50.007 Machine Learning

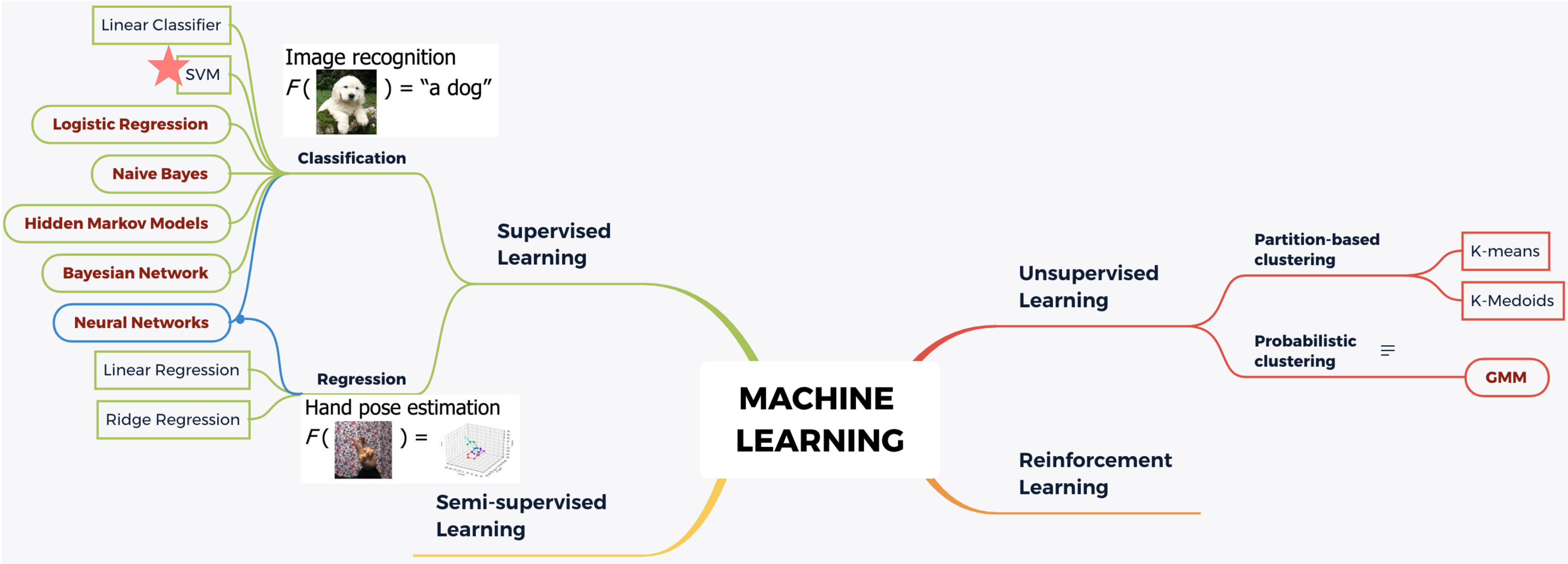
2026 Spring

7. Support Vector Machines (I)

Na Zhao
Assistant Professor, ISTD

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Roadmap



Linear Classifier

SVM

Logistic Regression

Naive Bayes

Hidden Markov Models

Bayesian Network

Neural Networks

Linear Regression

Ridge Regression

Image recognition

$F($) = "a dog"

Classification

Regression

Hand pose estimation

$F($) =

Supervised Learning

Semi-supervised Learning

MACHINE LEARNING

Unsupervised Learning

Partition-based clustering

K-means

K-Medoids

Probabilistic clustering

GMM

Reinforcement Learning

Learning Objectives

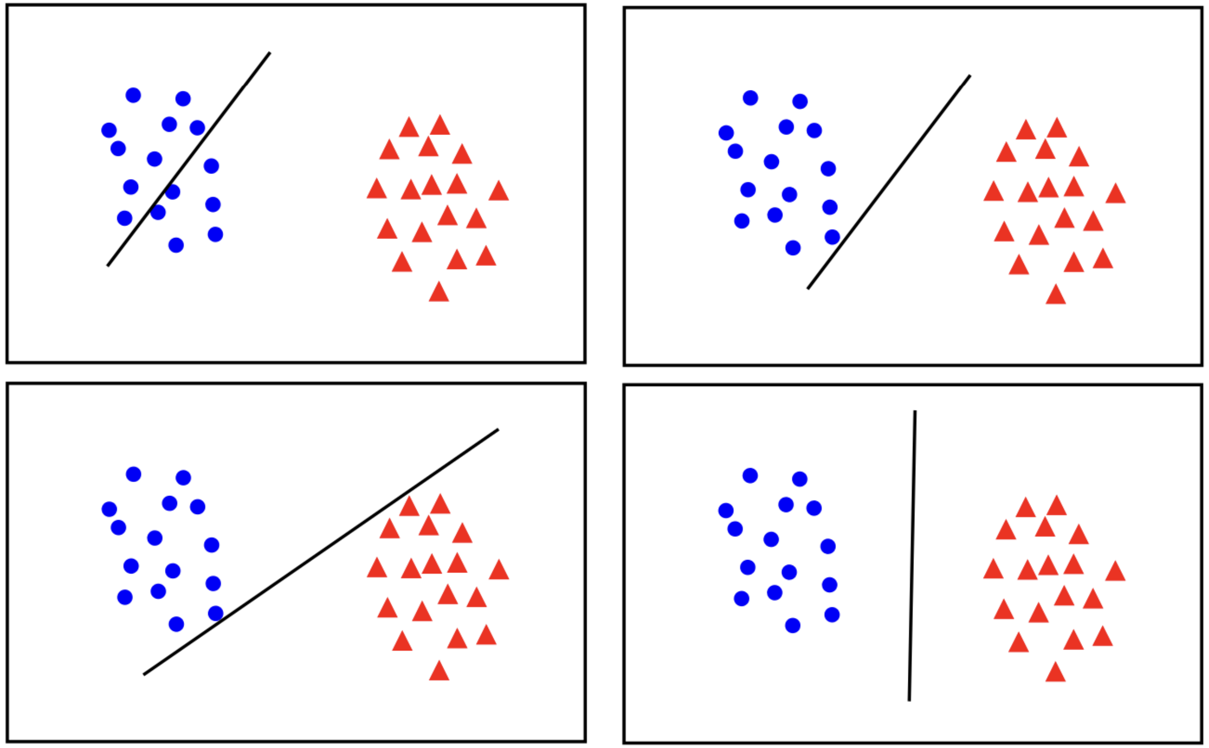
You should be able to know:

1. What is SVM and how it is formulated based on the notion of margin?
2. What is the primal form and dual form of SVM and their relations?
3. What is the definition of support vectors and how to check if a training instance is a support vector?

Overview of Support Vector Machine

Motivation

- What is the *best* decision boundary (*i.e.* separating hyperplane)?



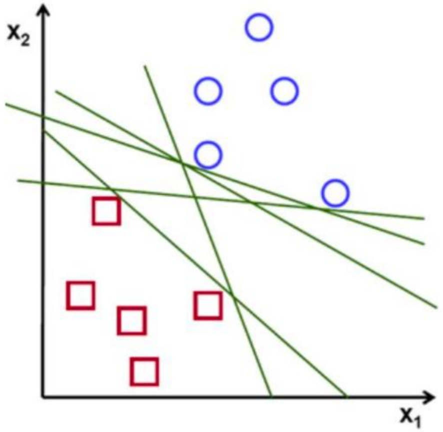
❑ Typically, if a data set is linearly separable, there are infinitely many separating hyperplanes.

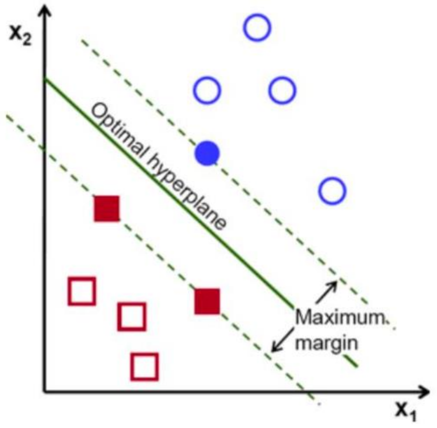
Support Vector Machine (SVM)

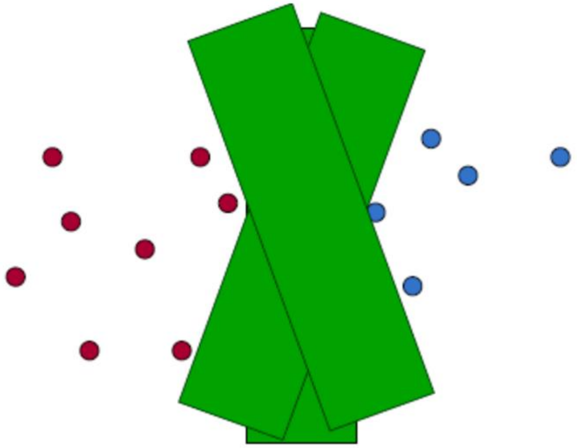
- **SVM** finds an optimal solution (*i.e.* hyperplane) that maximizes the distance between the hyperplane and the "*difficult points*" close to decision boundary.

- **One intuition**: if there are no points near the decision surface, then there are no uncertain classification decisions.

- **Another intuition**: If you have to place a fat separator between classes, you have fewer choices.



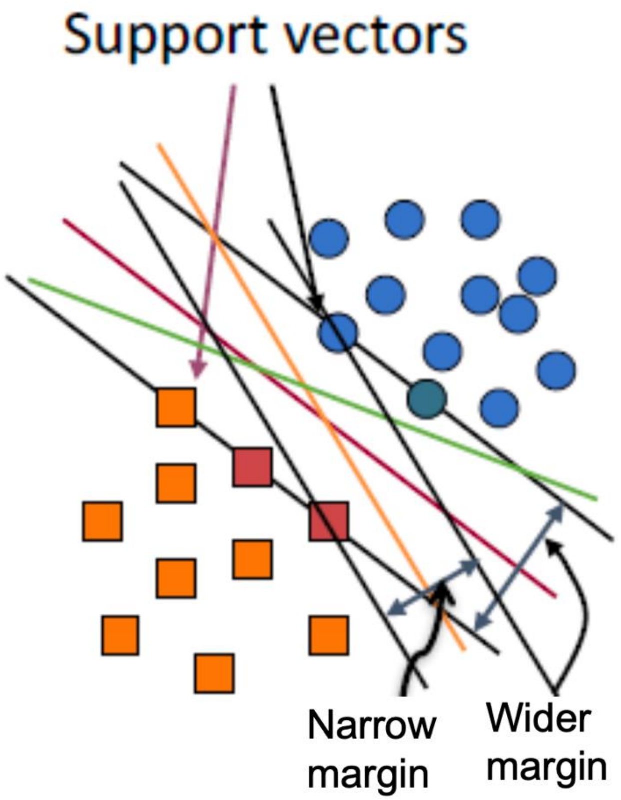




Support Vector Machine (SVM)

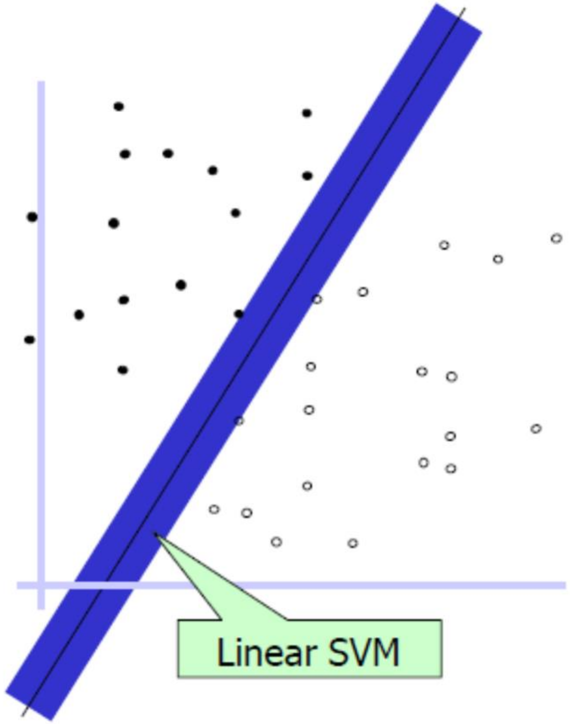
- The **SVM** algorithm finds a *linearly separable hyperplane* using **support vectors** and **margins**.
  - **Margin**: the width that the boundary can be increased before hitting a data point.
  - **Support vectors**: data points that lie exactly on the margin boundaries.

- The operation of the **SVM** algorithm is based on finding the **hyperplane** that *gives the largest distance to the support vectors, i.e.* to find the maximum margin.
  - SVM is also known as the **maximal margin classifier**.



SVM: Linearly separable data

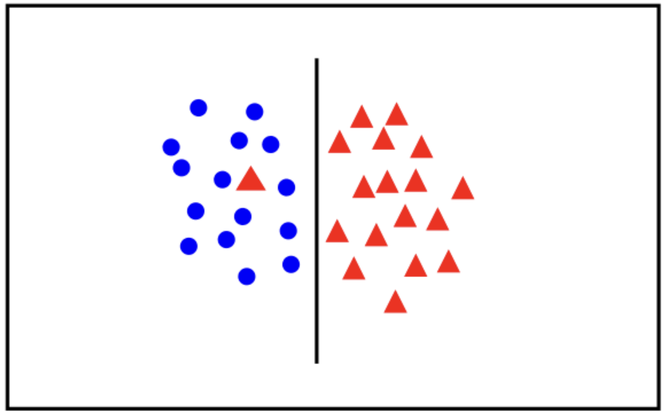
- When training data is linearly separable:
  - The simplest SVM (linear SVM) is the linear classifier with the maximum margin.



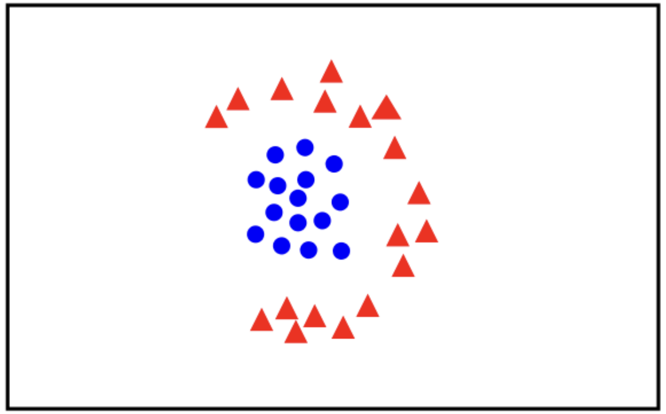
SVM: Non-linearly separable data

What if the data is not linearly separable?

- **Case A**: noisy data
  - introduce Slack variables
  - Allow a few points on the wrong side



- **Case B**: linear classifier not appropriate
  - Introduce Kernel trick
  - Map data to a higher dimensional space, do linear classification there.



SVM: more than two classes

N class problem: One-vs-All

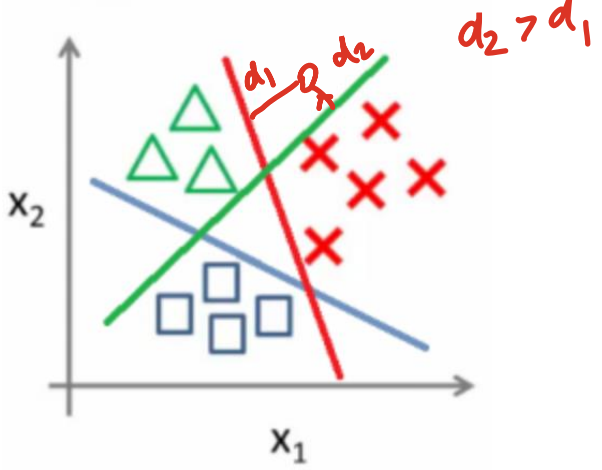
- Split the task into N binary tasks:

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

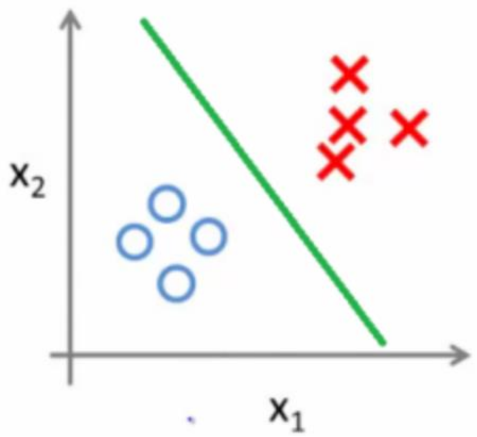
Finally, pick the class that put the point furthest into the positive region.

In this course, we mainly work with 2 classes.

Multi-class classification:



Binary classification:



Mathematical Intuition of Linear-SVM

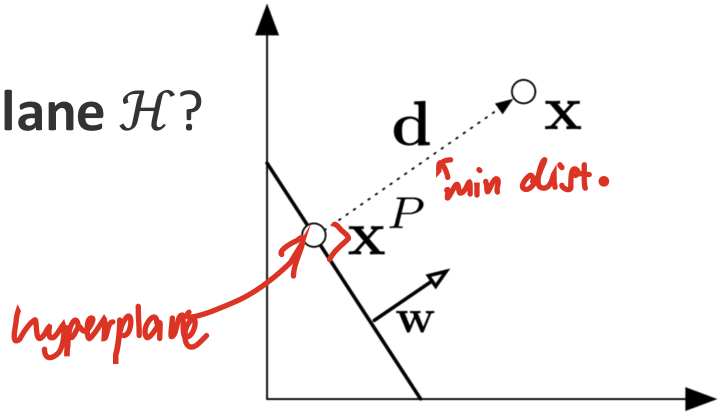
Notation

- A **hyperplane** is defined through $\overbrace{\mathbf{w}, b}^{\text{parameters}}$ as a set of points such that

$$\overset{\text{hyperplane}}{\mathcal{H}} = \{\mathbf{x} | \mathbf{w}^T\mathbf{x} + b = 0\}$$

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! ☺

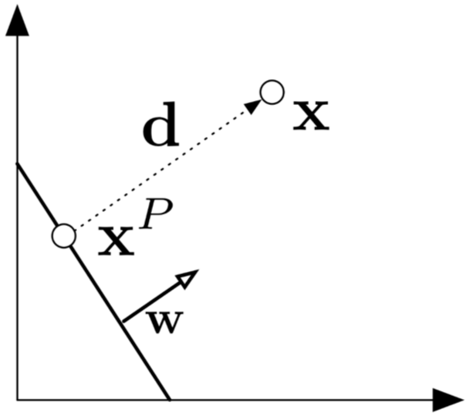
- The **geometric margin** γ is defined as the distance from the hyperplane to the *closest point* across both classes.
- **How to measure the distance of a point $\mathbf{x}$ to the hyperplane $\mathcal{H}$?**
  - Let $\mathbf{d}$ denote the vector from $\mathcal{H}$ to $\mathbf{x}$ of minimum length
  - let $\mathbf{x}^P$ be the projection of $\mathbf{x}$ onto $\mathcal{H}$.



$w^T x + b = 0$

Margin

- The **geometric margin** γ is defined as the distance from the hyperplane to the *closest* point across both classes.
- **Let us first measure the distance of a point $\mathbf{x}$ to the hyperplane $\mathcal{H}$:**



projection $\mathbf{x}^P = \mathbf{x} - \mathbf{d}$. min dist.

$\mathbf{d}$ is parallel to $\mathbf{w}$, so $\mathbf{d} = \alpha\mathbf{w}$ for some $\alpha \in \mathbb{R}$.

$\mathbf{x}^P \in \mathcal{H}$ which implies $\mathbf{w}^T\mathbf{x}^P + b = 0$

therefore $\mathbf{w}^T\mathbf{x}^P + b = \mathbf{w}^T(\mathbf{x} - \mathbf{d}) + b = \mathbf{w}^T(\mathbf{x} - \alpha\mathbf{w}) + b = 0$

which implies $\alpha = \frac{\mathbf{w}^T\mathbf{x}+b}{\mathbf{w}^T\mathbf{w}}$

- Thus, the **distance** (length of $\mathbf{d}$) is:

$$\|\mathbf{d}\|_2 = \sqrt{\mathbf{d}^T\mathbf{d}} = \sqrt{\alpha^2\mathbf{w}^T\mathbf{w}} = \frac{|\mathbf{w}^T\mathbf{x}+b|}{\sqrt{\mathbf{w}^T\mathbf{w}}} = \frac{|\mathbf{w}^T\mathbf{x}+b|}{\|\mathbf{w}\|_2}$$

The distance is a *function* of the parameters as well as the point itself.
b w & x

- The **geometric margin** γ is given by the minimum distance from the training set D.

$$\gamma(\mathbf{w}, b) = \min_{\mathbf{x}\in D} \frac{|\mathbf{w}^T\mathbf{x}+b|}{\|\mathbf{w}\|_2}$$

The margin (or hyperplane) is scale invariant: $\gamma(\beta\mathbf{w}, \beta b) = \gamma(\mathbf{w}, b), \forall\beta \neq 0$

Max Margin Classifier (SVM)

- Now, **SVM** that searches for the **maximum margin separating hyperplane** can be formulated as a **constrained optimization** problem.

- The **objective** is to maximize the margin under the constraints that all data points must lie on the correct side of the hyperplane (*correctly classified*):

$$\underbrace{\max_{\mathbf{w},b} \gamma(\mathbf{w},b)}_{maximize\ margin} \text{ such that } \underbrace{\forall i\ y_i(\mathbf{w}^T x_i + b) \geq 0}_{separating\ hyperplane}$$

↳ seperate linear classifiers

- If we plug in the definition of γ we obtain:

$$\underbrace{\max_{\mathbf{w},b} \underbrace{\frac{1}{\|\mathbf{w}\|_2} \min_{\mathbf{x}_i \in D} \left|\mathbf{w}^T \mathbf{x}_i + b\right|}_{\gamma(\mathbf{w},b)}}_{maximize\ margin} \quad s.t. \quad \underbrace{\forall i\ y_i(\mathbf{w}^T x_i + b) \geq 0}_{separating\ hyperplane}$$

↳ geometrical margin

Direct solution of this optimization problem would be *very complex*.

Max Margin Classifier (SVM)

$$\boxed{\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2} \min_{\mathbf{x}_i \in D} \left|\mathbf{w}^T\mathbf{x}_i + b\right| \quad s.t. \;\; \forall i\; y_i(\mathbf{w}^T x_i + b) \geq 0}$$

- Let us convert it into an *equivalent* problem that is much easier to solve:
  - Because the hyperplane is **scale invariant**, we can re-scale $\mathbf{w}, b$ anyway we want.
  - Let's be clever about it, and choose it such that (or any *positive* number on the right side):

$$\min_{\mathbf{x}\in D} \left|\mathbf{w}^T\mathbf{x} + b\right| = 1.$$

  - We can add this re-scaling as an equality constraint. Then our objective becomes:

$$\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2} \cdot 1 = \min_{\mathbf{w},b} \|\mathbf{w}\|_2 = \min_{\mathbf{w},b} \mathbf{w}^\top\mathbf{w} \qquad \text{s.t.} \quad \begin{aligned} \forall i,\; y_i(\mathbf{w}^T\mathbf{x}_i + b) &\geq 0 \\ \min_i \left|\mathbf{w}^T\mathbf{x}_i + b\right| &= 1 \end{aligned}$$

- The *new* optimization problem becomes:

$$\boxed{\min_{\mathbf{w},b} \mathbf{w}^\top\mathbf{w} \quad \text{s.t.} \quad \forall i\; y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1}$$

Max Margin Classifier (SVM)

$$\boxed{\min_{\mathbf{w},b} \mathbf{w}^\top \mathbf{w} \quad \text{s.t.} \quad \forall i \; y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1}$$

- This new formulation is a quadratic optimization problem (the objective is *quadratic* and the constraints are all *linear*) and there is a *unique* minimum (convex).

- The choice of 1 on the right hand side is arbitrary (*any positive* number would do).

- In the lecture notes, we have

$$\min \frac{1}{2} \|\theta\|^2 \;\text{ subject to } y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1, t = 1, \ldots, n$$

  - The factor of 1/2 is added for computation convenience and doesn't affect the optimal parameters.

SVM - Geometrical Interpretation

$$\boxed{\min \frac{1}{2}\|\theta\|^2 \text{ subject to } y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1, t = 1, \ldots, n}$$

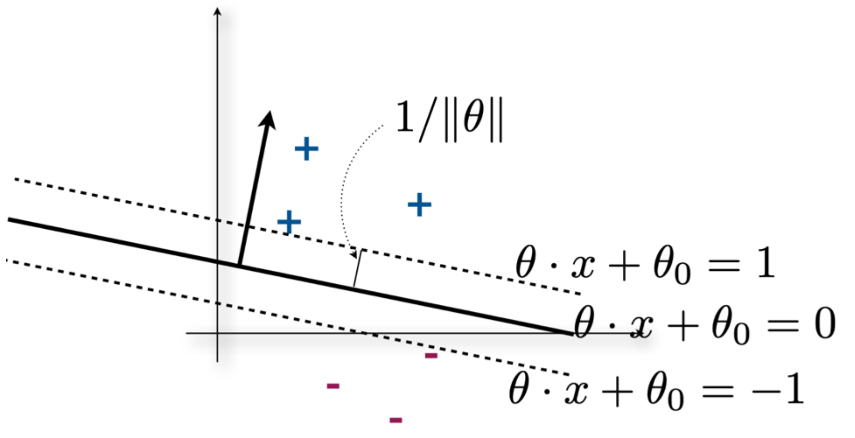
- Two **margin boundaries** are parallel to the decision boundary:

$$\theta \cdot x + \theta_0 = 1 \text{ and } \theta \cdot x + \theta_0 = -1$$

- The two **margin boundaries** lie exactly $1/\|\theta\|$ away from the decision boundary, on the opposite sides.
  - Let $x^{(0)}$ be a point that is exactly on the decision boundary, and let u = $\theta/\|\theta\|$ be a unit vector in the direction of θ.
  - Move from $x^{(0)}$ exactly length $1/\|\theta\|$ in the direction of u :

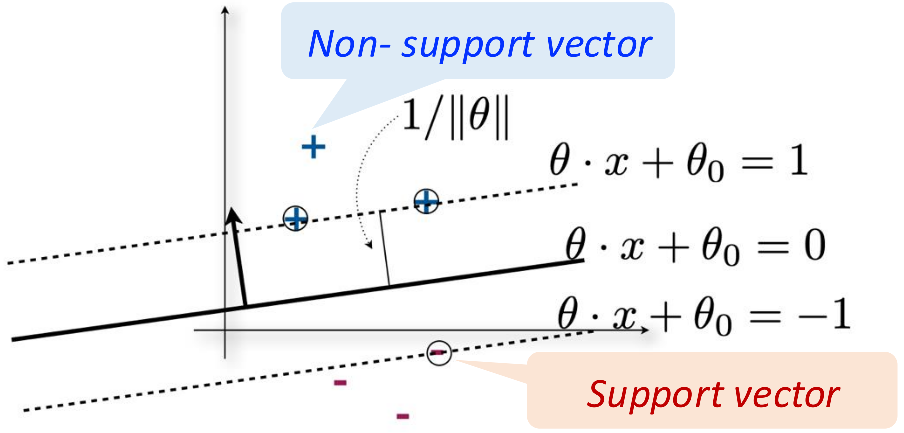
$$\theta \cdot (x^{(0)} + u/\|\theta\|) + \theta_0 = \theta \cdot x^{(0)} + \theta \cdot u/\|\theta\| + \theta_0 = \underbrace{\theta \cdot x^{(0)} + \theta_0}_{=0} + \underbrace{\theta \cdot \theta/\|\theta\|^2}_{=1} = 1$$

- Thus, by minimizing $\|\theta\|$, we push the margin boundaries apart.
- The margin boundaries will arrive the **unique maximum margin** solution *without* violating the classification constraints.



$1/\|\theta\|$
+
+
+
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = 0$
-
-
-
$\theta \cdot x + \theta_0 = -1$

a) A linear separator with margin boundaries that satisfy the classification constraints



Non- support vector
$1/\|\theta\|$
+
$\theta \cdot x + \theta_0 = 1$
$\theta \cdot x + \theta_0 = 0$
$\theta \cdot x + \theta_0 = -1$
-
-
Support vector

b) The **maximum** margin separator satisfying the classification constraints

only need for SVM

Minimizing $||\theta||$: Constrained optimization

$$\boxed{\min \frac{1}{2}\|\theta\|^2 \text{ subject to } y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1, t = 1, \ldots, n}$$

- In many cases, the dimension of the parameters (or examples) is quite large.
- This makes the quadratic programming problem a bit challenging to solve.
- We can, however, solve its *Lagrangian dual problem*.
  - **Lagrangian dual problem** obtained by forming the Lagrangian of a minimization problem that uses nonnegative Lagrange multipliers to add the constraints to the objective function.

Lagrangian Dual Problem
(Background)

Lagrangian Dual Problem

Given a minimization problem, called *primal*

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \quad i = 1, \ldots m \\ & \ell_j(x) = 0, \quad j = 1, \ldots r \end{aligned}$$

we defined the **Lagrangian:**

Lagrange multipliers

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

and **Lagrange dual function:**

$$g(u, v) = \min_{x\in\mathbb{R}^n} L(x, u, v)$$

The subsequent **dual problem** is:

$$\begin{aligned} \max_{u\in\mathbb{R}^m,\, v\in\mathbb{R}^r} \quad & g(u, v) \\ \text{subject to} \quad & u \ge 0 \end{aligned}$$

Important properties:

- Dual problem is always convex, i.e., g is always concave (even if primal problem is not convex)
- The primal and dual optimal values, $f^\star$ and $g^\star$, always satisfy weak duality: $f^\star \ge g^\star$
- Slater's condition: for convex primal, if there is an x such that

  $$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

  then **strong duality** holds: $f^\star = g^\star$. (Can be further refined to strict inequalities over nonaffine h_i, $i = 1, \ldots m$)

Source: CMU Optimization 10-725

Karush-Kuhn-Tucker conditions

Given general problem *(primal)*

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \ldots m \\ & \ell_j(x) = 0, \quad j = 1, \ldots r \end{aligned}$$

The subsequent **dual problem** is:

$$\begin{aligned} \max_{u \in \mathbb{R}^m, v \in \mathbb{R}^r} \; & \min_{x \in \mathbb{R}^n} L(x, u, v) \\ \text{subject to} \; & u \geq 0 \end{aligned}$$

where $L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$

Source: CMU Optimization 10-725

The **Karush-Kuhn-Tucker conditions** or **KKT conditions** are:

- $0 \in \partial f(x) + \sum_{i=1}^{m} u_i \partial h_i(x) + \sum_{j=1}^{r} v_j \partial \ell_j(x)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0, \; \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)

***KKT conditions** are first-order necessary conditions for solving constrained optimization problems*

$x^\star$ and $u^\star, v^\star$ are primal and dual solutions

$\Rightarrow$ $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

$x^\star$ and $u^\star, v^\star$ are primal and dual solutions

$\Leftrightarrow$ $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions

***KKT conditions** are necessary and sufficient when strong duality holds.*

Dual Problem for Solving SVM

Solving SVM

$1 - y^t (\theta - x^t)$

Let us drop the bias parameter θ_0 for simplicity.

- **Primal**

$$\boxed{\min \frac{1}{2}\|\theta\|^2 \text{ subject to } y^{(t)}(\theta \cdot x^{(t)}) \geq 1, t = 1, \ldots, n}$$

- Introduce Lagrange multipliers $\alpha_1 \geq 0, \ldots, \alpha_n \geq 0$ (serve to enforce the classification constraints)
- The **Lagrangian function**:

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{t=1}^{n} \alpha_t [1 - y^{(t)}(\theta \cdot x^{(t)})]$$

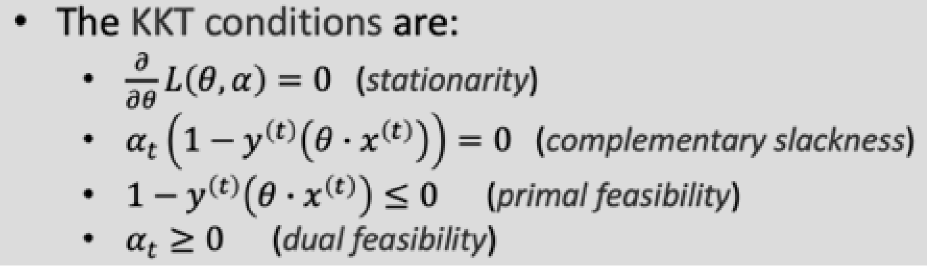
$$\max_{\{\alpha_i\}} \min_{\theta} L(\theta, \alpha)$$

- The KKT conditions are:
  - $\frac{\partial}{\partial \theta} L(\theta, \alpha) = 0$ (*stationarity*) $\Rightarrow$ $\frac{\partial}{\partial \theta} L(\theta, \alpha) = \theta - \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)} = 0$ $\Rightarrow$ $\theta = \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)}$
  - $\alpha_t \left(1 - y^{(t)}\left(\theta \cdot x^{(t)}\right)\right) = 0$ (*complementary slackness*)
  - $1 - y^{(t)}\left(\theta \cdot x^{(t)}\right) \leq 0$ (*primal feasibility*)
  - $\alpha_t \geq 0$ (*dual feasibility*)

Solving SVM

- Plugging $\theta = \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)}$ back into the Lagrangian function $L(\theta,\alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{t=1}^{n} \alpha_t [1 - y^{(t)}(\theta \cdot x^{(t)})]$
- We obtain the **dual** problem:

$$\boxed{\begin{array}{l}\max \sum_{t=1}^{n} \alpha_t - \frac{1}{2} \sum_{t=1}^{n} \sum_{t'=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} (x^{(t)} \cdot x^{(t')}) \\ \text{subject to } \alpha_t \geq 0, t = 1, \ldots, n\end{array}}$$

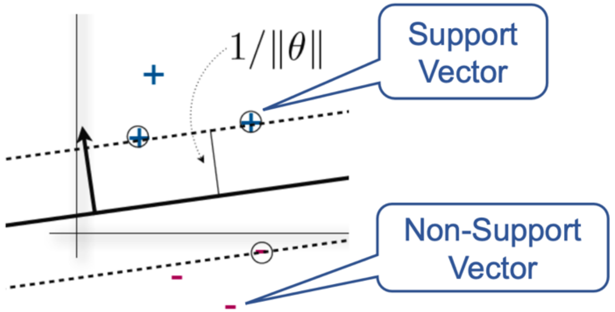


- The KKT conditions are:
  - $\frac{\partial}{\partial\theta} L(\theta,\alpha) = 0$ *(stationarity)*
  - $\alpha_t \left(1 - y^{(t)}(\theta \cdot x^{(t)})\right) = 0$ *(complementary slackness)*
  - $1 - y^{(t)}(\theta \cdot x^{(t)}) \leq 0$ *(primal feasibility)*
  - $\alpha_t \geq 0$ *(dual feasibility)*

- After solving the dual to get the optimal $\hat{\alpha}_t$, we obtain the optimal $\hat{\theta} = \sum_{t=1}^{n} \hat{\alpha}_t y^{(t)} x^{(t)}$
  - The solution satisfies the *complementary slackness* constraints:

$$\hat{\alpha}_t > 0 : y^{(t)} \left(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')} x^{(t')} \right) \cdot x^{(t)} = 1 \qquad \text{(support vector)}$$

$$\hat{\alpha}_t = 0 : y^{(t)} \left(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')} x^{(t')} \right) \cdot x^{(t)} \geq 1 \qquad \text{(non-support vector)}$$



Linear SVM with Offset

Solving Linear SVM with Offset

- **Primal**

Let us put the bias parameter θ_0 back.

$$\boxed{\min \frac{1}{2}\|\theta\|^2 \ \text{subject to}\ y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \geq 1, t = 1, \ldots, n}$$

- Introduce Lagrange multipliers $\alpha_1 \geq 0$, …., $\alpha_n \geq 0$, our **Lagrangian function** now is:

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{t=1}^{n} \alpha_t [1 - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)]$$

- The KKT conditions are:
  - $\frac{\partial}{\partial \theta} L(\theta, \theta_0, \alpha) = 0$ (*stationarity*) ➡ $\theta = \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)}$
  - $\frac{\partial}{\partial \theta_0} L(\theta, \theta_0, \alpha) = 0$ (*stationarity*) ➡ $\sum_{t=1}^{n} \alpha_t y^{(t)} = 0$
  - $\alpha_t \left(1 - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)\right) = 0$ (*complementary slackness*)
  - $1 - y^{(t)}(\theta \cdot x^{(t)} + \theta_0) \leq 0$ (*primal feasibility*)
  - $\alpha_t \geq 0$ (*dual feasibility*)

Solving SVM with Offset

- Plugging $\theta = \sum_{t=1}^{n} \alpha_t y^{(t)} x^{(t)}$ back into the Lagrangian function $L(\theta,\alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{t=1}^{n} \alpha_t[1 - y^{(t)}(\theta \cdot x^{(t)} + \theta_0)]$
- We obtain the **dual** problem:

$$\max \sum_{t=1}^{n} \alpha_t - \frac{1}{2}\sum_{t=1}^{n}\sum_{t'=1}^{n} \alpha_t \alpha_{t'} y^{(t)} y^{(t')} (x^{(t)} \cdot x^{(t')})$$

$$\text{subject to } \alpha_t \geq 0, t = 1, \ldots, n \text{ and } \sum_{t=1}^{n} \alpha_t y^{(t)} = 0$$

Solving $\hat{\alpha}_t$ →

$$\hat{\theta} = \sum_{t=1}^{n} \hat{\alpha}_t y^{(t)} x^{(t)}$$

- θ_0 does not appear anywhere in the dual. How do we get $\hat{\theta}_0$?
  - After we solve the dual, for support vectors ($\hat{\alpha}_t > 0$) the classification constraints must be satisfied with equality (according to the *complementary slackness* condition):

$$y^{(t)}(\hat{\theta} \cdot x^{(t)} + \hat{\theta}_0) = y^{(t)}(\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')} (x^{(t')} \cdot x^{(t)}) + \hat{\theta}_0) = 1$$

  - As $y^{(t)} \in \{-1,1\}$ and $(y^{(t)})^2 = 1$, we can multiply both sides by $y^{(t)}$ and get

$$\hat{\theta}_0 = y^{(t)} - (\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')} (x^{(t')} \cdot x^{(t)}))$$

$$\text{subject to } \hat{\alpha}_t > 0$$

→ more stable version

$$\hat{\theta}_0 = \frac{1}{|V|} \sum_{t \in V} \left\{ y^{(t)} - (\sum_{t'=1}^{n} \hat{\alpha}_{t'} y^{(t')} (x^{(t')} \cdot x^{(t)})) \right\}, \quad V = \{i | \hat{\alpha}_i > 0\}$$

Acknowledgements

- Some slides and content of this lecture are adopted from:
  - MIT 6.036 Introduction to Machine Learning
  - SUTD 50.007 Machine Learning (Asst Prof. Malika Meghjani & Berrak Sisman)
  - Cornell CS4780 Machine Learning for Intelligent Systems
  - McGill COMP-652 Machine Learning
  - Oxford C19 Machine Learning
  - Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. (C7)
  - Murphy K. P. (2012). Machine learning: a probabilistic perspective. MIT Press. (C14)