Convex Optimization

Lecture 12: Support Vector Machine

Lecturer: *Dr*. Wan-Lei Zhao

Autumn Semester 2025

Contact: wlzhao@xmu.edu.cn

Outline

Overview of discriminative classifer (1)

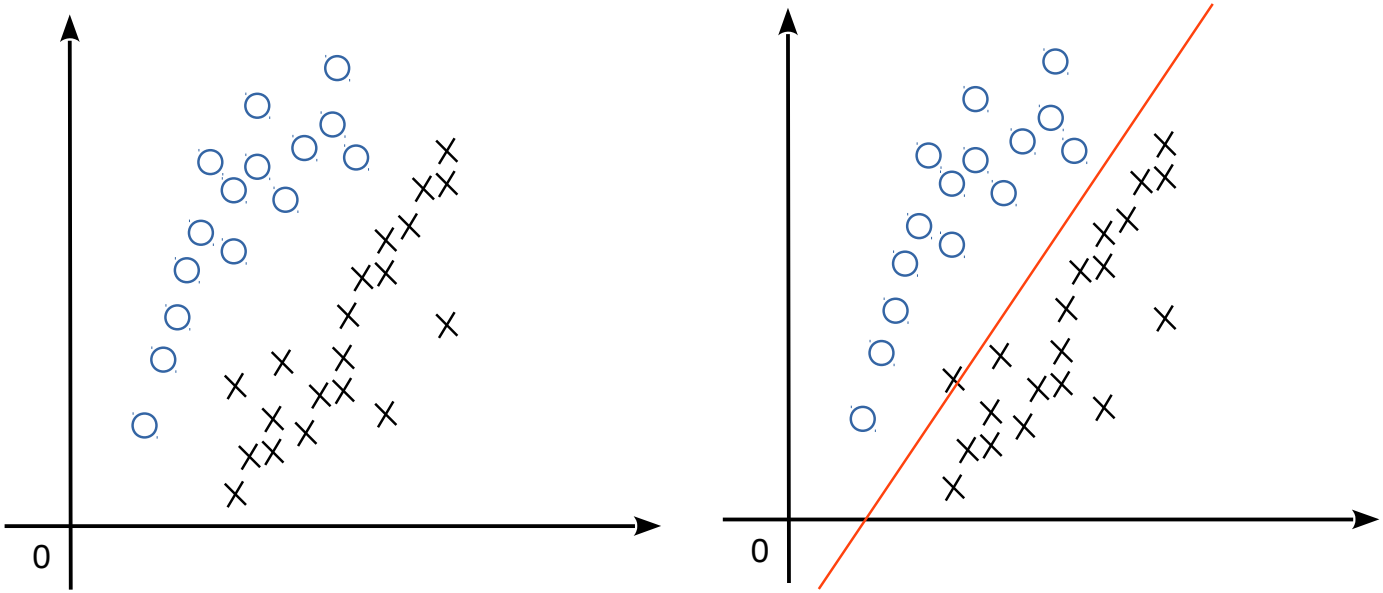
- Given a training set $(x_i, y_i)_{i=1\cdots m}$
- $x_i \in R^n$ is the observation, $y_i \in [-1, 1]$ is the class label
- A classifier is trained with this set
- Given a new instance $u \in R^n$
- The classifier makes the prediction

Overview of discriminative classifer (2)

Figure: Training examples for plane

Figure: Samples to be predicted, plane or not

Openning discussion: linear regression (1)



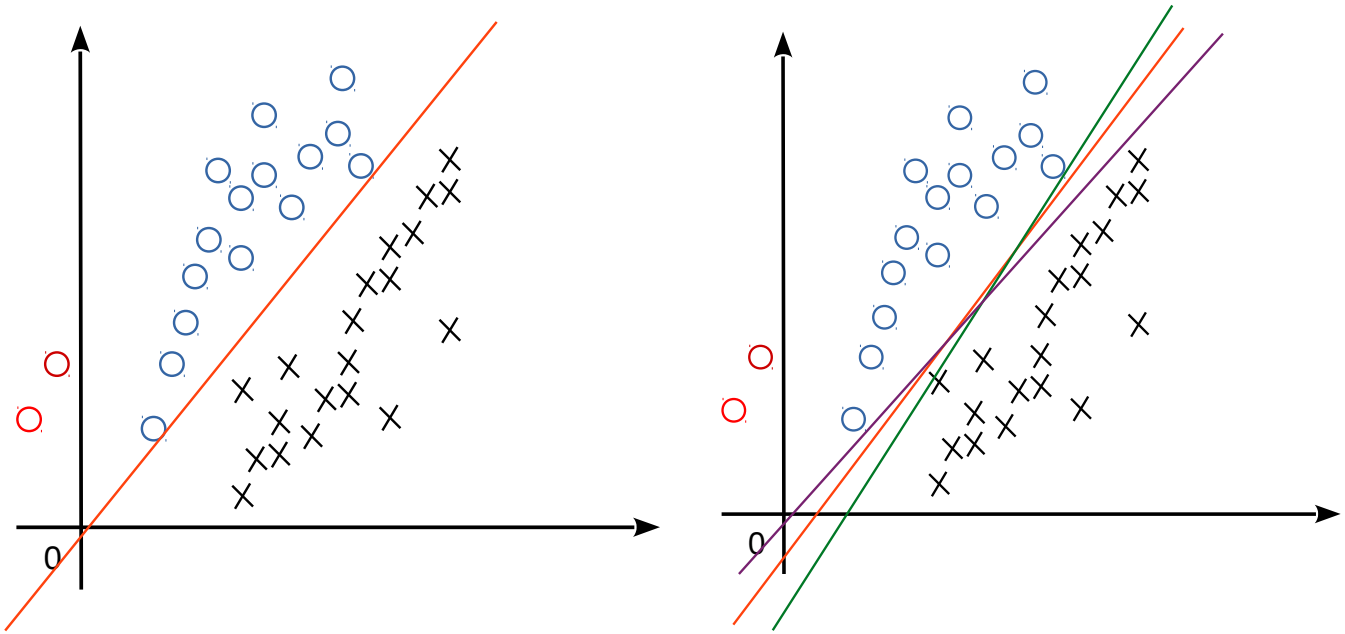
(a) points groups

(b) linear regression on points groups

Figure: Two groups of points

- Linear regression helps to address this problem

Openning discussion: linear regression (2)



(a) sensitive to outliers (b) many possible options

Figure: Two groups of points

- The results are not satisfactory

Discriminative classifer: the binary classification problem

- Let's now start with a simple problem
- Given any instance x, the classifier is asked to make a binary prediction
- Yes or No
- From another point of view, we want to train a function g(.)

$$g(x) = \begin{cases} 1 & c1 \\ -1 & otherwise \end{cases} \tag{1}$$

- To simplify the problem
- We say g(.) is a linear function, that is

$$g(x) = w^T x + b \tag{2}$$

- w and b are parameters to be trained, g(w^Tx)$\in [-1, 1]$
- By default, we are dealing with column vectors

SVM: the functional margin

- Now, it is clear that we want to maximize the margin
- Given $\gamma^{(i)}$, it is defined as

$$\gamma^{(i)} = y^{(i)}(w^T x^{(i)} + b) \tag{3}$$

- if $y^{(i)}$=1, $\gamma^{(i)}$ is going to be large only if $(w^T x^{(i)} + b)$ is large positive
- if $y^{(i)}$=-1, $\gamma^{(i)}$ is going to be large only if $(w^T x^{(i)} + b)$ is large negative
- However, this functional margin is not fully meaningful
- We can make it larger by $(2w^T x^{(i)} + 2b)$

SVM: the minimum functional margin (1)

- With w being normalized, the functional margin $\gamma^{(i)}$ is given as

$$\gamma^{(i)} = y^{(i)}(\frac{w^T}{||w||_2}x^{(i)} + b) \tag{4}$$

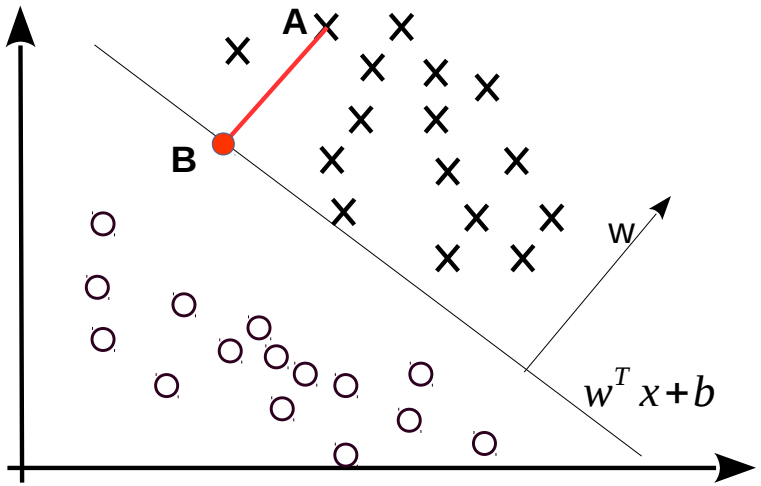


Figure: $\gamma^{(i)}$ is the distance from A to line $w^T x + b$

- It is equal to the distance between **A** and **B**
- Here, we consider the case y_i=1, similar analysis applies to y_i=-1

SVM: the minimum functional margin (2)

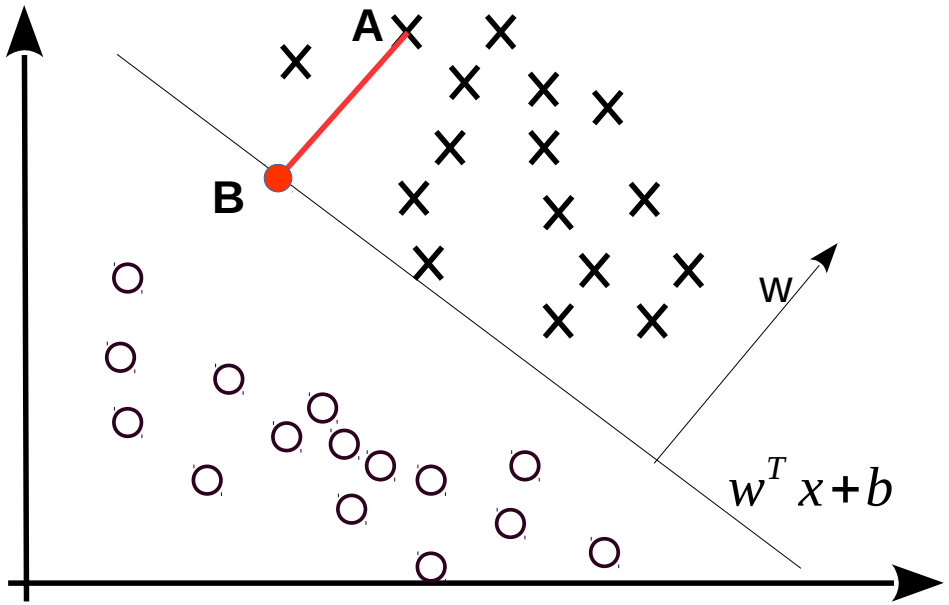


Figure: $\gamma^{(i)}$ is the distance from A to line $w^T x + b$

- Given $A = x^{(i)}$, **B** is on the line $w^T x + b$
- According to the theorem of distance between a point and a line/plain

$$r^{(i)} = \frac{w^T x^{(i)} + b}{\| w \|_2} \tag{5}$$

SVM: the minimum functional margin (3)

- So we have

$$r^{(i)} = \frac{w^T x^{(i)}}{\| w \|_2} + \frac{b}{\| w \|_2} \tag{6}$$

- Above equation is for positive training example
- For all the training examples, we have

$$r^{(i)} = y^{(i)}\left(\frac{w^T}{||w||_2} x^{(i)} + \frac{b}{||w||_2}\right) \tag{7}$$

SVM: the minimum functional margin (4)

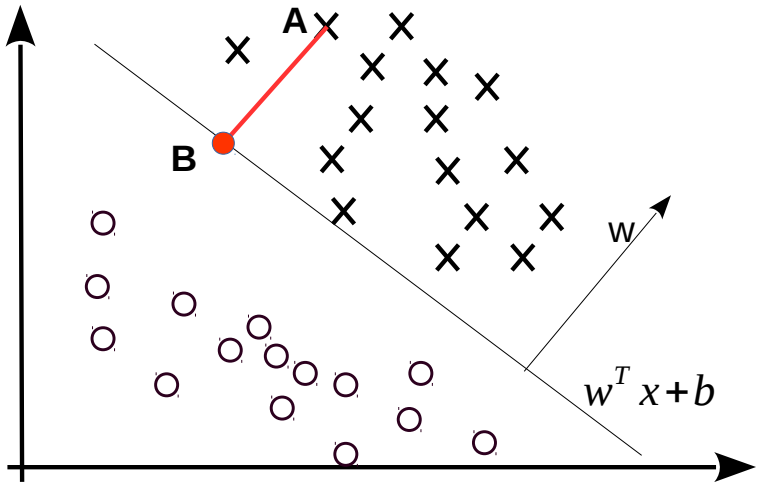


Figure: $\gamma^{(i)}$ is the distance from A to line $\frac{w^T}{||w||_2}x + b$

$$r^{(i)} = y^{(i)}(\frac{w^T}{||w||_2}x^{(i)} + \frac{b}{||w||_2}) \tag{10}$$

- We want to find the minimum margin, and maximize it

$$\gamma = \underset{i=1,\cdots,m}{\text{argmin}}\ \gamma^{(i)} \tag{11}$$

SVM: the model (1)

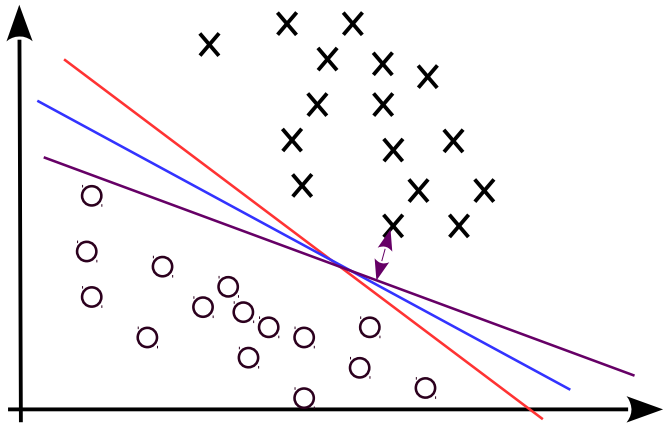


Figure: Searching for maximum γ

$$\gamma = \underset{i=1,\cdots,m}{\text{argmin}}\ \gamma^{(i)} \tag{11}$$

- Searching for w, b that maximizes γ

$$\begin{aligned} \underset{w,b,\gamma}{\text{Max.}} \quad & \gamma \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, \cdots, m \\ & ||w||_2 = 1. \end{aligned}$$

SVM: the model (2)

- Searching for w, b that maximizes γ

$$\begin{aligned} \underset{w,b,\gamma}{\text{Max.}} \quad & \gamma \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, \cdots, m \\ & ||w||_2 = 1. \end{aligned} \tag{11}$$

- Unfortunately, above problem is not solvable
- Constraint $||w||_2 = 1$ is not convex

$$\begin{aligned} \underset{w,b,\hat{\gamma}}{\text{Max.}} \quad & \hat{\gamma}/||w||_2 \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, i = 1, \cdots, m. \end{aligned} \tag{12}$$

- Unfortunately, above problem is not solvable either
- $\hat{\gamma}$ is functional margin, it is valid to scale it to $\hat{\gamma} = 1$

SVM: the model (3)

$$\begin{aligned} \underset{w,b,\hat{\gamma}}{\text{Max.}} \quad & \hat{\gamma}/||w||_2 \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, i = 1, \cdots, m. \end{aligned} \tag{12}$$

- $\hat{\gamma}$ is functional margin, it is valid to scale it to $\hat{\gamma} = 1$

$$\begin{aligned} \underset{w,b}{\text{Max.}} \quad & \frac{1}{||w||_2} \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \cdots, m. \end{aligned} \tag{13}$$

- This is equivalent to solving following **quadratic optimization** problem

$$\begin{aligned} \underset{w,b}{\text{Min.}} \quad & \frac{1}{2}||w||^2 \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \cdots, m. \end{aligned} \tag{14}$$

SVM: the model (4)

- Binary classification problem is now modeled as **quadratic optimization** problem

$$\begin{aligned} \underset{w,b}{\text{Min.}} \quad & \frac{1}{2}||w||^2 \\ \text{s. t.} \quad & 1 - y^{(i)}(w^T x^{(i)} + b) \leq 0, i = 1, \cdots, m. \end{aligned} \tag{14}$$

- The unknowns are w and **b**
- The target function is quadratic
- m constraints are linear
- We are going to solve it with **Lagrange multiplier** method

$$\begin{aligned} \underset{w,b,\gamma}{\text{Max.}} \quad & \gamma \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, \cdots, m \\ & ||w||_2 = 1. \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \underset{w,b,\hat{\gamma}}{\text{Max.}} \quad & \hat{\gamma}/||w||_2 \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, i = 1, \cdots, m. \end{aligned}$$

$$\Downarrow$$

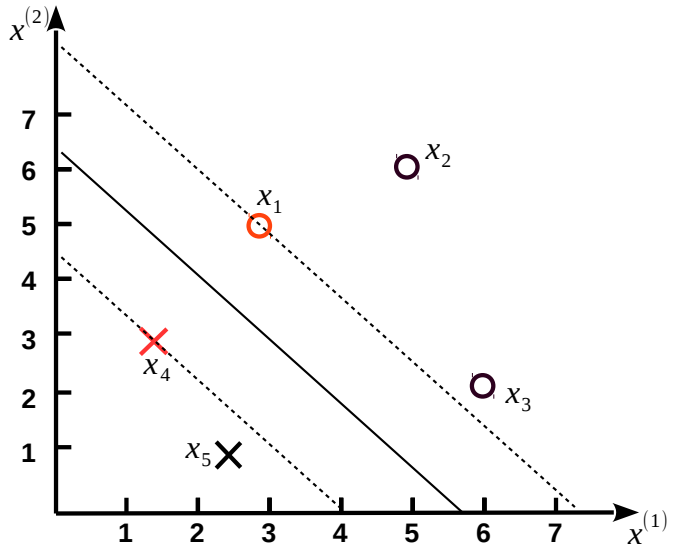
$$\begin{aligned} \underset{w,b}{\text{Max.}} \quad & \frac{1}{||w||_2} \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \cdots, m. \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \underset{w,b}{\text{Min.}} \quad & \frac{1}{2}||w||^2 \\ \text{s. t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \cdots, m. \end{aligned}$$

Outline

Solve it as a QP problem (1)

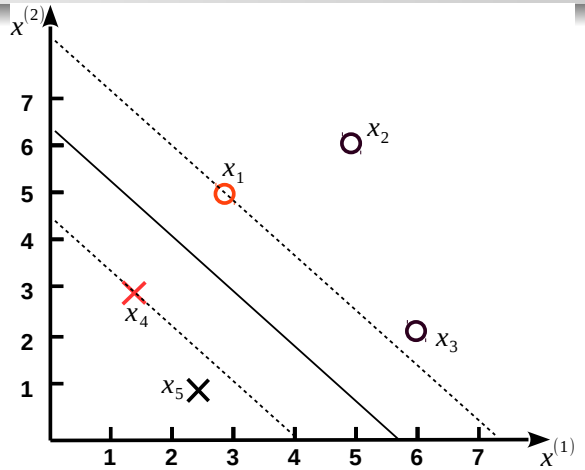


$$\underset{w,b}{\text{Min.}} \quad \frac{1}{2}||w||^2 \tag{1}$$

$$\text{sub. to} \quad 1 - y^{(i)}(w^T x^{(i)} + b) \leq 0, i = 1, \cdots, m.$$

1. The training samples (x_i, y_i)s are all known
2. $w \in R^d$ and $b \in R$ are unknown
3. All constraints are linear inequation, it is QP problem!!

Solve it as a QP problem (2)



$$\min_{w,b} \quad \frac{1}{2}||w||^2 \tag{2}$$

$$\text{sub. to} \quad 1 - y^{(i)}(w^T x^{(i)} + b) \leq 0, i = 1, \cdots, m.$$

$$\Downarrow$$

$$\min_{w,b} \quad \frac{1}{2}||w||^2 \tag{3}$$

$$\text{s. t.} \quad [-y^{(i)}x^{(i)\,T} \quad -y^{(i)}] \begin{bmatrix} w \\ b \end{bmatrix} \cdot \leq -1, i = 1, \cdots, m.$$

Solve it as a QP problem (3)

$$\min_{w,b} \quad \frac{1}{2}||w||^2 \tag{4}$$

$$\text{s. t.} \quad 1 - y^{(i)}(w^T x^{(i)} + b) \leq 0, i = 1, \cdots, m.$$

$$\Downarrow$$

$$\min_{w,b} \quad \frac{1}{2}||w||^2 \tag{5}$$

$$\text{s. t.} \quad [-y^{(i)}x^{(i)\,T} \quad -y^{(i)}] \begin{bmatrix} w \\ b \end{bmatrix} \cdot \leq -1, i = 1, \cdots, m.$$

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

$$\text{s.t.} \quad \begin{bmatrix} -y^{(1)}x_1^{(1)} & -y^{(1)}x_2^{(1)} & -y^{(1)} \\ -y^{(2)}x_1^{(2)} & -y^{(2)}x_2^{(2)} & -y^{(2)} \\ . & . & . \\ -y^{(m)}x_1^{(m)} & -y^{(m)}x_2^{(m)} & -y^{(m)} \end{bmatrix} \cdot \begin{bmatrix} w \\ b \end{bmatrix} \preceq \begin{bmatrix} -1 \\ -1 \\ . \\ -1 \end{bmatrix}$$

Lagrangian duality of SVM (1)

- Given binary classification problem is modeled as

$$\begin{aligned} \underset{w,b}{\text{Min.}} \quad & \frac{1}{2}||w||^2 \\ \text{s. t} \quad & 1 - y^{(i)}(w^T x^{(i)} + b) \leq 0, i = 1, \cdots, m. \end{aligned} \tag{1}$$

- Given $\alpha_i \geq 0$, $i = 1, \cdots, m$, we define

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] \tag{2}$$

- Take derivative on w, we have

$$\nabla_w L(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow \quad w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \tag{3}$$

Lagrangian duality of SVM (2)

- Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] \tag{4}$$

- Take derivative on w, we have

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \tag{5}$$

- Take derivative on b, we have

$$\frac{\partial}{\partial b} L(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \tag{6}$$

- Plug Eqn. 5 and Eqn. 6 into Eqn. 4, we have

$$L(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \tag{7}$$

Lagrangian duality of SVM (3)

- Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] \tag{4}$$

- Notice that we have new form for L(.)

$$L(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \tag{7}$$

- We say this is the dual problem of original L(w, b, α)

$$\begin{aligned} \max_{\alpha} \quad & G(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\ s.t. \quad & \alpha_i \geq 0, i = 1 \cdots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \end{aligned} \tag{8}$$

Lagrangian duality of SVM (4)

- Dual form of Lagrangian:

$$\begin{aligned} \underset{\alpha}{\text{Max.}} \quad & G(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\ s.t. \quad & \alpha_i \geq 0, i = 1 \cdots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \end{aligned} \tag{8}$$

- This problem is easier to solve, only α to be considered
- Conventionally, this problem is solved by **coordinate ascent**[1]
- Remember that once we get α, we have

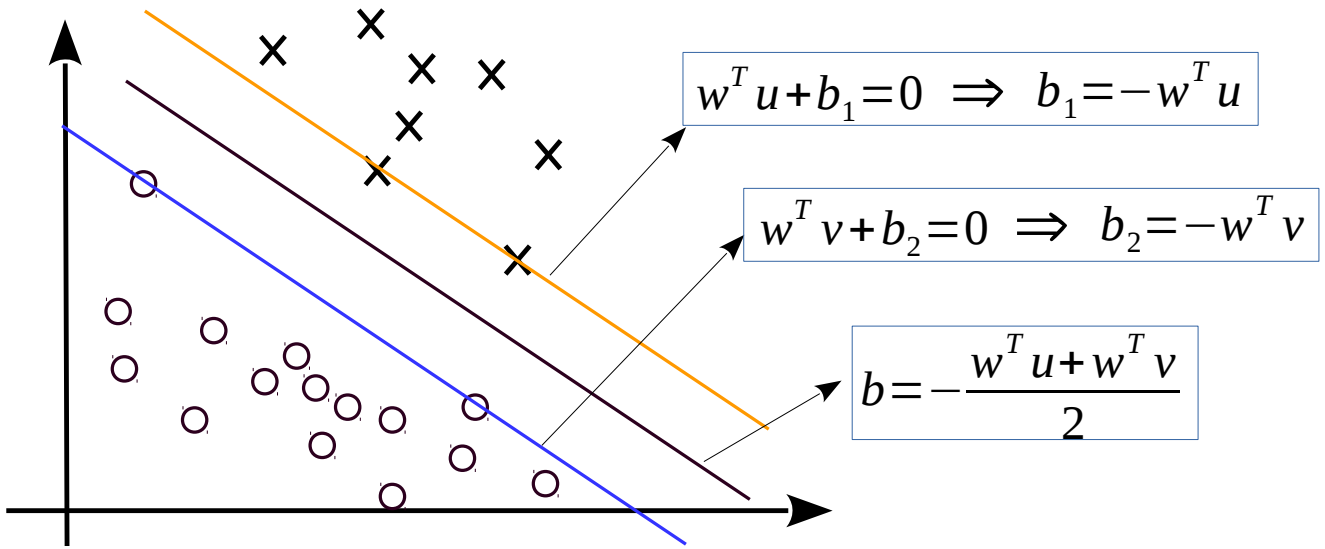
$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \quad b = -\frac{max_{i:y^{(i)}=-1} w^T x^{(i)} + min_{i:y^{(i)}=1} w^T x^{(i)}}{2}$$

[1] Lecture notes on Machine Learning, Andrew Ng. http://cs229.stanford.edu/

Lagrangian duality of SVM (5)

- Solution to the Dual form of Lagrangian:

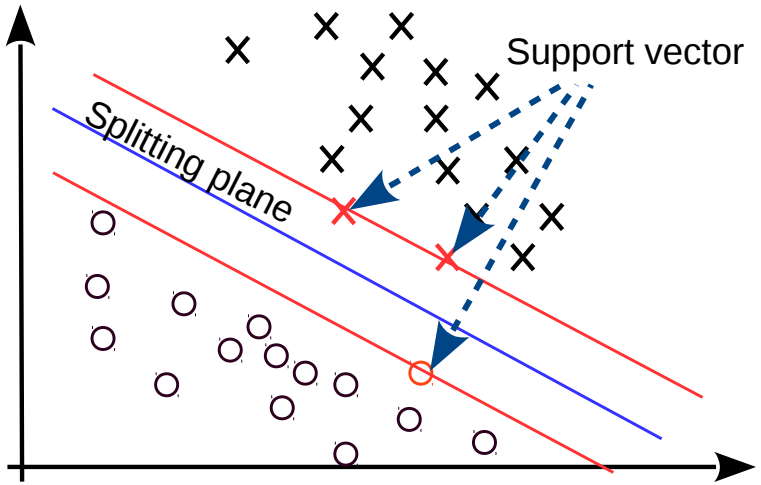
$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \quad b = -\frac{max_{i:y^{(i)}=-1} w^T x^{(i)} + min_{i:y^{(i)}=1} w^T x^{(i)}}{2}$$



SVM: brief procedure of the optimization

$$\begin{aligned} \underset{\alpha}{\text{Max.}} \quad & G(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\ s.t. \quad & \alpha_i \geq 0, i = 1 \cdots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \end{aligned} \tag{8}$$

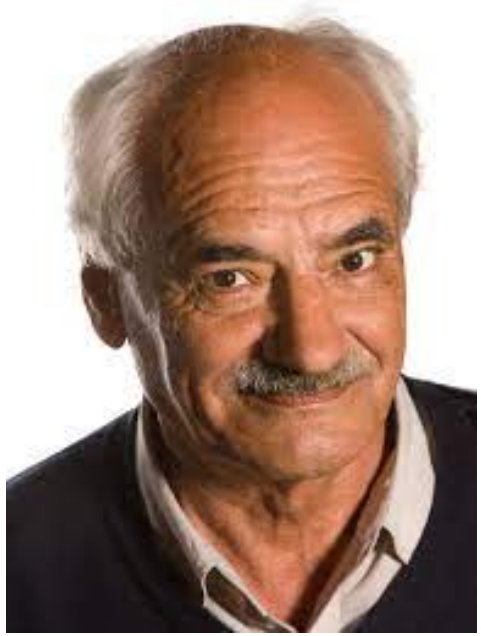
- Coordinate ascent: the optimization process takes α_i and α_j out
- Check whether change them will lead to larger G(.)



- The splitting plane (line) is determined by the support vectors

Comments about SVM

(a) Vladimir Vapnik (1936 -)



(b) A. Chervonenkis (1938 - 2014)

Figure: The scientists invented SVM.

- The most popular classifier before the advent of deep learning
- The theory is still shining

1. Top 10 algorithms in data mining, X. Wu, V. Kumar and et al. Knowledge and Information Systems, 2008, 14(1): 1-37
2. The Nature of Statistical Learning Theory, Vladimir N. Vapnik , Springer-Verlag, 1995.
3. Lecture notes on Machine Learning, Andrew Ng., http://cs229.stanford.edu/