L5: SVM

Shan Wang

Lingnan College, Sun Yat-sen University

2020 Data Mining and Machine Learning LN3119

https://wangshan731.github.io/DM-ML/

Last lecture

- Classification problem
- Logistic regression method
  - Model
  - Strategy
  - Algorithm
- Prediction & evaluation
- Multi-class logistic regression
- Application: diabetes care

Logistic regression revisit

- To predict the quality of care
  - The dependent variable is modelled as a binary variable
  - 1 if low-quality care, 0 if high-quality care
- This is a **categorical variable**
  - Typically a small number of possible outcomes, 2 (low-quality care and high-quality care) in this case
- **Logistic regression** would predict a **probability**

$$p_\theta(y|x) = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}'\boldsymbol{x}}}$$

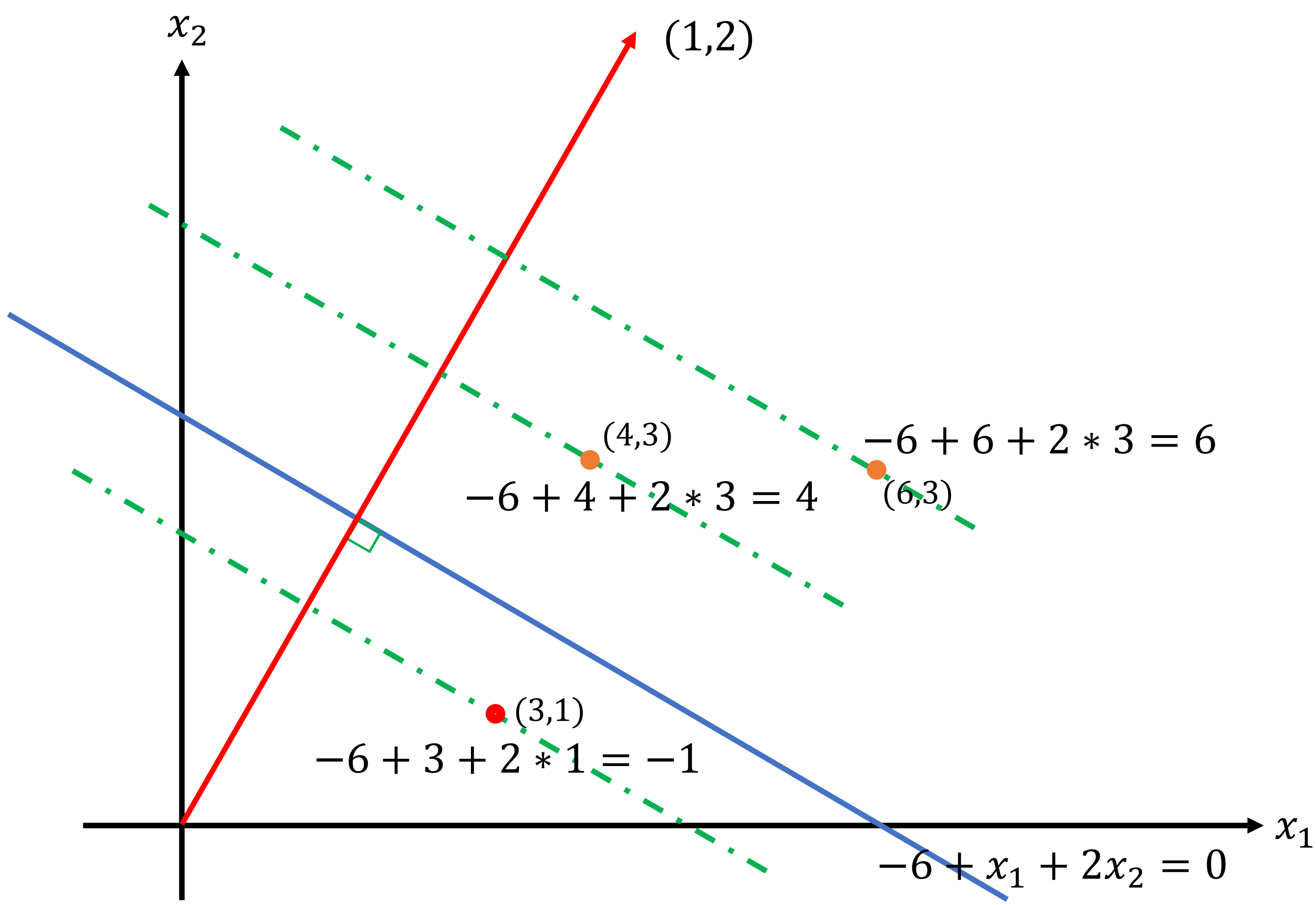
- A threshold is chosen

- **Logistic regression** would predict a **probability**

$$p_\theta(y|x) = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}'\boldsymbol{x}}}$$

- Consider the threshold 0.5, it is easy to check that
  - if $\boldsymbol{\theta}'\boldsymbol{x} > \mathbf{0}$, $p_\theta(y|x) > 0.5$
  - if $\boldsymbol{\theta}'\boldsymbol{x} < \mathbf{0}$, $p_\theta(y|x) < 0.5$
- Recall $\boldsymbol{\theta}'\boldsymbol{x} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_M x_M$
  - $\boldsymbol{\theta}'\boldsymbol{x} = 0$ is a hyperplane which divides the feature space to 2 parts
  - A hyperplane in M dimensions is a flat affine subspace of dimension $M-1$
- Then hyperplane $\boldsymbol{\theta}'\boldsymbol{x} = 0$ is a decision boundary
  - Given a data point $\boldsymbol{x}$, the value of $\boldsymbol{\theta}'\boldsymbol{x}$ is the distance from the data point to the hyperplane
  - Positive distance: 1
  - Negative distance: 0
  - The larger distance, the higher confidence (closer to 1 or 0)

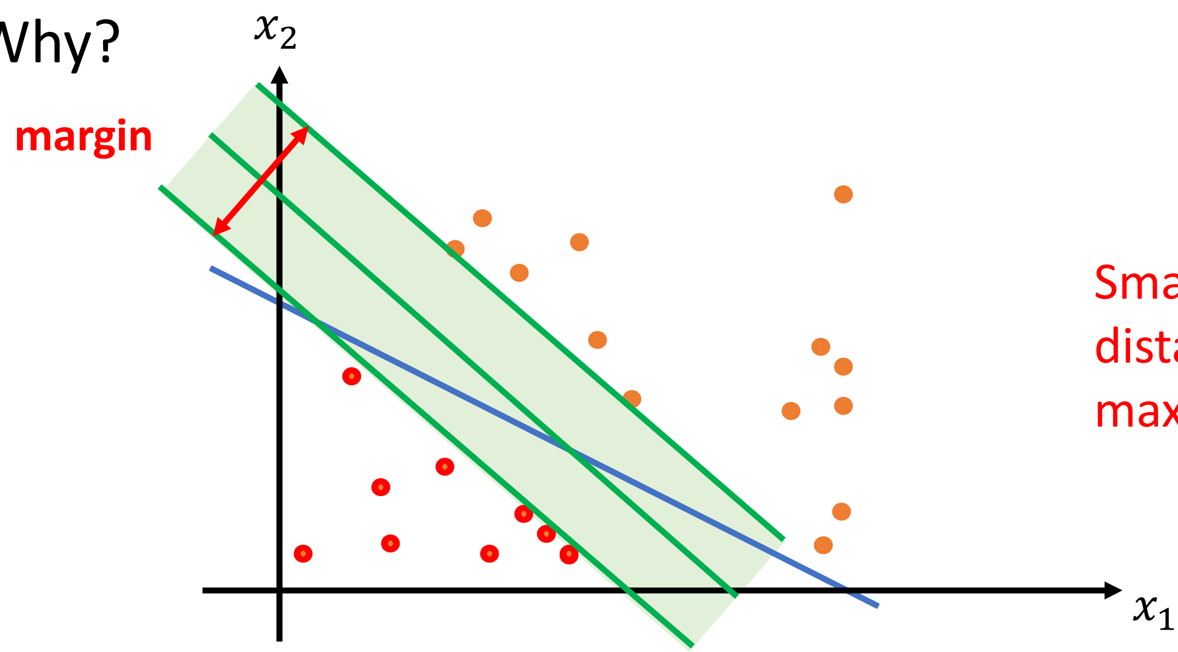
2-dimension example



- Recall the error function in logistic regression
  - Cross entropy

$$\min -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_\theta(1|\boldsymbol{x}_i) + (1-y_i)\log(1-p_\theta(1|\boldsymbol{x}_i))]$$

  - For all data points with label 1, we want $p_\theta(1|\boldsymbol{x}_i)$ is as large as possible
  - For all data points with label 0, we want $p_\theta(1|\boldsymbol{x}_i)$ is as small as possible
- May the green decision boundary be safer?
  - Why?



Smallest distance is maximized!

Course outline

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

This lecture

- Linear SVM
  - Model
  - Strategy
  - Algorithm
- Regularization
- Kernels
- Application: diabetes care revisit

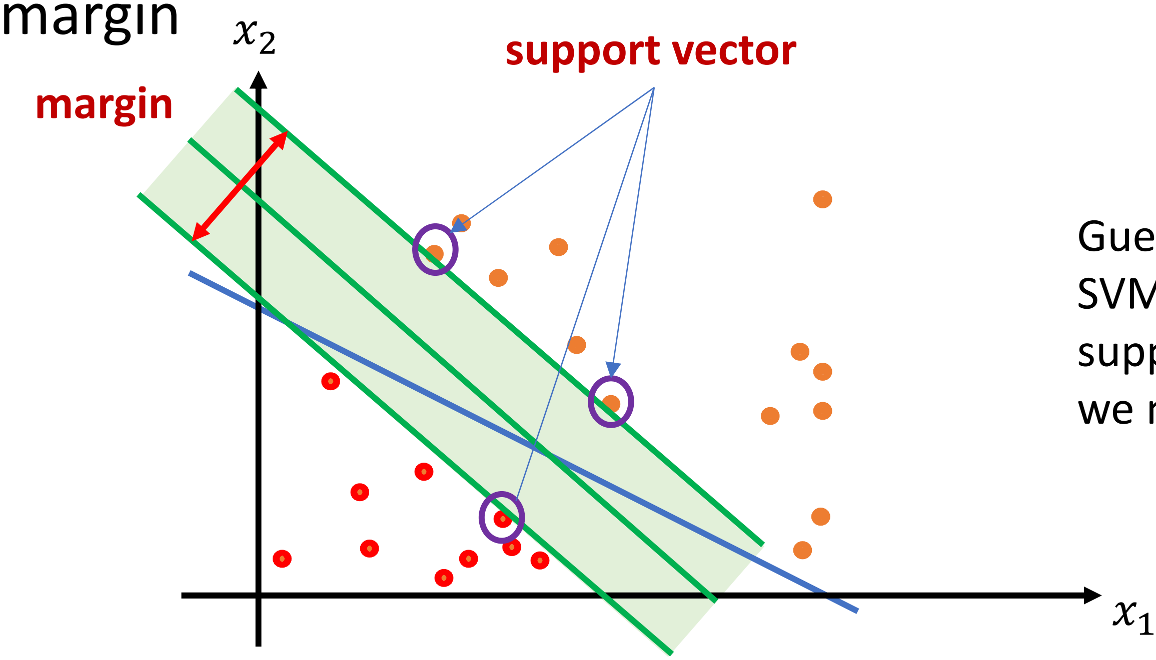
Reference: CS420, Weinan Zhang (SJTU)

Linear SVM

Model

What is support vector machine?

- A linear binary classifier with a decision boundary
  - Which is a separating hyperplane providing maximum margin



Guess: in linear SVM, how many support vectors we need?

- Support vectors: data points that the margin pushes up against

Model

- Feature vector: $\boldsymbol{x} = (x_1, x_2, \dots, x_M)$
- Class label: $y \in \{1, -1\}$
- Parameters
  - Intercept θ_0
  - Feature weight vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_M)$
  - In this lecture, we may drop it if we let extra $x_0 = 1$, and only use $\boldsymbol{\theta}$ to represent all parameters
- Model

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \begin{cases} +1, & \text{if } \boldsymbol{\theta}'\boldsymbol{x} + \theta_0 \geq 0 \\ -1, & \text{if } \boldsymbol{\theta}'\boldsymbol{x} + \theta_0 < 0 \end{cases}$$

  - Decision boundary: $\boldsymbol{\theta}'\boldsymbol{x} + \theta_0 = 0$
  - $y(\boldsymbol{\theta}'\boldsymbol{x} + \theta_0)$ means what?

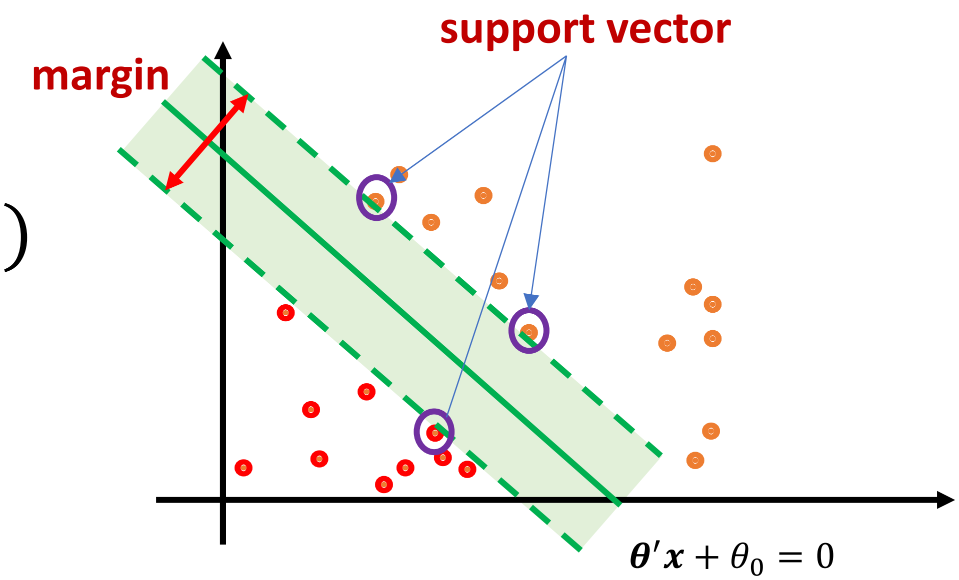
Linear SVM

Strategy

Margin

- Margin:

$$\min_{i=1,..,N} y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0)$$



- For a data point $(\boldsymbol{x_i}, y_i)$
  - When $y_i = 1$, large positive $\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0$ value would give a high confidence
  - When $y_i = -1$, large negative $\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0$ value would give a high confidence
  - $y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) > 0$ means correct prediction

Objective

- Objective: maximize the margin

$$\max_{\boldsymbol{\theta},\theta_0} \min_{i=1,..,N} [y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0)]$$

- Any problem?
  - Unbounded solution, add a constraint

$$\max_{\boldsymbol{\theta},\theta_0} \min_{i=1,..,N} [y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0)]$$

$$s.t.\ \|\boldsymbol{\theta}\| = 1$$

  - Non-convex problem
    - Convex programming: the objective function is convex function, the constraint set is a convex set
  - Transform to a convex problem?

Linear SVM

Algorithm

Convex reformulation

- Equivalent to

$$\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2} ||\boldsymbol{\theta}||^2$$

Quadratic programming

$$s.t.\ \ y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) \geq 1, \qquad i = 1, \dots, N$$

- Why?
  - From math
  - From intuition
    - Original problem: nomarlize the othorgonal vector, maximize the margin
    - Current problem: normarlize the margin to be 1, minimize the norm of the othorgonal vector
    - Functional margin: $y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0)$
    - Geometric margin: $y_i \frac{1}{||\boldsymbol{\theta}||}(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0)$

Lagrangian Dual

$$\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2$$
$$s.t.\ \ y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) \geq 1, i = 1, \ldots, N$$

- Lagrangian dual problem
  - For each constraint: dual variable α_i

$$\max_{\boldsymbol{\alpha}>\mathbf{0}} \min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + \sum_{i=1}^{N} \alpha_i[1 - y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0)]$$

  - Take the FOC on $\boldsymbol{\theta}, \theta_0$
    - $\boldsymbol{\theta} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i} = 0; \qquad \sum_{i=1}^{N} \alpha_i y_i = 0$

$$\max_{\boldsymbol{\alpha}>\mathbf{0}} \frac{1}{2}\left\|\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i}\right\|^2 + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i \left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i}\right)' \boldsymbol{x_i} + \theta_0 \sum_{i=1}^{N} \alpha_i y_i$$

$$\boxed{\begin{gathered}\max_{\boldsymbol{\alpha}\geq\mathbf{0}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_j}'\boldsymbol{x_i} \\ s.t. \sum_{i=1}^{N} \alpha_i y_i = 0\end{gathered}}$$

More on the dual

Prim

$$\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2$$

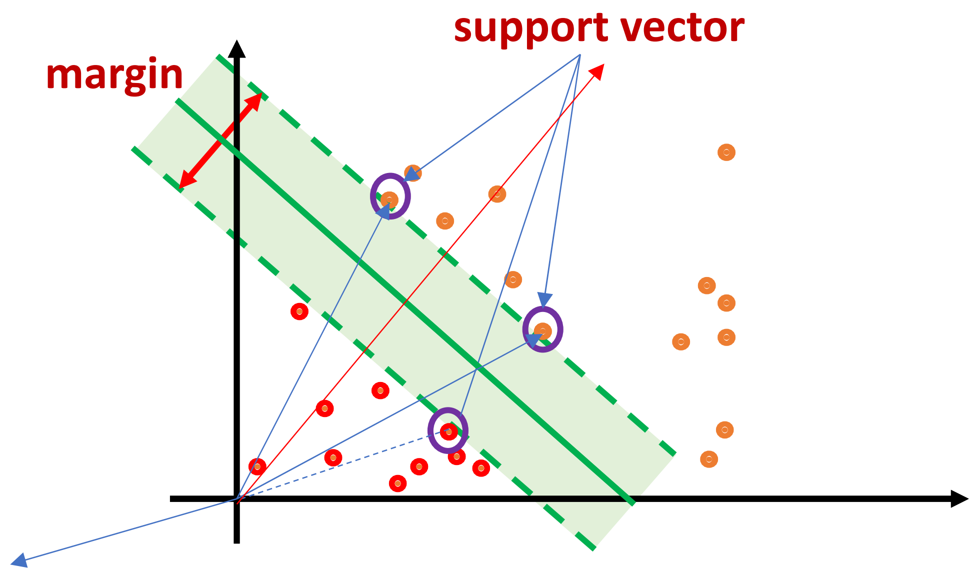
$$s.t.\;\; y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) \geq 1, i = 1, \dots, N$$

Dual

$$\max_{\boldsymbol{\alpha}\geq\mathbf{0}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \boldsymbol{x_j}'\boldsymbol{x_i}$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0$$

- If $y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) = 1, \alpha_i > 0$
- If $y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) > 1, \alpha_i = 0$
- Only support vectors (data points against the margin has positive α_i)



- $\boldsymbol{\theta} = \sum_{i \text{ is support vector}} \alpha_i y_i \boldsymbol{x_i}$
- $\theta_0 = -\frac{1}{2}(\min_{i:y_i=1} \boldsymbol{\theta}'\boldsymbol{x_i} + \max_{j:y_j=-1} \boldsymbol{\theta}'\boldsymbol{x}_j)$

Coordinate ascent algorithm

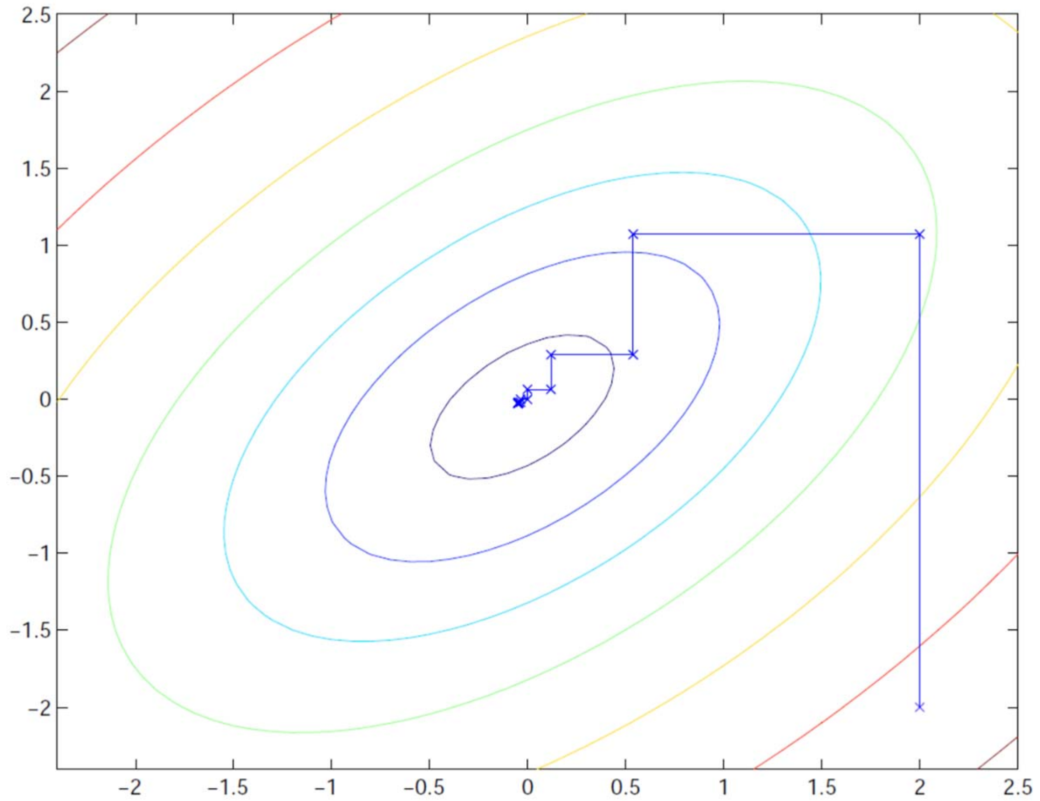
[Coordinate Ascent]

- For optimization problem

$$\max_{\boldsymbol{\alpha}} W(\alpha_1, \alpha_2, \dots, \alpha_N)$$

- Loop until convergence:
  - For $i = 1, \dots, N$:
    - Fix all $\alpha_j \; s.t. \; j \neq i$
    - Update α_i such that,

$$\alpha_i = \operatorname*{argmax}_{\alpha_i} W(\alpha_1, \alpha_2, \dots, \alpha_N)$$



SMO algorithm

- SMO: Sequential Minimal Optimization
- For SVM dual problem: cannot directly apply coordinate ascent algorithm because

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_j}' \boldsymbol{x_i}$$

$$\boxed{s.t. \sum_{i=1}^{N} \alpha_i y_i = 0}$$

- If you fix all α_j $s.t.\, j \neq i$, α_i is also fixed
- Any idea?

SMO algorithm (cont.)

- Update two variables each time
- Loop until convergence:
  - For $i, j = 1, \ldots, N$:
    - Fix all $\alpha_k \; s.t. \, k \neq j, k \neq i$
    - Update α_i, α_j such that, $(\alpha_i, \alpha_j) =$

$$\underset{\alpha_i, \alpha_j \geq \mathbf{0}}{\operatorname{argmax}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_j}' \boldsymbol{x_i}$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0$$

- Convergence test: whether the change of $W(\alpha)$ is smaller than a predefined value (e.g. 0.01)
- Key advantage: the update of α_i, α_j is efficient

SMO algorithm (cont.)

- Without loss of generality, we assume $\alpha_3, \ldots, \alpha_N$ are fixed
- Optimization on α_1, α_2 :

$$\underset{\alpha_i, \alpha_j \geq \mathbf{0}}{\operatorname{argmax}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_j}' \boldsymbol{x_i}$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0$$

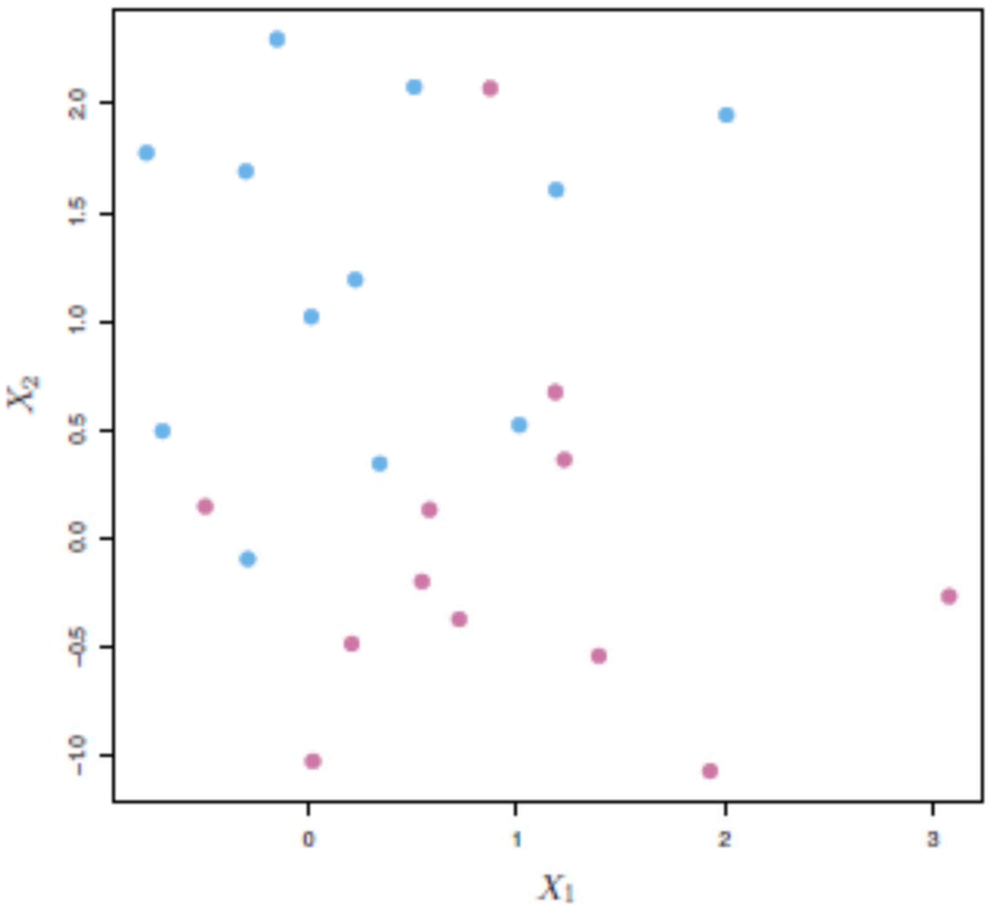
- Rewrite the constraint: let ω denote $-\sum_{i=3}^{N} \alpha_i y_i$
  - $\alpha_2 = \frac{\omega - \alpha_1 y_1}{y_2} = y_2(\omega - \alpha_1 y_1)$
- Maximize a quadratic function
- FOC on α_1:

$$\alpha_1 = \frac{1 - y_1 y_2 - \omega y_1 (\boldsymbol{x_1} - \boldsymbol{x_2})' \boldsymbol{x_2} - \sum_{j=3}^{N} \alpha_j y_j y_1 \boldsymbol{x_j}' (\boldsymbol{x_1} - \boldsymbol{x}_2)}{\boldsymbol{x_1}' \boldsymbol{x_1} + \boldsymbol{x_2}' \boldsymbol{x_2} - 2 \boldsymbol{x_1}' \boldsymbol{x_2}}$$

Regularization

Non-separable Data

- Linear SVM assumes linearly separable data

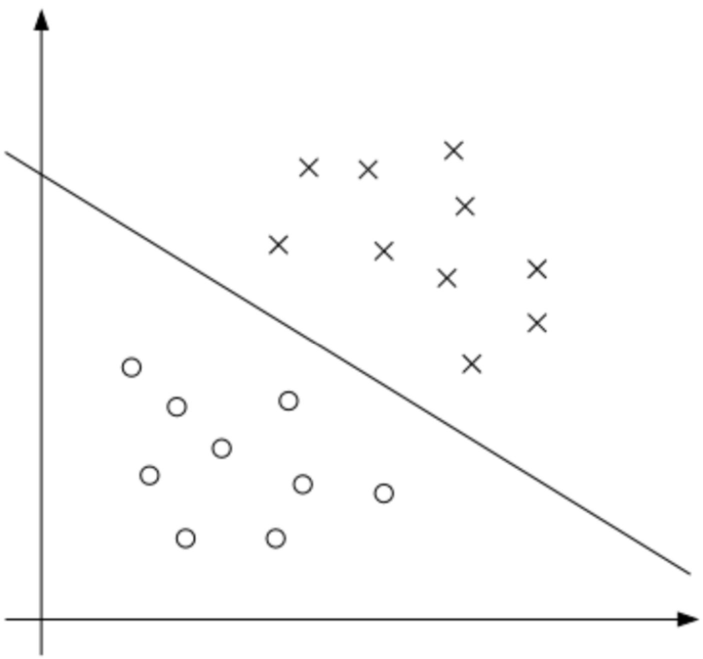


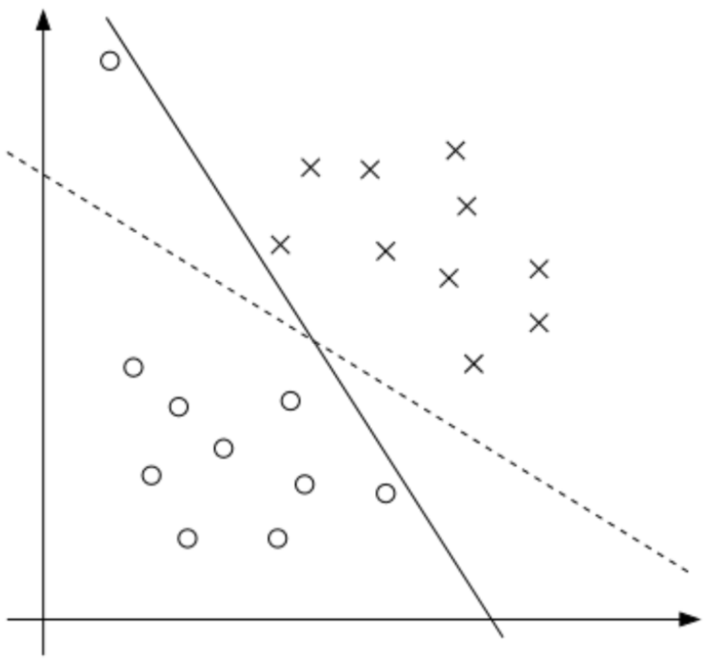
- The data on the left are not separable by a linear boundary

- This is often the case

Noisy Data

- Sometimes the data are separable, but noisy.
- This can lead to a poor solution for the linear SVM.





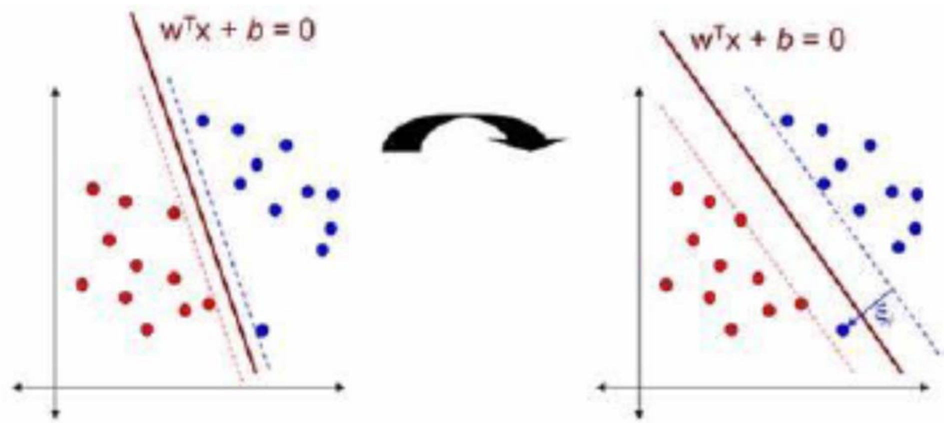
Soft margin

- To make the algorithm
  - work for non-linearly separable datasets
  - be less sensitive to outliers
- Allow some data points violates constraint

$$\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + \boxed{C\sum_{i=1}^{N}\xi_i}$$

L1 regularization

$$s.t. \quad y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) \geq 1 - \xi_i, \qquad i = 1, \ldots, N$$
$$\xi_i \geq 0, \qquad i = 1, \ldots, N$$



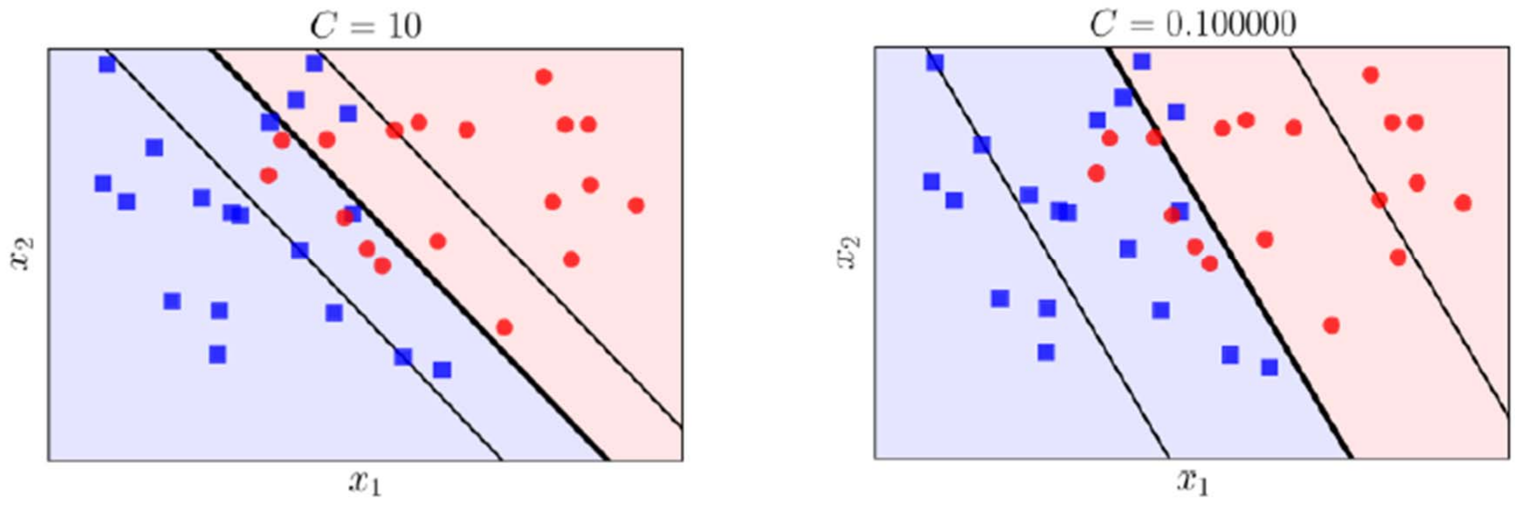
Soft margin (cont.)

$$\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t.\ \ y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) \geq 1 - \xi_i, \qquad i = 1, \dots, N$$

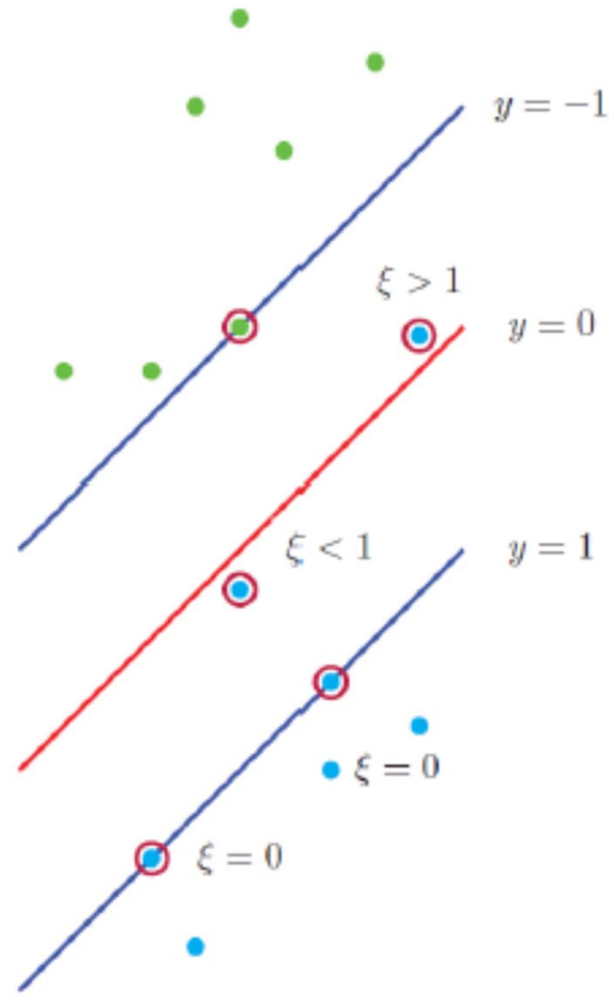
$$\xi_i \geq 0, \qquad i = 1, \dots, N$$

- C is a regularization parameter
  - The smaller C, the softer margin
  - The larger C, the narrower margin



Soft margin (cont.)

- Correctly classified points beyond / on the support line with $\xi = 0$
- Correctly classified points inside the margin with $0 < \xi \leq 1$
- The misclassified points inside the margin with slack $1 < \xi \leq 2$
- The misclassified points outside the margin with slack $\xi > 2$



Lagrangian Dual

Prim

$$\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t.\ y_i(\boldsymbol{\theta}'\boldsymbol{x_i} + \theta_0) \geq 1 - \xi_i, i = 1, \ldots, N$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, N$$

Dual

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x_j}'\boldsymbol{x_i}$$

$$s.t. \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\boxed{0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, N}$$

Compared with the dual of linear SVM, the only change is the upper bound for dual variable

Dual problem: can be efficiently solved by SMO algorithm

Maximize a quadratic function

Bonus question:

Can you derive the dual problem by yourself?

Lecture 5 wrap-up

- ✓ Linear SVM
  - ✓ Model
  - ✓ Strategy
  - ✓ Algorithm
- ✓ Regularization
- Kernels
- Application: diabetes care revisit

Assignment 4

- No official assignment

- One bonus question (NOT required):
  - Derive the Lagrangian Dual for the problem with soft margin
  - Send your answer to TA (any form, e.g., word, pdf, photo …)

- Due: TBD
- TA: Mr. Xiong, xiongym3@mail2.sysu.edu.cn

Questions?

Shan Wang (王杉)

https://wangshan731.github.io/