MAT3007 Tutorial1

School of Data Science
The Chinese University of Hong Kong, Shenzhen

Information

- TA Information:
  - Name: Xinyang Feng
  - Email:120090445@link.cuhk.edu.cn
  - Office Hour: Tuesday 4:00-5:00pm, Zhixin 411
- Tutorial Information:
  - T01: Tuesday and Thursday 6:00-6:50 pm, Teaching A 101 .
  - T02: Tuesday and Thursday 7:00-7:50 pm, Teaching A 101.

Support Vector Machines — Setup

General Setup:

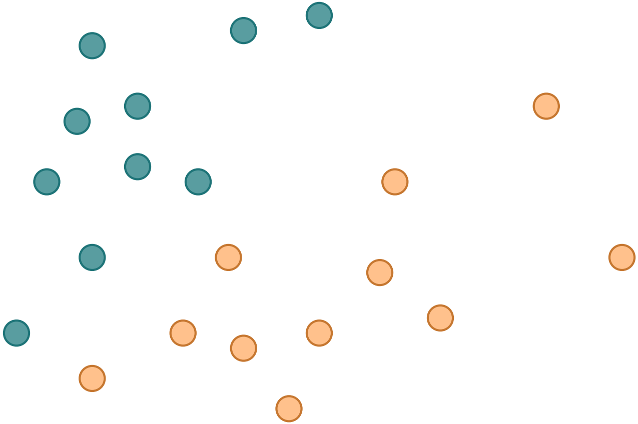
- Given: m objects represented by vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m \in \mathbb{R}^n$ with labels $b_i \in \{-1, 1\}$.
- The two labels ± 1 indicate that the data can be separated into two classes $A \equiv +1$ and $B \equiv -1$.
- **Idea**: Learn a function $\ell : \mathbb{R}^n \to \{-1, 1\}$ based on the training samples $(\boldsymbol{a}_1, b_1), \ldots, (\boldsymbol{a}_m, b_m)$.

$\rightsquigarrow$ Predict the label of a new object $\boldsymbol{a}$ via $\ell(\boldsymbol{a})$.

- **Example:** Blind taste — Does the taste determine the color?

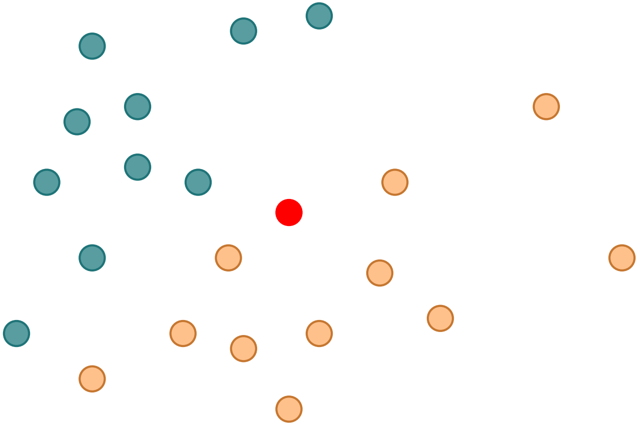
$$\boldsymbol{a}_i \equiv \begin{bmatrix} \text{juicy/fresh} \\ \text{body} \\ \text{acidity} \\ \vdots \end{bmatrix}, \quad b_i = \begin{cases} +1 & \text{white wine,} \\ -1 & \text{red wine.} \end{cases}$$

Classification: Illustration



- Consider two labeled sets of points (green and orange).

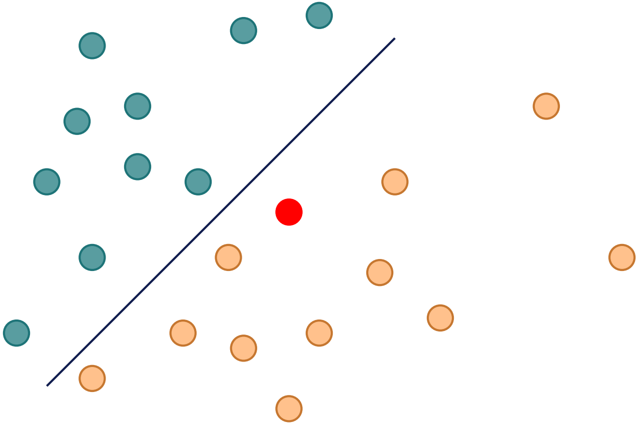
Classification: Illustration



- Consider two labeled sets of points (green and orange).

$\rightsquigarrow$ **Question:** Can we predict the label of a newly added point?

Classification: Illustration



- **Idea:** Separate the two data sets with a hyperplane!

Linear Classification

Decision: Find a hyperplane $\ell(\boldsymbol{a}) := \boldsymbol{x}^\top \boldsymbol{a} + y$ defined by $(\boldsymbol{x}, y)$ separating the datapoints such that:

$$b_i = \begin{cases} +1 & \text{if } \ell(\boldsymbol{a}_i) > 0, \\ -1 & \text{if } \ell(\boldsymbol{a}_i) \leq 0. \end{cases}$$

This is equivalent to choosing $(\boldsymbol{x}, y)$ such that:

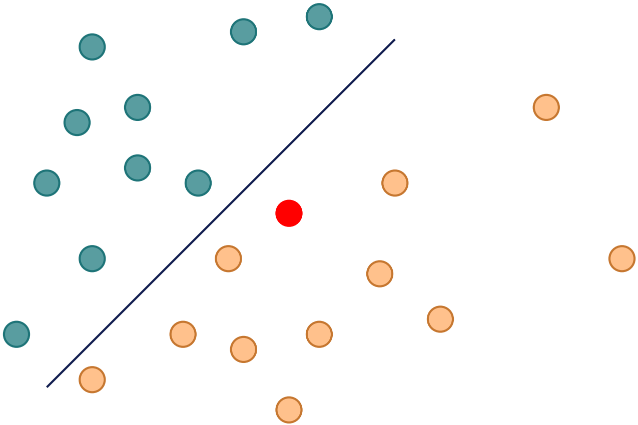
$$b_i = \begin{cases} +1 & \text{if } \ell(\boldsymbol{a}_i) \geq +1, \\ -1 & \text{if } \ell(\boldsymbol{a}_i) \leq -1, \end{cases} \quad \forall\ i = 1, \ldots, m.$$

The associated optimization problem is given by:

$$\text{minimize}_{\boldsymbol{x}, y}\ 0 \quad \text{s.t.} \quad b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y) \geq 1, \quad \forall\ i.$$

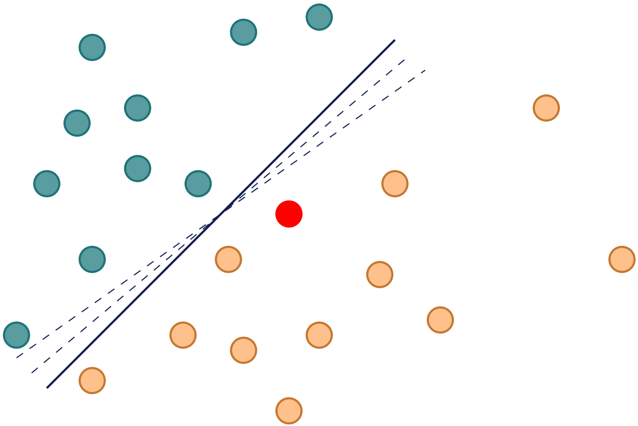
- This problem is a feasibility problem: feasibility problems are a special kind of optimization problem.
- SVMs are used in pattern recognition, machine learning, etc.

Support Vector Machines: Illustration



- Is this a good optimization problem or formulation?

Support Vector Machines: Illustration



- Is this a good optimization problem or formulation?
- **Problem:** The separating hyperplane might not be unique!

Support Vector Machines: Model Improvement

- Select the "best" hyperplane to separate the two groups.
- Distance between the hyperplanes $\{\boldsymbol{a} : \boldsymbol{x}^\top \boldsymbol{a} + y = 1\}$ and $\{\boldsymbol{a} : \boldsymbol{x}^\top \boldsymbol{a} + y = -1\}$ is $2/\|\boldsymbol{x}\|$ (why?).

Maximize the possible margin (distance between the datasets):

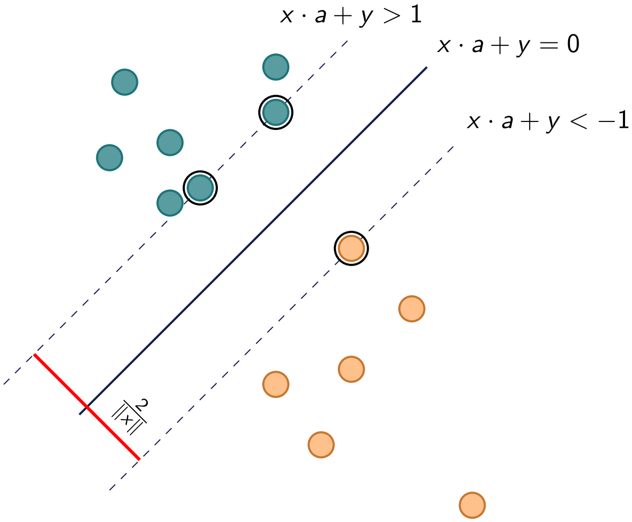
$$\text{maximize}_{\boldsymbol{x},y} \quad \frac{2}{\|\boldsymbol{x}\|} \quad \text{s.t.} \quad b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y) \geq 1, \quad \forall\, i.$$

Compact and equivalent formulation:

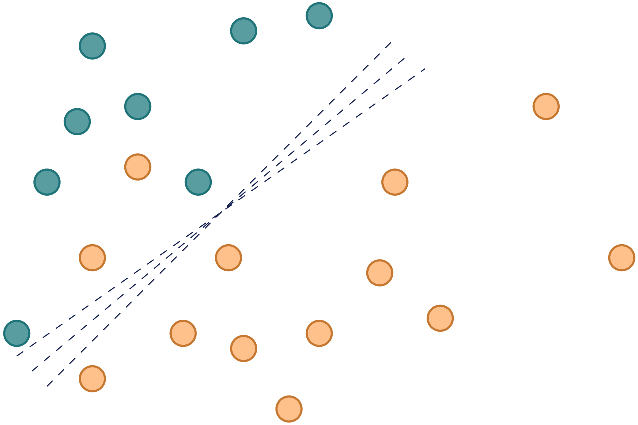
$$\text{minimize}_{\boldsymbol{x},y} \quad \frac{1}{2}\|\boldsymbol{x}\|^2 \quad \text{s.t.} \quad b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y) \geq 1, \quad \forall\, i.$$

- We prefer to use $\|\boldsymbol{x}\|^2$ instead of $\|\boldsymbol{x}\|$ because $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|$ is not differentiable at 0. ($\rightsquigarrow$ Later!).
- Nonlinear (quadratic objective), constrained, continuous.

Support Vector Machines: Illustration



Support Vector Machines: Misclassification



- **Question:** What to do if the training set can not be perfectly separated by a hyperplane?

Handling Misclassification

Strategy and the Full SVM Problem:

⇝ Try to minimize the total or misclassification error:

$$\sum_{i=1}^{m} \max\{0, 1 - b_i \ell(\boldsymbol{a}_i)\} \quad \text{(Hinge-Loss Function).}$$

- SVM combines large margin and small misclassification error:

$$\min_{\boldsymbol{x},y} \quad \frac{\lambda}{2}\|\boldsymbol{x}\|^2 + \sum_{i=1}^{m} \max\{0, 1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y)\}, \quad \lambda > 0.$$

- λ is chosen to balance margin and misclassification. (Typically: $\lambda = \frac{1}{m} \rightsquigarrow$ fine-tuning ...).
- This is an unconstrained, nonsmooth, nonlinear, continuous problem. Can be extremely large-scale (depending on the data set).
- We now show how to equivalently rewrite it as a linear optimization problem in the case $\lambda = 0$.

A Linear Optimization Formulation for SVMs – I

Define $t_i = \max\{0, 1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y)\} =: (1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y))^+$.

We can first rewrite the SVM problem as follows:

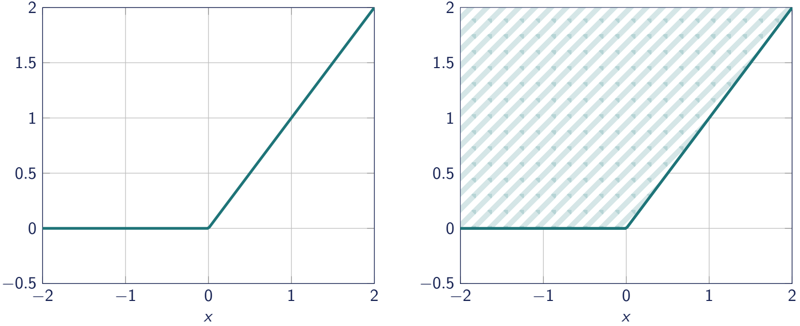
$$\begin{array}{rc} \text{minimize}_{\boldsymbol{x},y,\boldsymbol{t}} & \sum_i t_i \\ \text{subject to} & t_i = (1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y))^+, \quad \forall\ i. \end{array}$$

We claim that we can relax "$=$" to "$\geq$" (why?):

$$\begin{array}{rc} \text{minimize}_{\boldsymbol{x},y,\boldsymbol{t}} & \sum_i t_i \\ \text{subject to} & t_i \geq (1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y))^+, \quad \forall\ i. \end{array}$$

Visualizing the Relaxation "$=$" $\to$ "$\geq$"

$\max\{0, x\} = t \quad \Longrightarrow \quad \max\{0, x\} \leq t$



A Linear Optimization Formulation for SVMs - II

Furthermore, $t_i \geq (1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y))^+$ is equivalent to:

$$t_i \geq 1 - b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y), \qquad t_i \geq 0.$$

Therefore, the optimization problem can be reformulated as:

$$\begin{array}{rll} \text{minimize}_{\boldsymbol{x},y,\boldsymbol{t}} & \sum_i t_i & \\ \text{subject to} & b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y) + t_i \geq 1, & \forall\ i \\ & t_i \geq 0 & \forall\ i. \end{array}$$

- This is a linear optimization problem with decision variables $\boldsymbol{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$, and $\boldsymbol{t} \in \mathbb{R}^n$.

Reformulation Principles

What we have done here:

- Introducing auxiliary variables ($\rightsquigarrow \boldsymbol{t}$).
- Relaxation of the binding constraints.

$\rightsquigarrow$ Minimization of the objective function "pushes" the solution towards the desired direction and original constraints.

- Identifying the correct equivalence.

Similar techniques can be applied to many other linear programming reformulation examples.

Support Vector Machines - Standard LP

$$
\begin{array}{rll}
\text{minimize}_{\boldsymbol{x},y,\boldsymbol{t}} & \sum_i t_i & \\
\text{subject to} & b_i(\boldsymbol{a}_i^\top \boldsymbol{x} + y) + t_i \geq 1, & \forall\ i \\
& t_i \geq 0 & \forall\ i.
\end{array}
$$

- Define $\boldsymbol{x} = \boldsymbol{x}^+ - \boldsymbol{x}^-$, $y = y^+ - y^-$, with $\boldsymbol{x}^+$, $\boldsymbol{x}^- \geq \mathbf{0}$, y^+,$y^- \geq 0$.
- Add slack variables to eliminate inequality constraints.

Standard Form for SVMs

$$
\begin{aligned}
\min_{\boldsymbol{x}^+,\boldsymbol{x}^-,y^+,y^-,\boldsymbol{t},\boldsymbol{s}} \quad & \textstyle\sum_i t_i \\
\text{subject to} \quad & b_i(\boldsymbol{a}_i^\top \boldsymbol{x}^+ - \boldsymbol{a}_i^\top \boldsymbol{x}^- + y^+ - y^-) + t_i - s_i = 1 \quad \forall\ i \\
& \boldsymbol{x}^+, \boldsymbol{x}^- \geq \mathbf{0}, \quad y^+, y^- \geq 0 \\
& t_i, s_i \geq 0 \quad \forall\ i.
\end{aligned}
$$

Exercise 1: Chebyshev center

- Consider a set P described by linear inequality Constraints, i.e. $P = \{x \in \mathbb{R}^n \mid a_i^\top x \leq b_i, i = 1, \ldots, m\}$.
- We are interested in finding a ball with the largest possible radius, which is entirely contained with the set P (the center of this ball is called the **Chebyshev center** of P). Provide a concise formulation for this optimization problem.

Solution to Exercise 1

Analysis:

Mathematically, a ball is defined with two elements: center y and radius r:

- Decision variable: the ball.
- Objective: radius r (as large as possible).
- Constraint: the ball entirely contained with set P:
  - Center y is in P : $a_i^\top y \leq b_i, i = 1, \ldots, m$.
  - The distance between y and the boundary of set P is bigger than radius r:
    - the boundary of set P: $a_i^\top x = b_i, i = 1, \ldots, m$;
    - In $\mathbb{R}^n$, the distance between a point y and $a_i^T x = b_i$: $d = \dfrac{|a_i^\top y - b_i|}{||a_i||}$.

Solution to Exercise 1

- We then have the following formulation:

$$\begin{aligned} \min_{y \in \mathbb{R}^n, r \in \mathbb{R}} \quad & -r \\ \text{s.t.} \quad & a_i^\top y \leq b_i, \qquad i = 1, \ldots, m \\ & r \leq \frac{|a_i^\top y - b_i|}{||a_i||}, i = 1, \ldots, m. \end{aligned}$$

Thank You!