$$

- The L is called the Lagrangian of the problem and the μ_j are called the Lagrange multipliers.

- We now look at one method of sloving the constrained optimization problem.
- Given our optimization problem, define

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{r} \mu_j (\mathbf{a}_j^T \mathbf{x} + b_j)$$

- The L is called the Lagrangian of the problem and the μ_j are called the Lagrange multipliers.
- Essentially, the constrained optimization problem can be solved through unconstrained optimization of L.

Kuhn-Tucker Conditions

- Consider the optimization problem with f convex.

Kuhn-Tucker Conditions

- Consider the optimization problem with f convex.
- Any $\mathbf{x}^*$ is a global minimum if and only if $\mathbf{x}^*$ is feasible and there exist $\mu_j^*, \ j = 1, \cdots, r$, such that

Kuhn-Tucker Conditions

- Consider the optimization problem with f convex.
- Any $\mathbf{x}^*$ is a global minimum if and only if $\mathbf{x}^*$ is feasible and there exist $\mu_j^*,\ j = 1, \cdots, r$, such that

1. $\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}^*) = 0$

Kuhn-Tucker Conditions

- Consider the optimization problem with f convex.
- Any $\mathbf{x}^*$ is a global minimum if and only if $\mathbf{x}^*$ is feasible and there exist $\mu_j^*,\ j = 1, \cdots, r$, such that

1. $\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}^*) = 0$
2. $\mu_j^* \geq 0, \ \ \forall j$

Kuhn-Tucker Conditions

- Consider the optimization problem with f convex.
- Any $\mathbf{x}^*$ is a global minimum if and only if $\mathbf{x}^*$ is feasible and there exist $\mu_j^*,\ j = 1, \cdots, r$, such that

1. $\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}^*) = 0$
2. $\mu_j^* \geq 0, \ \ \forall j$
3. $\mu_j^*(\mathbf{a}_j^T \mathbf{x}^* + b_j) = 0, \ \forall j$

Kuhn-Tucker Conditions

- Consider the optimization problem with f convex.
- Any $\mathbf{x}^*$ is a global minimum if and only if $\mathbf{x}^*$ is feasible and there exist $\mu_j^*,\ j = 1, \cdots, r$, such that
  1. $\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}^*) = 0$
  2. $\mu_j^* \geq 0, \ \ \forall j$
  3. $\mu_j^*(\mathbf{a}_j^T \mathbf{x}^* + b_j) = 0, \ \forall j$
- These are the so called Kuhn-Tucker conditions for our optimization problem with convex cost function and linear constraints.

- We can use the above conditions to obtain a $\mathbf{x}^*$ which is a minimum of the optimization problem.

- We can use the above conditions to obtain a $\mathbf{x}^*$ which is a minimum of the optimization problem.
- We can also solve the constrained optimization problem using the so called dual of this problem.

- We can use the above conditions to obtain a $\mathbf{x}^*$ which is a minimum of the optimization problem.
- We can also solve the constrained optimization problem using the so called dual of this problem.
- This is the approach taken in SVM algorithm.

- We can use the above conditions to obtain a $\mathbf{x}^*$ which is a minimum of the optimization problem.
- We can also solve the constrained optimization problem using the so called dual of this problem.
- This is the approach taken in SVM algorithm.
- Duality is an important concept in optimization.

- We can use the above conditions to obtain a $\mathbf{x}^*$ which is a minimum of the optimization problem.
- We can also solve the constrained optimization problem using the so called dual of this problem.
- This is the approach taken in SVM algorithm.
- Duality is an important concept in optimization.
- Here we discuss only one way of formulating the dual which is useful when the objective function is convex and constraints are linear.

- Our optimization problem is

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{a}_j^T\mathbf{x} + b_j \leq 0, \quad j = 1, \ldots, r \end{array}$$

where $f : \Re^m \rightarrow \Re$ is a continuously differentiable convex function, and

$\mathbf{a}_j \in \Re^m$, $b_j \in \Re$, $j = 1, \cdots, r$.

- Our optimization problem is

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{a}_j^T \mathbf{x} + b_j \leq 0, \quad j = 1, \ldots, r \end{array}$$

  where $f : \Re^m \rightarrow \Re$ is a continuously differentiable convex function, and

  $\mathbf{a}_j \in \Re^m$, $b_j \in \Re$, $j = 1, \cdots, r$.

- This is known as the **primal** problem.

- Our optimization problem is

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{a}_j^T\mathbf{x} + b_j \leq 0, \quad j = 1, \ldots, r \end{array}$$

  where $f : \Re^m \rightarrow \Re$ is a continuously differentiable convex function, and

  $\mathbf{a}_j \in \Re^m$, $b_j \in \Re$, $j = 1, \cdots, r$.
- This is known as the **primal** problem.
- Here the optimization variables are $\mathbf{x} \in \Re^m$.

- Recall that the Lagrangian is

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{r} \mu_j(\mathbf{a}_j^T \mathbf{x} + b_j)$$

Here, $\mathbf{x} \in \Re^m$ and $\boldsymbol{\mu} \in \Re^r$.

- Recall that the Lagrangian is

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{r} \mu_j(\mathbf{a}_j^T\mathbf{x} + b_j)$$

Here, $\mathbf{x} \in \Re^m$ and $\boldsymbol{\mu} \in \Re^r$.

- Define the *dual function*, $q : \Re^r \rightarrow [-\infty,\ \infty)$ by

$$q(\boldsymbol{\mu}) = \inf_{\mathbf{x}}\ L(\mathbf{x}, \boldsymbol{\mu})$$

- Recall that the Lagrangian is

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{r} \mu_j(\mathbf{a}_j^T\mathbf{x} + b_j)$$

Here, $\mathbf{x} \in \Re^m$ and $\boldsymbol{\mu} \in \Re^r$.

- Define the *dual function*, $q : \Re^r \to [-\infty,\ \infty)$ by

$$q(\boldsymbol{\mu}) = \inf_{\mathbf{x}}\ L(\mathbf{x}, \boldsymbol{\mu})$$

- If for a particular $\boldsymbol{\mu}$, if the infimum is not attained then $q(\boldsymbol{\mu})$ would take value $-\infty$.

The Dual problem

- The **dual** problem is:

$$\begin{array}{ll} \text{maximize} & q(\boldsymbol{\mu}) \\ \text{subject to} & \mu_j \geq 0, \quad j = 1, \ldots, r \end{array}$$

The Dual problem

- The **dual** problem is:

$$\begin{array}{ll} \text{maximize} & q(\boldsymbol{\mu}) \\ \text{subject to} & \mu_j \geq 0, \quad j = 1, \ldots, r \end{array}$$

- This is also a constrained optimization problem.

The Dual problem

- The **dual** problem is:

$$\begin{array}{ll} \text{maximize} & q(\boldsymbol{\mu}) \\ \text{subject to} & \mu_j \geq 0, \quad j = 1, \ldots, r \end{array}$$

- This is also a constrained optimization problem.
- Here the optimization is over $\Re^r$ and $\boldsymbol{\mu} \in \Re^r$ are the optimization variables.

The Dual problem

- The **dual** problem is:

$$\begin{array}{ll} \text{maximize} & q(\boldsymbol{\mu}) \\ \text{subject to} & \mu_j \geq 0, \quad j = 1, \ldots, r \end{array}$$

- This is also a constrained optimization problem.
- Here the optimization is over $\Re^r$ and $\boldsymbol{\mu} \in \Re^r$ are the optimization variables.
- There is a nice connection between the primal and dual problems.

Primal-Dual Relationship

- Now we have the following.

Primal-Dual Relationship

- Now we have the following.

  1. If the primal has a solution so does the dual and the optimal values are equal.

Primal-Dual Relationship

- Now we have the following.
  1. If the primal has a solution so does the dual and the optimal values are equal.
  2. $\mathbf{x}^*$ is optimal for primal and $\boldsymbol{\mu}^*$ is optimal for dual if and only if

Primal-Dual Relationship

- Now we have the following.
  1. If the primal has a solution so does the dual and the optimal values are equal.
  2. $\mathbf{x}^*$ is optimal for primal and $\boldsymbol{\mu}^*$ is optimal for dual if and only if
     - $\mathbf{x}^*$ is feasible for primal and $\boldsymbol{\mu}^*$ is feasible for dual,

Primal-Dual Relationship

- Now we have the following.

1. If the primal has a solution so does the dual and the optimal values are equal.
2. $\mathbf{x}^*$ is optimal for primal and $\boldsymbol{\mu}^*$ is optimal for dual if and only if
   - $\mathbf{x}^*$ is feasible for primal and $\boldsymbol{\mu}^*$ is feasible for dual,
   - $f(\mathbf{x}^*) = L(\mathbf{x}^*,\ \boldsymbol{\mu}^*) = \min_{\mathbf{x}}\ L(\mathbf{x},\ \boldsymbol{\mu}^*)$.

Primal-Dual Relationship

- Now we have the following.
  1. If the primal has a solution so does the dual and the optimal values are equal.
  2. $\mathbf{x}^*$ is optimal for primal and $\boldsymbol{\mu}^*$ is optimal for dual if and only if
     - $\mathbf{x}^*$ is feasible for primal and $\boldsymbol{\mu}^*$ is feasible for dual,
     - $f(\mathbf{x}^*) = L(\mathbf{x}^*,\ \boldsymbol{\mu}^*) = \min_{\mathbf{x}}\ L(\mathbf{x},\ \boldsymbol{\mu}^*)$.
- We would be using the dual formulation for the optimization problem in SVM

The optimization problem for SVM

- The optimal hyperplane is a solution of the following constrained optimization problem.

The optimization problem for SVM

- The optimal hyperplane is a solution of the following constrained optimization problem.
- Find $W \in \Re^m$, $b \in \Re$ to

$$\begin{array}{rl} \text{minimize} & \frac{1}{2} W^T W \\ \text{subject to} & 1 - y_i(W^T X_i + b) \leq 0, \quad i = 1, \ldots, n \end{array}$$

The optimization problem for SVM

- The optimal hyperplane is a solution of the following constrained optimization problem.
- Find $W \in \Re^m$, $b \in \Re$ to

$$\begin{array}{rl} \text{minimize} & \frac{1}{2} W^T W \\ \text{subject to} & 1 - y_i(W^T X_i + b) \leq 0, \quad i = 1, \ldots, n \end{array}$$

- Quadratic cost function and linear (inequality) constraints.

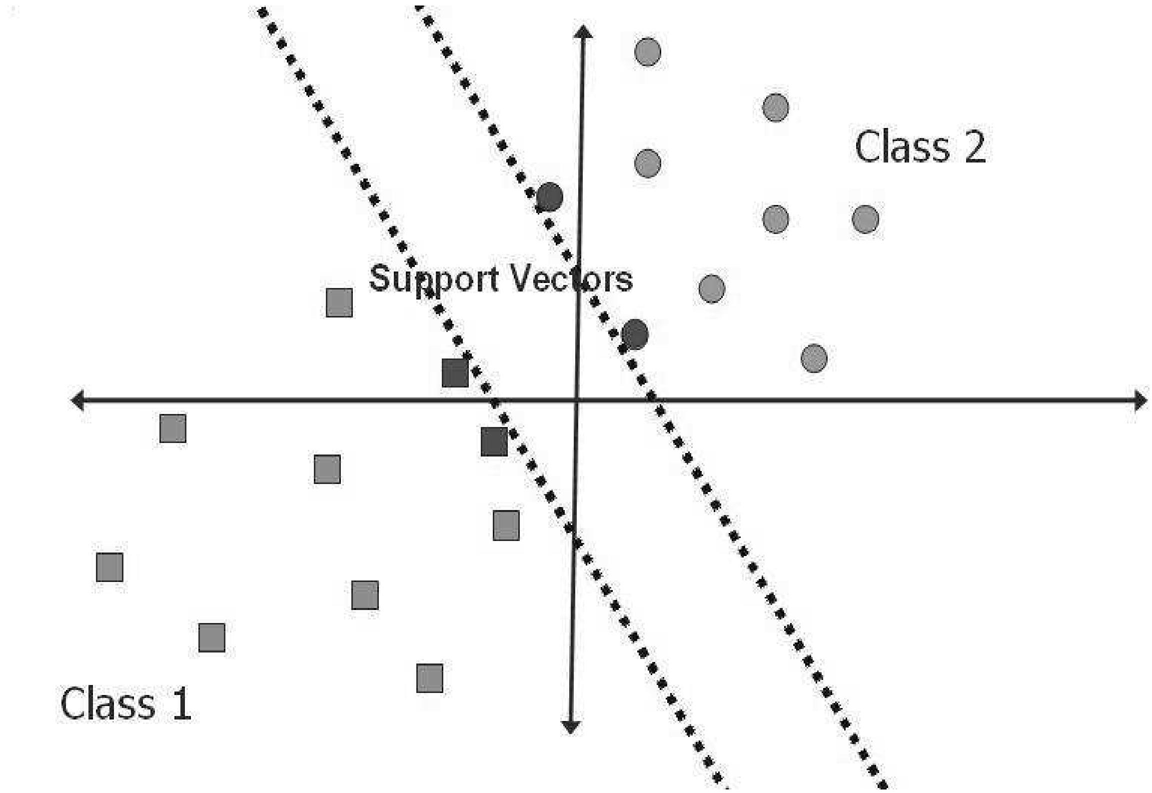
The optimization problem for SVM

- The optimal hyperplane is a solution of the following constrained optimization problem.
- Find $W \in \Re^m$, $b \in \Re$ to

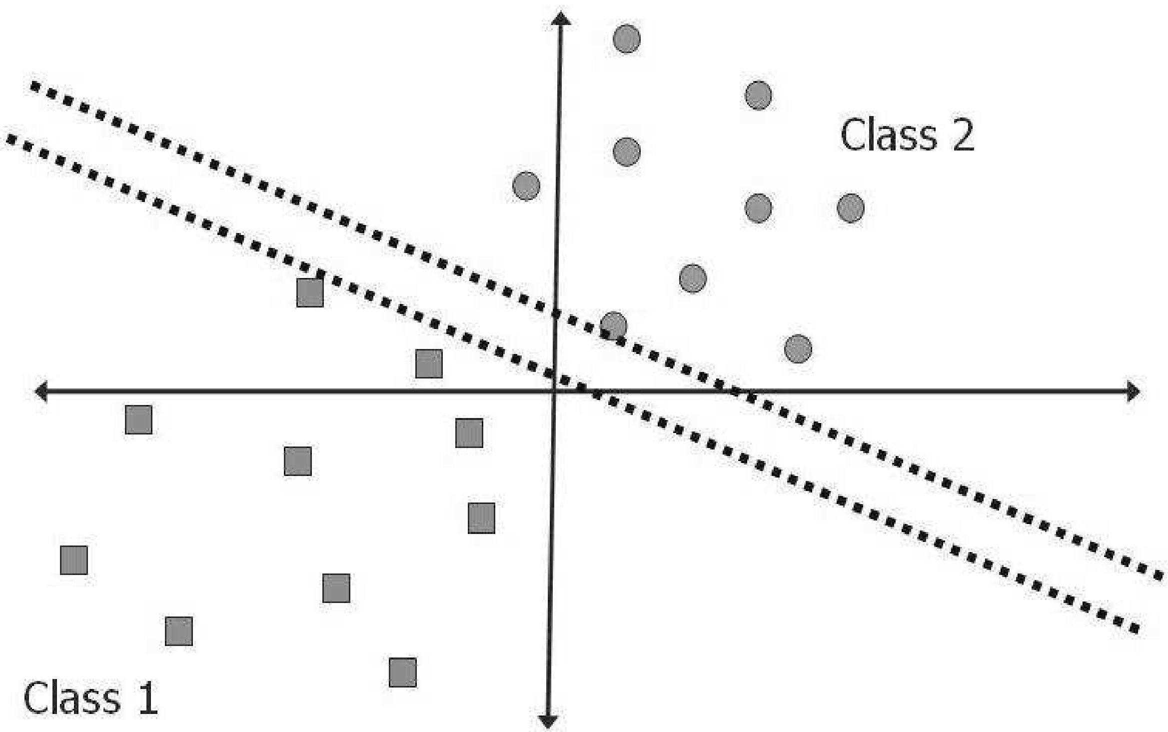
$$\begin{array}{rl} \text{minimize} & \frac{1}{2} W^T W \\ \text{subject to} & 1 - y_i(W^T X_i + b) \leq 0, \quad i = 1, \ldots, n \end{array}$$

- Quadratic cost function and linear (inequality) constraints.
- Kuhn-Tucker conditions are necessary and sufficient. Every local minimum is global minimum.

Optimal hyperplane



Non-optimal hyperplane



- The Lagrangian is given by

$$L(W,\ b,\ \boldsymbol{\mu}) = \frac{1}{2}W^TW + \sum_{i=1}^{n} \mu_i \left[1 - y_i(W^TX_i + b)\right]$$

- The Lagrangian is given by

$$L(W,\ b,\ \boldsymbol{\mu}) = \frac{1}{2}W^T W + \sum_{i=1}^{n} \mu_i \left[1 - y_i(W^T X_i + b)\right]$$

- The Kuhn-Tucker conditions give

- The Lagrangian is given by

$$L(W,\ b,\ \boldsymbol{\mu}) = \frac{1}{2}W^TW + \sum_{i=1}^{n}\ \mu_i\ [1 - y_i(W^TX_i + b)]$$

- The Kuhn-Tucker conditions give

$$\nabla_W L = 0 \Rightarrow W^* = \sum_{i=1}^{n}\ \mu_i^* y_i X_i$$

- The Lagrangian is given by

$$L(W,\ b,\ \boldsymbol{\mu}) = \frac{1}{2}W^TW + \sum_{i=1}^{n} \mu_i \left[1 - y_i(W^TX_i + b)\right]$$

- The Kuhn-Tucker conditions give

$$\nabla_W L = 0 \Rightarrow W^* = \textstyle\sum_{i=1}^{n}\ \mu_i^* y_i X_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \textstyle\sum_{i=1}^{n} \mu_i^* y_i = 0$$

- The Lagrangian is given by

$$L(W,\ b,\ \boldsymbol{\mu}) = \frac{1}{2}W^TW + \sum_{i=1}^{n}\ \mu_i\ [1 - y_i(W^TX_i + b)]$$

- The Kuhn-Tucker conditions give

$$\nabla_W L = 0 \Rightarrow W^* = \sum_{i=1}^{n}\ \mu_i^* y_i X_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \mu_i^* y_i = 0$$

$$1 - y_i(X_i^T W^* + b^*) \leq 0,\ \ \forall i$$

- The Lagrangian is given by

$$L(W,\ b,\ \boldsymbol{\mu}) = \frac{1}{2}W^TW + \sum_{i=1}^{n} \mu_i \left[1 - y_i(W^TX_i + b)\right]$$

- The Kuhn-Tucker conditions give

$\nabla_W L = 0 \Rightarrow W^* = \sum_{i=1}^n \ \mu_i^* y_i X_i$

$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \mu_i^* y_i = 0$

$1 - y_i(X_i^T W^* + b^*) \leq 0, \ \ \forall i$

$\mu_i^* \geq 0, \ \ \& \ \ \mu_i^*[1 - y_i(X_i^T W^* + b^*)] = 0, \ \ \forall i$

- Let $S = \{i \mid \mu_i^* > 0\}$.

- Let $S = \{i \mid \mu_i^* > 0\}$.
- By complementary slackness condition,

$$i \in S \quad \Rightarrow \quad y_i(X_i^T W^* + b^*) = 1$$

- Let $S = \{i \mid \mu_i^* > 0\}$.
- By complementary slackness condition,

$$i \in S \quad \Rightarrow \quad y_i(X_i^T W^* + b^*) = 1$$

Implies X_i is closest to separating hyperplane.

- Let $S = \{i \mid \mu_i^* > 0\}$.
- By complementary slackness condition,

$$i \in S \quad \Rightarrow \quad y_i(X_i^T W^* + b^*) = 1$$

  Implies X_i is closest to separating hyperplane.
- $\{X_i \mid i \in S\}$ are called Support vectors.

- Let $S = \{i \mid \mu_i^* > 0\}$.
- By complementary slackness condition,

$$i \in S \quad \Rightarrow \quad y_i(X_i^T W^* + b^*) = 1$$

Implies X_i is closest to separating hyperplane.

- $\{X_i \mid i \in S\}$ are called Support vectors. We have

$$W^* = \sum_i \mu_i^* y_i X_i = \sum_{i \in S} \mu_i^* y_i X_i$$

- Let $S = \{i \mid \mu_i^* > 0\}$.
- By complementary slackness condition,

$$i \in S \quad \Rightarrow \quad y_i(X_i^T W^* + b^*) = 1$$

  Implies X_i is closest to separating hyperplane.
- $\{X_i \mid i \in S\}$ are called Support vectors. We have

$$W^* = \sum_i \mu_i^* y_i X_i = \sum_{i \in S} \mu_i^* y_i X_i$$

- Optimal W is a linear combination of Support vectors.

- Let $S = \{i \mid \mu_i^* > 0\}$.
- By complementary slackness condition,

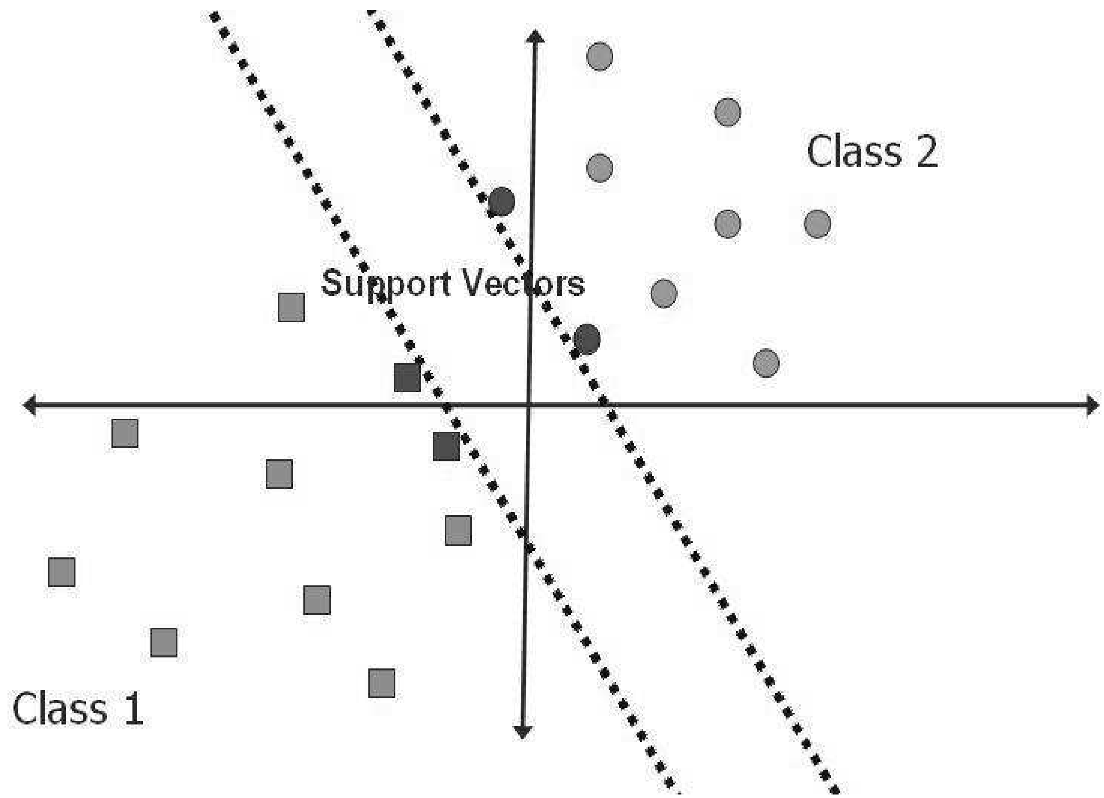
$$i \in S \quad \Rightarrow \quad y_i(X_i^T W^* + b^*) = 1$$

  Implies X_i is closest to separating hyperplane.
- $\{X_i \mid i \in S\}$ are called Support vectors. We have

$$W^* = \sum_i \, \mu_i^* y_i X_i = \sum_{i \in S} \, \mu_i^* y_i X_i$$

- Optimal W is a linear combination of Support vectors.
- Support vectors constitute a very useful output of the method.

Optimal hyperplane



The SVM solution

- The optimal hyperplane – W^*, b^* given by:

The SVM solution

- The optimal hyperplane – W^*, b^* given by:

$$W^* = \sum_i \ \mu_i^* y_i X_i = \sum_{i \in S} \ \mu_i^* y_i X_i$$

The SVM solution

- The optimal hyperplane – W^*, b^* given by:

$$W^* = \sum_i \ \mu_i^* y_i X_i = \sum_{i \in S} \ \mu_i^* y_i X_i$$

$$b^* = y_j - X_j^T W^*, \quad j \ \text{s.t.} \ \mu_j^* > 0$$

(Note that $\mu_j^* > 0 \ \Rightarrow \ y_j(X_j^T W^* + b^*) = 1$)

The SVM solution

- The optimal hyperplane – W^*, b^* given by:

$$W^* = \sum_i \; \mu_i^* y_i X_i = \sum_{i \in S} \; \mu_i^* y_i X_i$$

$$b^* = y_j - X_j^T W^*, \quad j \text{ s.t. } \mu_j^* > 0$$

(Note that $\mu_j^* > 0 \;\; \Rightarrow \;\; y_j(X_j^T W^* + b^*) = 1$)

- Thus, W^*, b^* are determined by $\mu_i^*, \;\; i = 1, \ldots, n$.

The SVM solution

- The optimal hyperplane – W^*, b^* given by:

$$W^* = \sum_i \ \mu_i^* y_i X_i = \sum_{i \in S} \ \mu_i^* y_i X_i$$

$$b^* = y_j - X_j^T W^*, \quad j \text{ s.t. } \mu_j^* > 0$$

(Note that $\mu_j^* > 0 \ \ \Rightarrow \ \ y_j(X_j^T W^* + b^*) = 1$)

- Thus, W^*, b^* are determined by $\mu_i^*, \ \ i = 1, \ldots, n$.
- We can use the dual of the optimization problem to get μ_i^*.

Dual optimization problem for SVM

- The dual function is

$$q(\boldsymbol{\mu}) = \inf_{W,b} \left\{ \frac{1}{2} W^T W + \sum_{i=1}^{n} \mu_i [1 - y_i (W^T X_i + b)] \right\}$$

Dual optimization problem for SVM

- The dual function is

$$q(\boldsymbol{\mu}) = \inf_{W,b} \left\{ \frac{1}{2} W^T W + \sum_{i=1}^{n} \mu_i [1 - y_i(W^T X_i + b)] \right\}$$

- If $\sum \mu_i y_i \neq 0$ then $q(\boldsymbol{\mu}) = -\infty$.

Dual optimization problem for SVM

- The dual function is

$$q(\boldsymbol{\mu}) = \inf_{W,b} \left\{ \frac{1}{2} W^T W + \sum_{i=1}^{n} \mu_i [1 - y_i (W^T X_i + b)] \right\}$$

- If $\sum \mu_i y_i \neq 0$ then $q(\boldsymbol{\mu}) = -\infty$.
- Hence we need to maximize q only over those $\boldsymbol{\mu}$ s.t. $\sum \mu_i y_i = 0$.

Dual optimization problem for SVM

- The dual function is

$$q(\boldsymbol{\mu}) = \inf_{W,b} \left\{ \frac{1}{2} W^T W + \sum_{i=1}^{n} \mu_i [1 - y_i (W^T X_i + b)] \right\}$$

- If $\sum \mu_i y_i \neq 0$ then $q(\boldsymbol{\mu}) = -\infty$.
- Hence we need to maximize q only over those $\boldsymbol{\mu}$ s.t. $\sum \mu_i y_i = 0$.
- Infimum w.r.t. W is attained at $W = \sum \mu_i y_i X_i$.

Dual optimization problem for SVM

- The dual function is

$$q(\boldsymbol{\mu}) = \inf_{W,b} \left\{ \frac{1}{2} W^T W + \sum_{i=1}^{n} \mu_i [1 - y_i (W^T X_i + b)] \right\}$$

- If $\sum \mu_i y_i \neq 0$ then $q(\boldsymbol{\mu}) = -\infty$.
- Hence we need to maximize q only over those $\boldsymbol{\mu}$ s.t. $\sum \mu_i y_i = 0$.
- Infimum w.r.t. W is attained at $W = \sum \mu_i y_i X_i$.
- We obtain the dual by substituting $W = \sum \mu_i y_i X_i$ and imposing $\sum \mu_i y_i = 0$.

- By substituting $W = \sum \mu_i y_i X_i$ and $\sum \ \mu_i y_i = 0$ we get

- By substituting $W = \sum \mu_i y_i X_i$ and $\sum \mu_i y_i = 0$ we get

$$q(\boldsymbol{\mu}) \ = \ \frac{1}{2}W^T W + \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \mu_i y_i (W^T X_i + b)$$

- By substituting $W = \sum \mu_i y_i X_i$ and $\sum \ \mu_i y_i = 0$ we get

$$
\begin{aligned}
q(\boldsymbol{\mu}) &= \frac{1}{2}W^T W + \sum_{i=1}^{n} \ \mu_i - \sum_{i=1}^{n} \ \mu_i y_i (W^T X_i + b) \\
&= \frac{1}{2}\left(\sum_i \mu_i y_i X_i\right)^T \sum_j \mu_j y_j X_j + \sum_i \ \mu_i \\
&\quad - \sum_i \ \mu_i y_i X_i^T (\sum_j \mu_j y_j X_j)
\end{aligned}
$$

- By substituting $W = \sum \mu_i y_i X_i$ and $\sum \mu_i y_i = 0$ we get

$$
\begin{aligned}
q(\boldsymbol{\mu}) &= \frac{1}{2}W^T W + \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \mu_i y_i (W^T X_i + b) \\
&= \frac{1}{2}\left(\sum_i \mu_i y_i X_i\right)^T \sum_j \mu_j y_j X_j + \sum_i \mu_i \\
&\quad - \sum_i \mu_i y_i X_i^T (\sum_j \mu_j y_j X_j) \\
&= \sum_i \mu_i - \frac{1}{2}\sum_i \sum_j \mu_i y_i \mu_j y_j X_i^T X_j
\end{aligned}
$$

- Thus, the dual problem is:

$$\max_{\boldsymbol{\mu}} \qquad q(\boldsymbol{\mu}) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j X_i^T X_j$$

subject to $\quad \mu_i \geq 0, \;\; i = 1, \ldots, n, \;\; \sum_{i=1}^{n} y_i \mu_i = 0$

- Thus, the dual problem is:

$$\max_{\boldsymbol{\mu}} \qquad q(\boldsymbol{\mu}) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j X_i^T X_j$$

$$\text{subject to} \qquad \mu_i \geq 0, \;\; i = 1, \ldots, n, \;\; \sum_{i=1}^{n} y_i \mu_i = 0$$

- Quadratic cost function and linear constraints

- Thus, the dual problem is:

$$\max_{\boldsymbol{\mu}} \qquad q(\boldsymbol{\mu}) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j X_i^T X_j$$

$$\text{subject to} \qquad \mu_i \geq 0, \;\; i = 1, \ldots, n, \;\; \sum_{i=1}^{n} y_i \mu_i = 0$$

- Quadratic cost function and linear constraints
- Training data vectors appear only as innerproduct

- Thus, the dual problem is:

$$\max_{\boldsymbol{\mu}} \qquad q(\boldsymbol{\mu}) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j X_i^T X_j$$

subject to $\qquad \mu_i \geq 0, \;\; i = 1, \ldots, n, \;\; \sum_{i=1}^{n} y_i \mu_i = 0$

- Quadratic cost function and linear constraints
- Training data vectors appear only as innerproduct
- Optimization is over $\Re^n$ irrespective of the dimension of X_i.