Machine Learning

CMPT 726

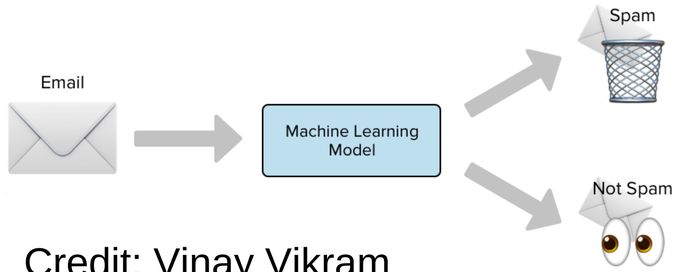
Mo Chen
SFU School of Computing Science
2022-11-08

Classification

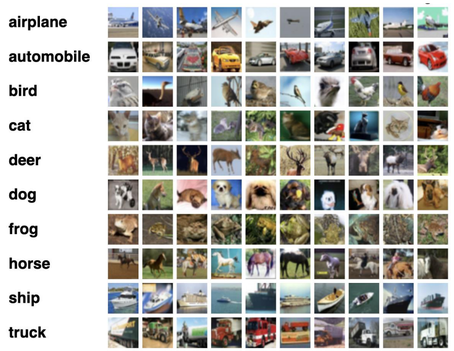
Motivation

In many problems, we would like to categorize an observation:

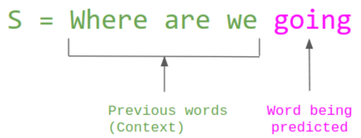
- Is an email spam?
- What object is depicted in an image?
- What will be the next word?



Credit: Vinay Vikram



Credit: Alex Krizhevsky



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

Credit: The Gradient

Binary Classification

Unlike regression, the goal of classification is to classify the input into one of multiple discrete classes.

A regression model produces a real number or in the case of multiple output regression, a real vector.

A classification model produces a class prediction.

Such a model is known as a **classifier**.

In binary classification, the goal is to classify into one of **two** discrete classes.

A binary classification model is known as a **binary classifier**.

Without loss of generality, we call one class the **positive class** and the other the **negative class**.

Data points whose labels are positive are known as **positive examples** and data points whose labels are negative are known as **negative examples**.

Support Vector Machines

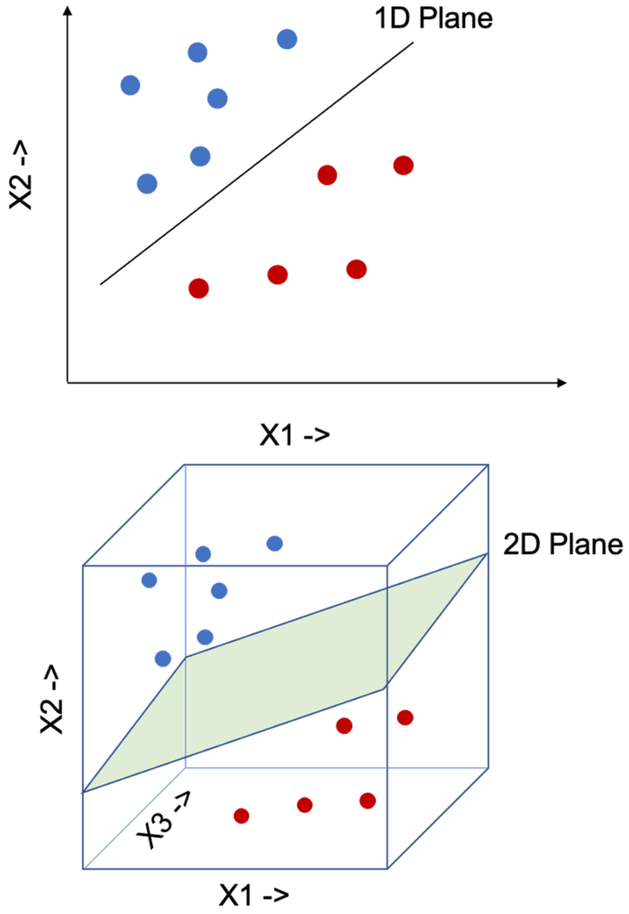
Given a dataset of input-output pairs (a.k.a. “observations”), $\{(\vec{x}_i, y_i)\}_{i=1}^{N}$, where $\vec{x}_i \in \mathbb{R}^{n-1}$ and $y_i \in \{-1,1\}$

We will construct a model called the support vector machine (SVM) to predict the label y from the data point $\vec{x}$.

The model is simply a line (in the case of 2D data), a plane (in the case of 3D data) or more generally, a hyperplane (in the case of higher dimensional data) that separates the data points.

For a new data point on one side, we predict the positive label.

For a new data point on the other side, we predict the negative label.



Credit: Abhisek Jana

Support Vector Machines

The **decision boundary** is the boundary that separates the region where the model generates positive predictions from the region where the model generates negative predictions.

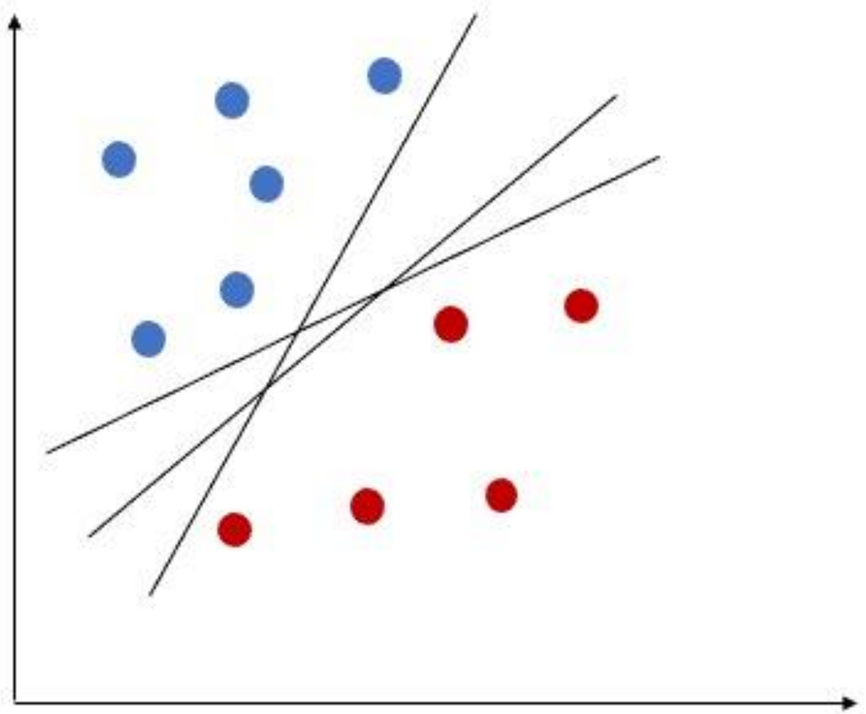
A **linear classifier** whose decision boundary is a hyperplane.

The support vector machine is an example of a linear classifier.

Support Vector Machines

There are many hyperplanes that would classify a training dataset perfectly.

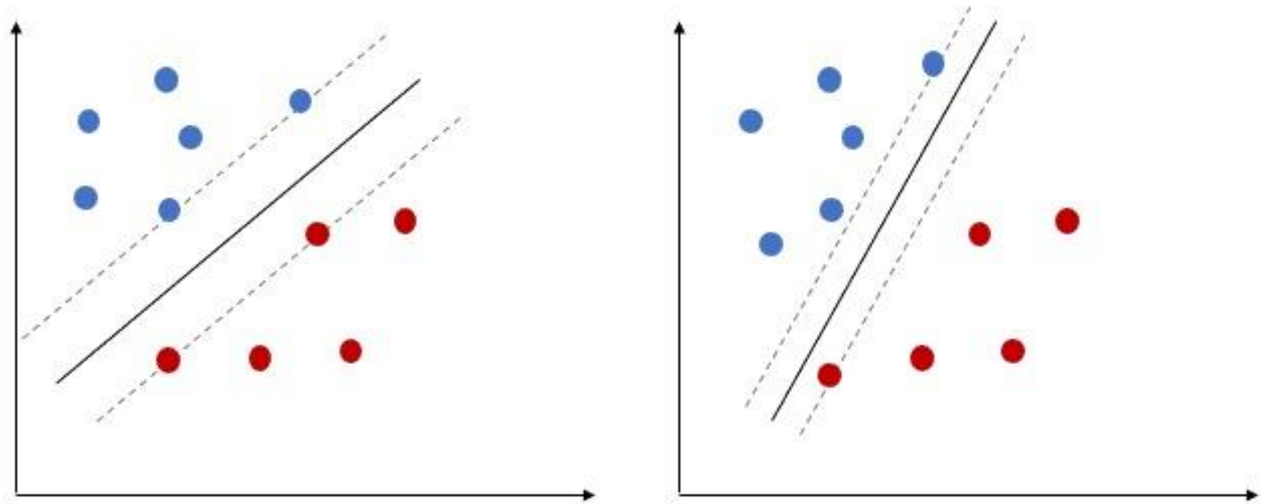
Which one should we choose?



Support Vector Machines

A boundary that is as far away from data points as possible is more robust to perturbations to the data points.

Intuitively, such a boundary is less prone to overfitting.



Preferred Less Preferred

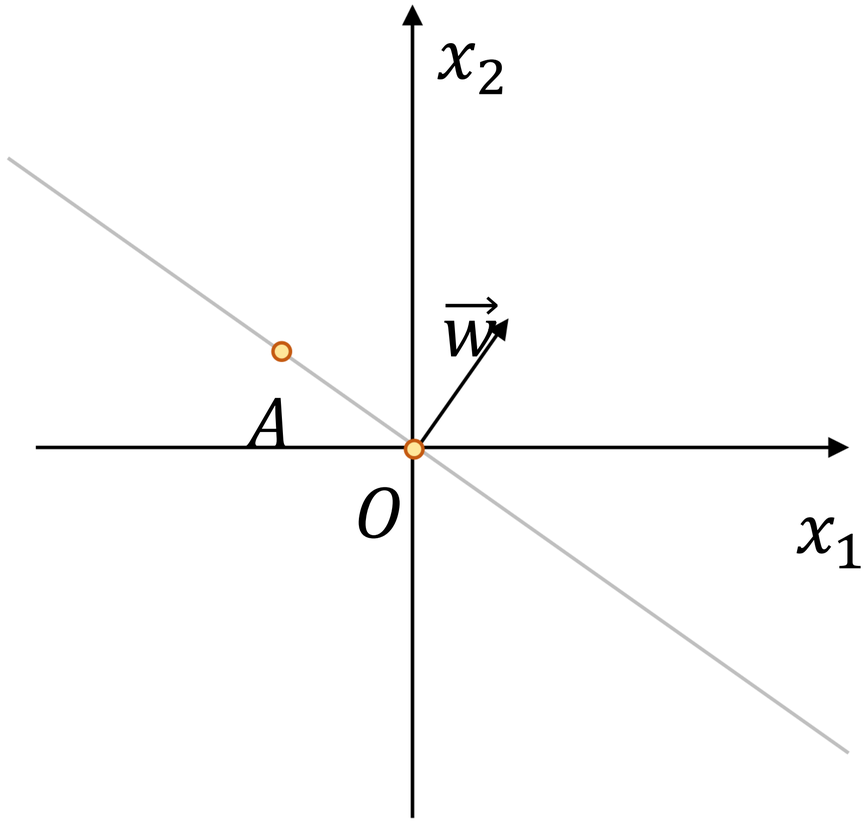
Let’s formulate an optimization problem to find such a boundary.

Hyperplanes

Consider a vector $\vec{w}$, and a hyperplane that is perpendicular to it (shown on the right).

For any A on the hyperplane, $\overrightarrow{OA}$ is perpendicular to $\vec{w}$.

Hence, $\vec{w}^{\top}(\overrightarrow{OA}) = 0$. So, this hyperplane corresponds to the set $\{\vec{x} | \vec{w}^{\top}\vec{x} = 0\}$.



Special case when $b = 0$

Hyperplanes

We shift the hyperplane in parallel (as shown on the right).

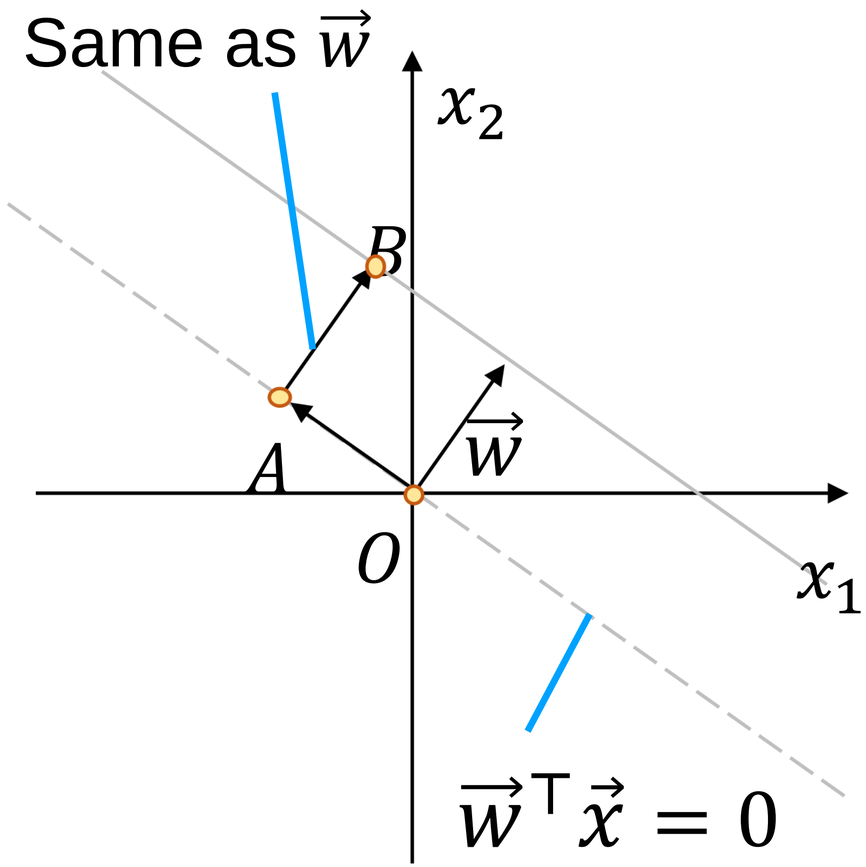
Consider any point B on the hyperplane and the vector $\overrightarrow{OB}$.

$$\begin{aligned}\vec{w}^{\top}\left(\overrightarrow{OB}\right) &= \vec{w}^{\top}\left(\overrightarrow{OA} + \overrightarrow{AB}\right)\\ &= \vec{w}^{\top}\left(\overrightarrow{OA}\right) + \vec{w}^{\top}\left(\overrightarrow{AB}\right)\\ &= 0 + \vec{w}^{\top}\vec{w}\\ &= \|\vec{w}\|_2^2\end{aligned}$$

So, for any B on the hyperplane, $\vec{w}^{\top}(\overrightarrow{OB}) = \|\vec{w}\|_2^2$.

Hence, the hyperplane corresponds to the following set:

$$\{\vec{x} | \vec{w}^{\top}\vec{x} = \|\vec{w}\|_2^2\}$$

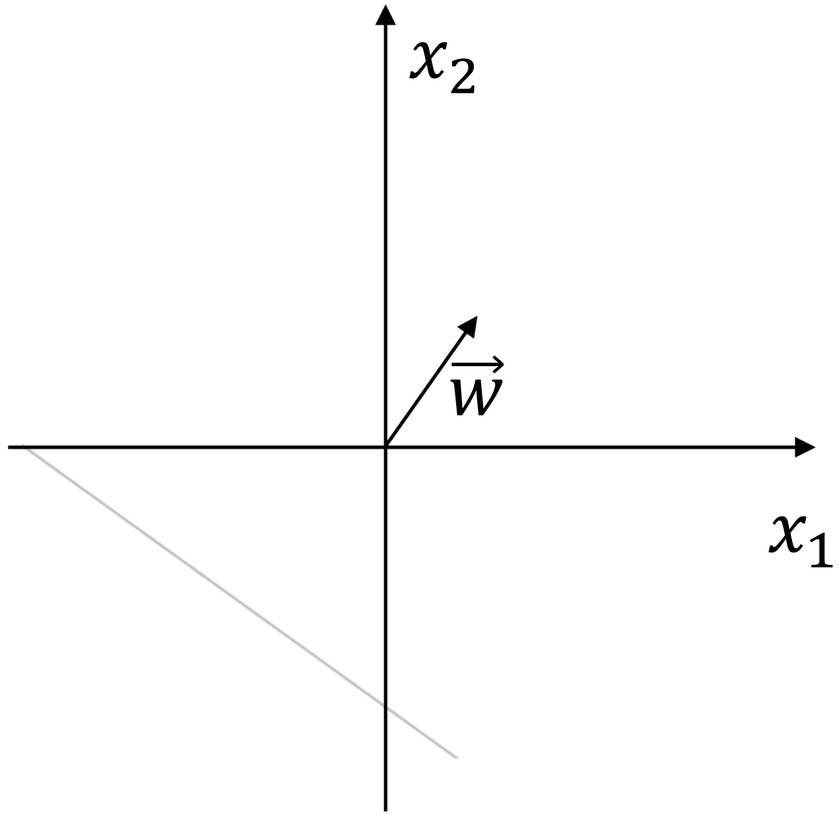


Special case when $b = \|\vec{w}\|_2^2$

Hyperplanes

So, in general:

As b changes, $\{\vec{x}|\vec{w}^{\top}\vec{x} = b\}$ corresponds to shifting the hyperplane in parallel.



Distance to the Hyperplane

Consider an arbitrary points on the hyperplane, $\vec{z}$.

The distance to the hyperplane from $\vec{x}$ can be obtained by projecting the vector $\vec{x} - \vec{z}$ along the vector $\vec{w}$) which is orthogonal to the hyperplane).

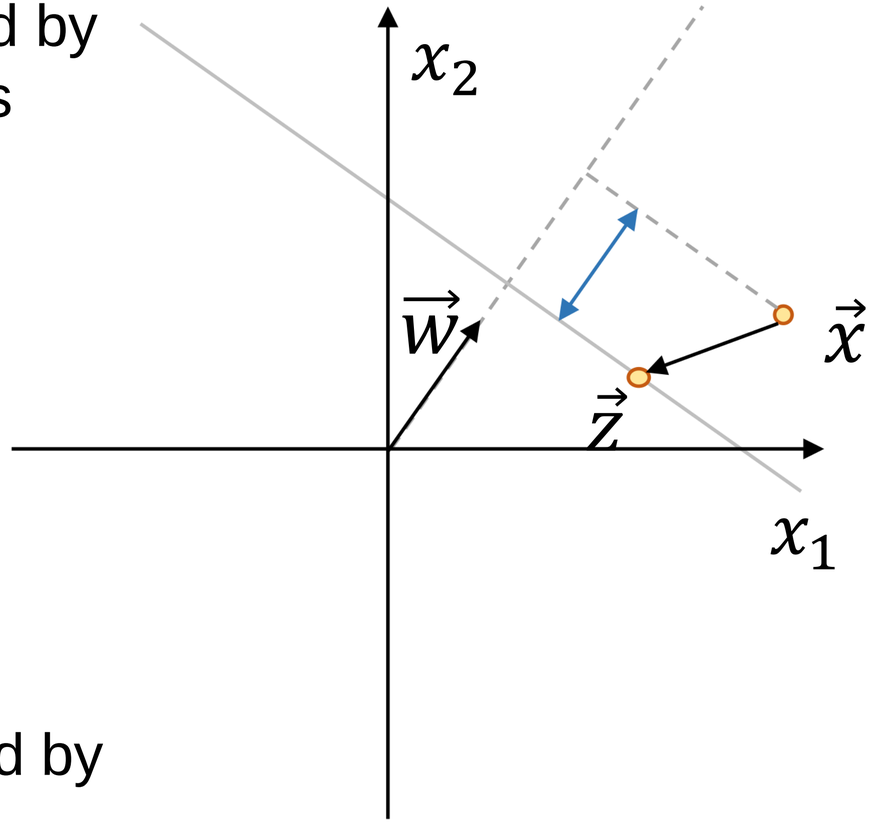
The length of the projection is given by:

$$\left|\left\langle \vec{x} - \vec{z}, \frac{\vec{w}}{\|\vec{w}\|_2} \right\rangle\right| = \frac{1}{\|\vec{w}\|_2} |\langle \vec{x} - \vec{z}, \vec{w} \rangle|$$

$$= \frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x} - \vec{w}^\top \vec{z}|$$

Because $\vec{z}$ is on the hyperplane, which is characterized by $\{\vec{x} | \vec{w}^\top \vec{x} = b\}$, $\vec{w}^\top \vec{z} = b$.

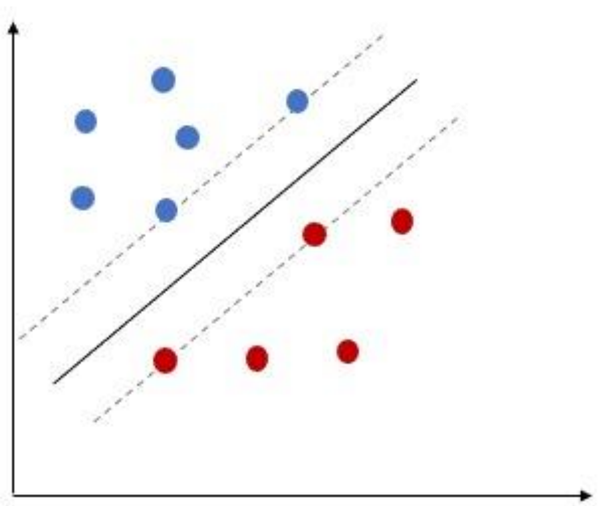
So, the distance to the hyperplane from $\vec{x}$ is $\frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x} - b|$.



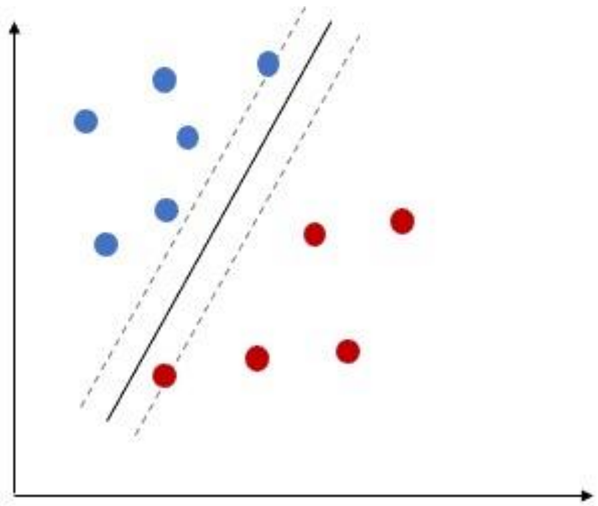
Goals of SVM

Recall: We would like to find a hyperplane that:

(1) Separates the positive data points from the negative data points
(2) Is as far away from the data points as possible



Preferred



Less Preferred

Margin

We define the width of the buffer on each side of the hyperplane as the margin.

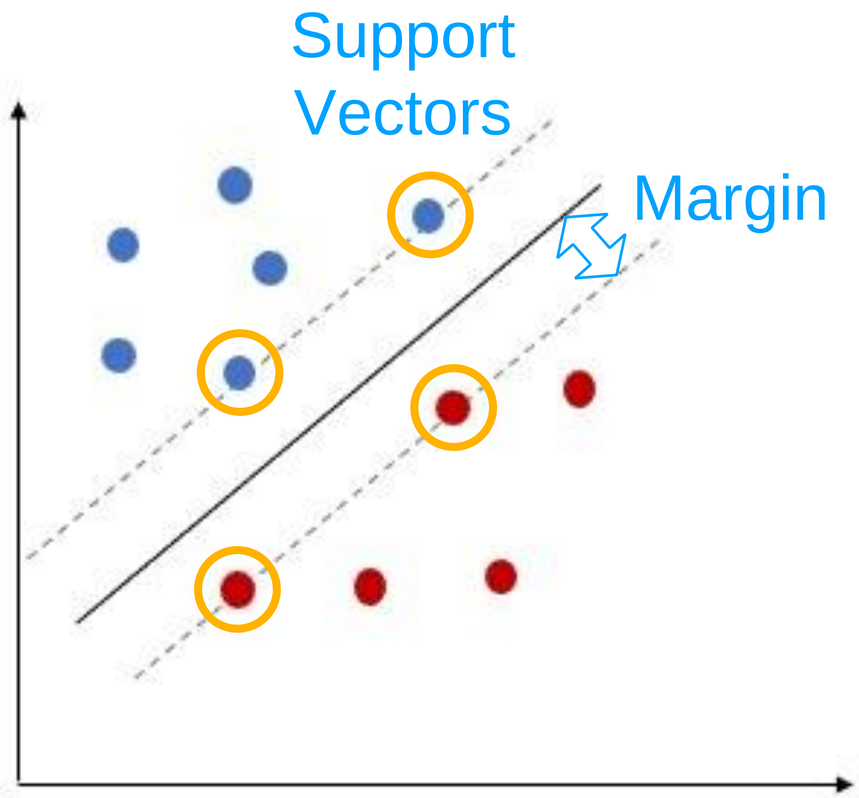
Let’s derive a mathematical expression for the margin.

The margin is determined by the data points closest to the hyperplane.

These data points are known as **support vectors**.

The margin is the distance from support vectors to the hyperplane:

$$m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x}_i - b| \right\}$$



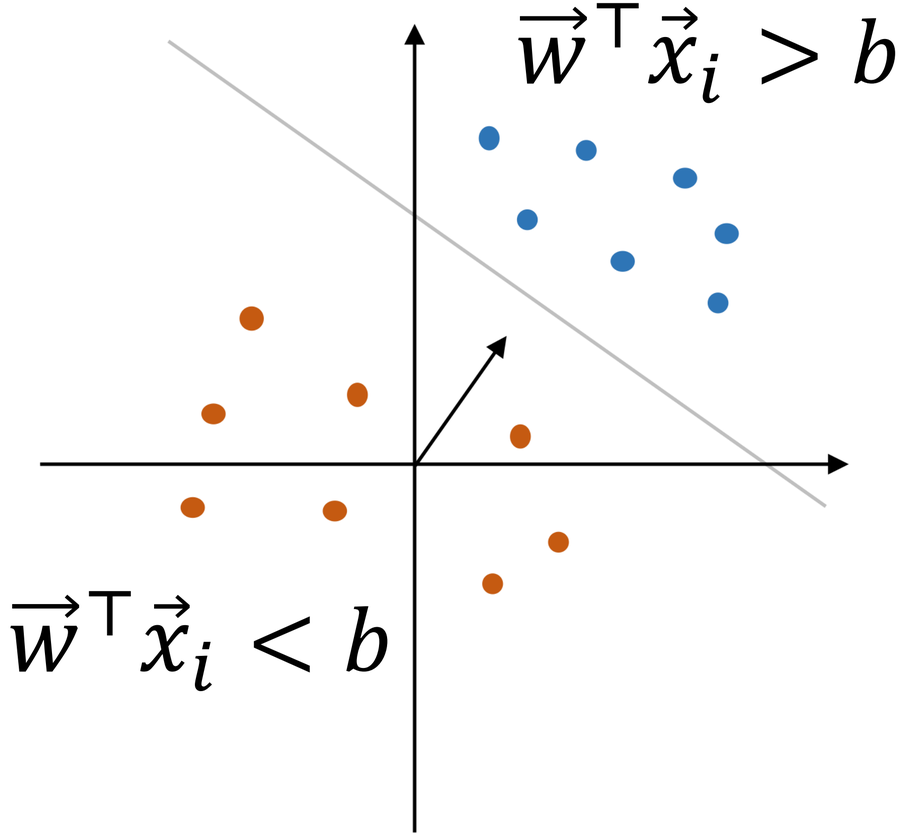
Formulation

Recall: We would like to find a hyperplane that:

(1) Separates the positive data points from the negative data points
$\vec{w}^\top \vec{x}_i > b$ for all i such that $y_i = 1$
$\vec{w}^\top \vec{x}_i < b$ for all i such that $y_i = -1$

(2) Is as far away from the data points as possible

Maximize the margin $m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x}_i - b| \right\}$



Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i > b \text{ for all } i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i < b \text{ for all } i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x}_i - b| \right\}$$

The definition of margin

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \text{ for all } i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \text{ for all } i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$m = \min_i \left\{ \frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x}_i - b| \right\}$$

The definition of margin

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \; \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \; \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\frac{1}{\|\vec{w}\|_2} |\vec{w}^\top \vec{x}_i - b| \geq m$$

The margin is by definition less than or equal to the distance from any data point to the hyperplane

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \; \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \; \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$|\vec{w}^\top \vec{x}_i - b| \geq m\|\vec{w}\|_2 \; \forall i$$

The margin is by definition less than or equal to the distance from any data point to the hyperplane

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \; \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \; \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b \geq m\|\vec{w}\|_2 \text{ or } \vec{w}^\top \vec{x}_i - b \leq -m\|\vec{w}\|_2 \; \forall i$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

Maximize margin

subject to

$$\vec{w}^\top \vec{x}_i - b > 0 \; \forall i \text{ such that } y_i = 1,$$

For positive examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b < 0 \; \forall i \text{ such that } y_i = -1,$$

For negative examples, should lie on one side of the hyperplane

$$\vec{w}^\top \vec{x}_i - b \geq m\|\vec{w}\|_2 \; \forall i \text{ such that } y_i = 1,$$

$$\vec{w}^\top \vec{x}_i - b \leq -m\|\vec{w}\|_2 \; \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

Maximize margin

$$\max_{m,\vec{w},b} m$$

subject to

$$\vec{w}^\top \vec{x}_i - b \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$\vec{w}^\top \vec{x}_i - b \leq -m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

Maximize margin

$$\max_{m,\vec{w},b} m$$

subject to

$$1 \cdot (\vec{w}^{\top}\vec{x}_i - b) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$-1 \cdot (\vec{w}^{\top}\vec{x}_i - b) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

Maximize margin

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = 1,$$

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq m\|\vec{w}\|_2 \ \forall i \text{ such that } y_i = -1,$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i\left((\alpha\vec{w})^\top \vec{x}_i - (\alpha b)\right) \geq m\|(\alpha\vec{w})\|_2$$

$$m \geq 0$$

It turns out that the scale of $\begin{pmatrix}\vec{w}\\ b\end{pmatrix}$ doesn't matter, in the sense that if $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ is a feasible solution, then $\begin{pmatrix}\alpha\vec{w}^*\\ \alpha b^*\\ m^*\end{pmatrix}$ is feasible and achieves the same objective value as $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ for any $\alpha > 0$.

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

$$\max_{m,\vec{w},b} m$$

subject to

$$\alpha y_i(\vec{w}^\top \vec{x}_i - b) \geq m|\alpha|\|\vec{w}\|_2$$

$$m \geq 0$$

It turns out that the scale of $\begin{pmatrix}\vec{w}\\ b\end{pmatrix}$ doesn't matter, in the sense that if $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ is a feasible solution, then $\begin{pmatrix}\alpha\vec{w}^*\\ \alpha b^*\\ m^*\end{pmatrix}$ is feasible and achieves the same objective value as $\begin{pmatrix}\vec{w}^*\\ b^*\\ m^*\end{pmatrix}$ for any $\alpha > 0$.

Formulation

$$\max_{m,\vec{w},b} m$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq m\|\vec{w}\|_2$$

$$m \geq 0$$

Therefore, without loss of generality, we can set the scale of $\vec{w}$.

We set $\|\vec{w}\|_2 = \frac{1}{m}$, so $m = \frac{1}{\|\vec{w}\|_2}$

Formulation

$$\max_{\vec{w},b} \frac{1}{\|\vec{w}\|_2}$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq 1$$

Since $x \mapsto \frac{1}{x}$ is strictly decreasing, we can apply the transformation to the objective and change the max to a min.

Formulation

$$\min_{\vec{w},b} \|\vec{w}\|_2$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq 1$$

Want to make the objective function convex. Will see why this is useful later.

Since $x \mapsto \frac{1}{2}x^2$ is strictly increasing for $x \geq 0$, we can apply the transformation to the objective.

Formulation

$$\min_{\vec{w},b} \frac{1}{2} \|\vec{w}\|_2^2$$

subject to

$$y_i(\vec{w}^\top \vec{x}_i - b) \geq 1$$

Since $x \mapsto \frac{1}{2}x^2$ is strictly increasing for $x \geq 0$, we can apply the transformation to the objective.

This is an example of a constrained optimization problem

Lagrangian Duality

Constrained a generic constrained optimization problem:

$\min_{\vec{\theta}} f(\vec{\theta})$ Objective function

subject to

$g_i(\vec{\theta}) \leq 0 \; \forall i \in \{1, \dots, k\}$ Inequality constraints

$h_i(\vec{\theta}) = 0 \; \forall i \in \{1, \dots, l\}$ Equality constraints

$S := \{\vec{\theta} | g_i(\vec{\theta}) \leq 0 \; \forall i \in \{1, \dots, k\} \text{ and } h_i(\vec{\theta}) = 0 \; \forall i \in \{1, \dots, l\}\}$ is known as the **feasible region**. A solution $\vec{\theta}$ is the feasible region is known as a **feasible solution**.

Generalized Lagrangian

The generalized Lagrangian turns a constrained optimization problem to an unconstrained optimization problem.

$$\mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu}) = f(\vec{\theta}) + \sum_{i=1}^{k} \lambda_i g_i(\vec{\theta}) + \sum_{i=1}^{l} \nu_i h_i(\vec{\theta})$$

Consider the function $\Phi(\vec{\theta}) = \max_{\vec{\lambda},\vec{\nu}:\lambda_i \geq 0 \ \forall i} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$

Observe that:

$$\Phi(\vec{\theta}) = \begin{cases} f(\vec{\theta}), & g_i(\vec{\theta}) \leq 0 \ \forall i \in \{1, \dots, k\} \quad \text{and} \quad h_i(\vec{\theta}) = 0 \ \forall i \in \{1, \dots, l\} \\ \infty, & \text{otherwise} \end{cases}$$

Generalized Lagrangian

$$p^* = \min_{\vec{\theta}} \Phi(\vec{\theta}) = \min_{\vec{\theta}} \max_{\vec{\lambda},\vec{\nu}:\lambda_i \geq 0 \; \forall i} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$$

Since $\Phi(\vec{\theta}) = f(\vec{\theta})$ in the feasible region and is ∞ otherwise, the above is equivalent to the original constrained optimization problem.

The above is known as the **primal problem**.

Now consider the following optimization problem, which is different in that the min and max are swapped:

$$d^* = \max_{\vec{\lambda},\vec{\nu}:\lambda_i \geq 0 \; \forall i} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$$

The above is known as the **dual problem**.

Weak Duality

Weak duality relates $p^* = \min_{\vec{\theta}} \max_{\vec{\lambda},\vec{\nu}:\lambda_i \geq 0\ \forall i} \mathcal{L}(\vec{\theta},\vec{\lambda},\vec{\nu})$ to $d^* = \max_{\vec{\lambda},\vec{\nu}:\lambda_i \geq 0\ \forall i} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda},\vec{\nu})$.

$\forall \vec{\theta}', \mathcal{L}(\vec{\theta}',\vec{\lambda},\vec{\nu}) \geq \min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda},\vec{\nu})$ at **every point** in the space of $\vec{\lambda}$ and $\vec{\nu}$. (1)

Consider a point $(\vec{\lambda}_0, \vec{\nu}_0)$ that maximizes the function $\min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda},\vec{\nu})$.

So, from (1), in particular,$\forall \vec{\theta}', \mathcal{L}(\vec{\theta}',\vec{\lambda}_0,\vec{\nu}_0) \geq \min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda}_0,\vec{\nu}_0) = \max_{\vec{\lambda},\vec{\nu}} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda},\vec{\nu})$. (2)

On the other hand, $\forall \vec{\theta}', \max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}',\vec{\lambda},\vec{\nu}) \geq \mathcal{L}(\vec{\theta}',\vec{\lambda}_0,\vec{\nu}_0)$. (3)

Combining (2) and (3), we have $\forall \vec{\theta}', \max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}',\vec{\lambda},\vec{\nu}) \geq \mathcal{L}(\vec{\theta}',\vec{\lambda}_0,\vec{\nu}_0) \geq \min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda}_0,\vec{\nu}_0) = \max_{\vec{\lambda},\vec{\nu}} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta},\vec{\lambda},\vec{\nu})$. (4)

Weak Duality

From the previous slide, we have: $\forall \vec{\theta}', \max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}', \vec{\lambda}, \vec{\nu}) \geq \max_{\vec{\lambda},\vec{\nu}} \min_{\vec{\theta}} \mathcal{L}\left(\vec{\theta}, \vec{\lambda}, \vec{\nu}\right)$. (4)

Consider a point $\vec{\theta}_0'$ that minimizes the function $\max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}', \vec{\lambda}, \vec{\nu})$.

From (4), in particular, $\max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}_0', \vec{\lambda}, \vec{\nu}) \geq \max_{\vec{\lambda},\vec{\nu}} \min_{\vec{\theta}} \mathcal{L}\left(\vec{\theta}, \vec{\lambda}, \vec{\nu}\right)$.

Therefore, $p^* = \min_{\vec{\theta}} \max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu}) \geq \max_{\vec{\lambda},\vec{\nu}} \min_{\vec{\theta}} \mathcal{L}\left(\vec{\theta}, \vec{\lambda}, \vec{\nu}\right) = d^*$

Strong Duality

The quantity $p^* - d^*$ is known as the **duality gap**.

Under some conditions, the duality gap is zero, i.e. $p^* = d^*$.

In this case, we say **strong duality** holds.

Strong duality is nice because if it holds, solving the primal problem of $\min_{\vec{\theta}} \max_{\vec{\lambda},\vec{\nu}} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$ is equivalent to solving the dual problem of $\max_{\vec{\lambda},\vec{\nu}} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$.

Sometimes the dual problem can be solved more efficiently than the primal problem.

Sometimes it reveals relationships between the optimal solution and the quantities given in the problem that are not obvious from the primal problem.

In the case of SVM, we will see both when we look at the dual problem. It also results in a natural way to generalize an SVM.

Dual Form of the SVM

Recall the primal form of the SVM: $\min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|_2^2$ s.t. $y_i(\vec{w}^\top \vec{x}_i - b) \geq 1 \; \forall i$

Recall: For $\min_{\vec{\theta}} f(\vec{\theta})$ s.t. $g_i(\vec{\theta}) \leq 0 \; \forall i \in \{1, \ldots, k\}$ and $h_i(\vec{\theta}) = 0 \; \forall i \in \{1, \ldots, l\}$, the dual problem is

$\max_{\vec{\lambda},\vec{\nu}:\lambda_i \geq 0 \; \forall i} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$, where $\mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu}) = f(\vec{\theta}) + \sum_{i=1}^{k} \lambda_i g_i(\vec{\theta}) + \sum_{i=1}^{l} \nu_i h_i(\vec{\theta})$

In the case of the SVM, $\vec{\theta} = \begin{pmatrix} \vec{w} \\ b \end{pmatrix}$, and we don't have $\vec{\nu}$ because we only have inequality constraints.

$$\mathcal{L}(\vec{w}, b, \vec{\lambda}) = \frac{1}{2}\|\vec{w}\|_2^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i(\vec{w}^\top \vec{x}_i - b)\right)$$

The dual problem is

$$\max_{\vec{\lambda}:\lambda_i \geq 0} \min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda}) = \max_{\vec{\lambda}:\lambda_i \geq 0} \min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|_2^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i(\vec{w}^\top \vec{x}_i - b)\right)$$

Dual Form of the SVM

Let's solve the inner optimization problem of $\min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda})$:

$$\mathcal{L}(\vec{w}, b, \vec{\lambda}) = \frac{1}{2}\|\vec{w}\|_2^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i(\vec{w}^\top \vec{x}_i - b)\right)$$

$$= \frac{1}{2}\vec{w}^\top \vec{w} + \sum_{i=1}^{N} (\lambda_i - \lambda_i y_i \vec{w}^\top \vec{x}_i + \lambda_i y_i b)$$

Recall: $\frac{\partial(\vec{x}^\top A \vec{x})}{\partial \vec{x}} = (A + A^\top)\vec{x}$ and $\frac{\partial(\vec{a}^\top \vec{x})}{\partial \vec{x}} = \vec{a}$

$$0 = \frac{\partial \mathcal{L}}{\partial \vec{w}} = \frac{1}{2}(I + I^\top)\vec{w} - \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i = \frac{1}{2}(2I)\vec{w} - \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i = \vec{w} - \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i \Longrightarrow \vec{w} = \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i \quad (1)$$

$$0 = \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i \quad (2)$$

Both equations must hold at the critical point of $\mathcal{L}$.

Dual Form of the SVM

Let's check if the critical point is indeed a global minimum.

$$\frac{\partial \mathcal{L}}{\partial \vec{w}} = \vec{w} - \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i \text{ , and } \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i$$

Recall: $\frac{\partial (A\vec{x})}{\partial \vec{x}} = A^\top$

$$\frac{\partial^2 \mathcal{L}}{\partial \vec{w} \partial \vec{w}^\top} = I, \frac{\partial^2 \mathcal{L}}{\partial \vec{w} \partial b} = \vec{0}^\top, \frac{\partial^2 \mathcal{L}}{\partial b \partial \vec{w}^\top} = \vec{0}, \text{and} \frac{\partial^2 \mathcal{L}}{\partial b \partial b} = 0$$

Recall $\vec{\theta} = \begin{pmatrix} \vec{w} \\ b \end{pmatrix}$, so $\frac{\partial^2 \mathcal{L}}{\partial \vec{\theta} \partial \vec{\theta}^\top} = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \vec{w} \partial \vec{w}^\top} & \frac{\partial^2 \mathcal{L}}{\partial b \partial \vec{w}^\top} \\ \frac{\partial^2 \mathcal{L}}{\partial \vec{w} \partial b} & \frac{\partial^2 \mathcal{L}}{\partial b \partial b} \end{pmatrix} = \begin{pmatrix} I & \vec{0} \\ \vec{0}^\top & 0 \end{pmatrix} \succcurlyeq 0$

So $\mathcal{L}$ is convex in $\vec{\theta}$, and so any critical point must be a global minimum.

Dual Form of the SVM

$$\mathcal{L}(\vec{w}, b, \vec{\lambda}) = \frac{1}{2}\vec{w}^\top\vec{w} + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \vec{w}^\top \vec{x}_i + \lambda_i y_i b\right)$$

To find $\min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda})$, let's plug the equations that must hold at the critical point into the generalized Lagrangian.

Recall we have two equations that hold at the critical point: (1) $\vec{w} = \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i$ and (2) $\sum_{i=1}^{N} \lambda_i y_i = 0$.

$$\min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda}) = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i\right)^\top \left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i\right) + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j\right)^\top \vec{x}_i + \lambda_i y_i b\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i^\top\right)\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i\right) + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right) \vec{x}_i + \lambda_i y_i b\right)$$

Dual Form of the SVM

$$\min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda}) = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i^\top\right)\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i\right) + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right)\vec{x}_i + \lambda_i y_i b\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_i^\top \vec{x}_j\right) + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right)\vec{x}_i + \lambda_i y_i b\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_i^\top \vec{x}_j\right) + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right)\vec{x}_i + \sum_{i=1}^{N} \lambda_i y_i b$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_i^\top \vec{x}_j\right) + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i + b\sum_{i=1}^{N} \lambda_i y_i$$

$$= \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i + b\sum_{i=1}^{N} \lambda_i y_i$$

Recall eq. (2) from a few slides ago: $0 = \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i$. Hence the last term is 0.

Dual Form of the SVM

$$\min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda}) = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i^\top\right)\left(\sum_{i=1}^{N} \lambda_i y_i \vec{x}_i\right) + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right)\vec{x}_i + \lambda_i y_i b\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_i^\top \vec{x}_j\right) + \sum_{i=1}^{N}\left(\lambda_i - \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right)\vec{x}_i + \lambda_i y_i b\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_i^\top \vec{x}_j\right) + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j \vec{x}_j^\top\right)\vec{x}_i + \sum_{i=1}^{N} \lambda_i y_i b$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_i^\top \vec{x}_j\right) + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i + b\sum_{i=1}^{N} \lambda_i y_i$$

$$= \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i \text{ subject to the condition that } \sum_{i=1}^{N} \lambda_i y_i = 0$$

Recall eq. (2) from a few slides ago: $0 = \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i$. Hence the last term is 0.

Dual Form of the SVM

Recall the dual problem is $\max_{\vec{\lambda}:\lambda_i \geq 0\ \forall i} \min_{\vec{\theta}} \mathcal{L}(\vec{\theta}, \vec{\lambda}, \vec{\nu})$.

From the previous slide,

$$\min_{\vec{w},b} \mathcal{L}(\vec{w}, b, \vec{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i \text{ s.t. } \sum_{i=1}^{N} \lambda_i y_i = 0$$

So, the dual form of the SVM is:

$$\max_{\vec{\lambda}} \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \vec{x}_j^\top \vec{x}_i$$

subject to

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

$$\lambda_i \geq 0\ \forall i$$

Slater's Condition

Recall: Strong duality does not always hold.

One *sufficient* condition for strong duality to hold is *Slater's condition*.

Recall: We have a constrained optimization problem of the following form:

$\min_{\vec{\theta}} f(\vec{\theta})$ s.t. $g_i(\vec{\theta}) \leq 0 \; \forall i \in \{1, \dots, k\}$ and $h_i(\vec{\theta}) = 0 \; \forall i \in \{1, \dots, l\}$

If $f(\vec{\theta})$ and $g_i(\vec{\theta})$ are convex in$\vec{\theta}$ $\forall i \in \{1, \dots, k\}$ and $h_i(\vec{\theta})$ is linear in$\vec{\theta}$ $\forall i \in \{1, \dots, l\}$, and there exists a point $\vec{\theta}_0$ such that $g_i(\vec{\theta}) < 0 \; \forall i \in \{1, \dots, k\}$ and $h_i(\vec{\theta}) = 0 \forall i \in \{1, \dots, l\}$,then strong duality holds.

Strict inequality!