Support Vector Machines II

COMPCSI 361

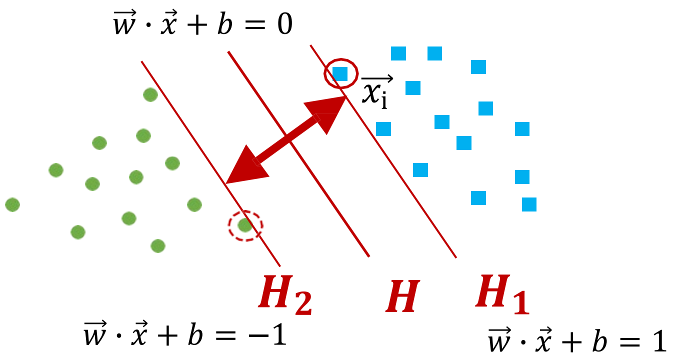
Instructor: Thomas Lacombe

Based on slides from Meng-Fen Chiang

WEEK 9

RECAP: Margin Maximization Hyperplane (MMH)

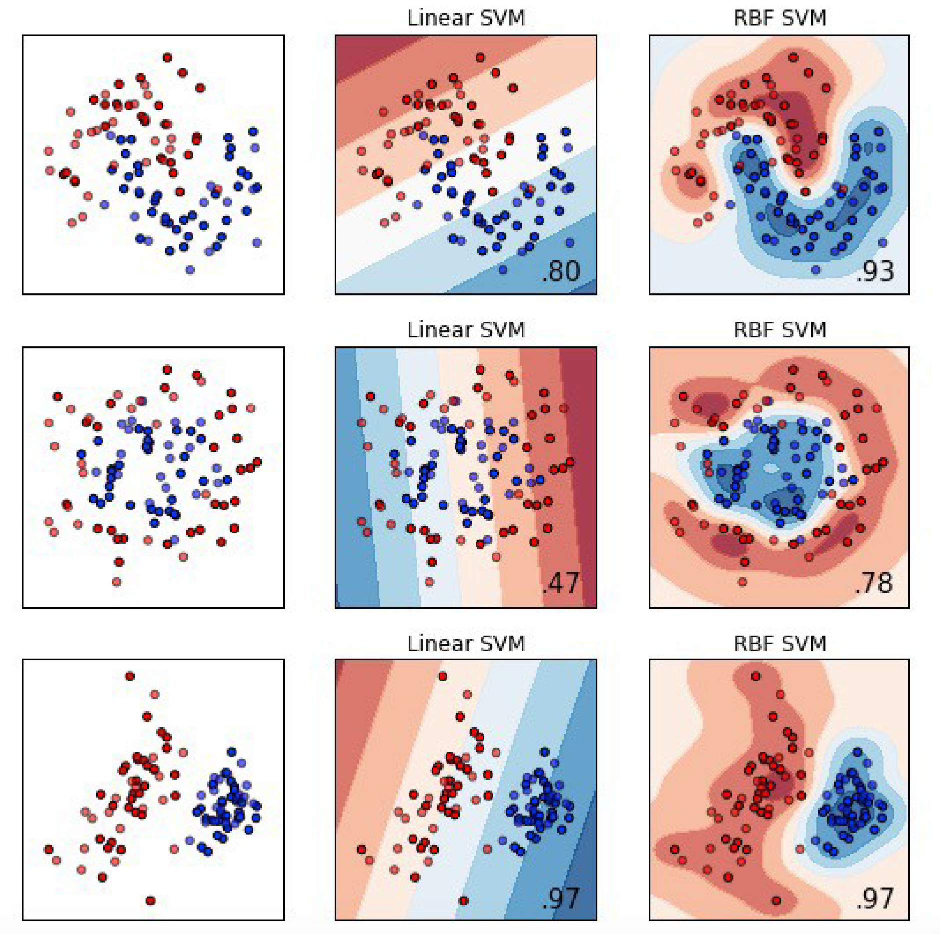
- A separating hyperplane (H) can be formally defined as $\vec{w} \cdot \vec{x} + b = 0$
  - $\vec{w} = \{w_1, w_2, \dots, w_n\}$ is a weight vector and $\boldsymbol{b}$ a scalar (bias)



- For 2-D it can be written as: $\overrightarrow{w_1} \cdot \overrightarrow{x_{i,1}} + \overrightarrow{w_2} \cdot \overrightarrow{x_{i,2}} + b = 0$
- The hyperplanes defining the sides of the margin:
  - $\boldsymbol{H_1}$: $\overrightarrow{w_1} \cdot \overrightarrow{x_{i,1}} + \overrightarrow{w_2} \cdot \overrightarrow{x_{i,2}} + b \geq 1$, for $\mathrm{y_i} = +1$, and
  - $\boldsymbol{H}_2$: $\overrightarrow{w_1} \cdot \overrightarrow{x_{i,1}} + \overrightarrow{w_2} \cdot \overrightarrow{x_{i,2}} + b \leq 1$, for $y_i = -1$
- Any training tuples that fall on margins $\boldsymbol{H_1}$ or $\boldsymbol{H_2}$ (i.e., the hyperplanes defining the margin) are support vectors

OUTLINE

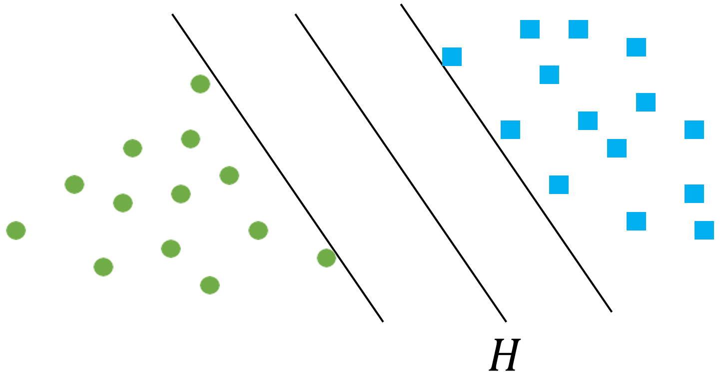
- Data Characteristics
  - Linearly Separable Data
  - Non-Linearly separable Data
- SVM
  - Linearly Separable Data: Hard-margin SVMs (9.1)
  - Non-Linearly Separable Data: Soft-margin SVMs (9.2)
  - Non-Linearly Separable Data: Kernelized SVMs (9.3)
- Summary



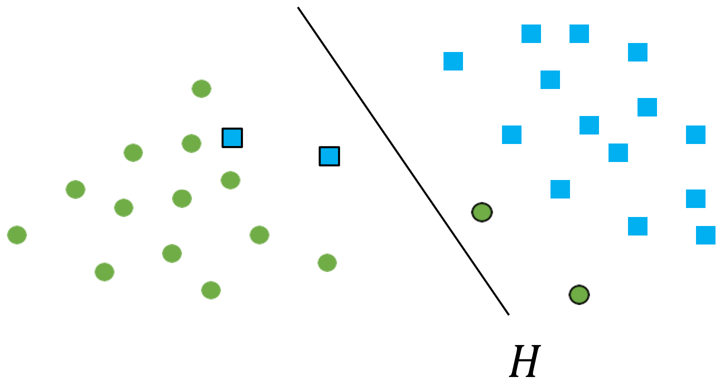
Hard-margin v.s. Soft-margin

- The difference between a hard margin and a soft margin in SVMs lies in the separability of the data.
    1. Case: If our data is linearly separable, we go for a hard margin.
    2. Case: Otherwise, we would have to be more lenient and let some of the data points to be misclassified. In this case, a soft margin SVM is appropriate.

Hard Margin

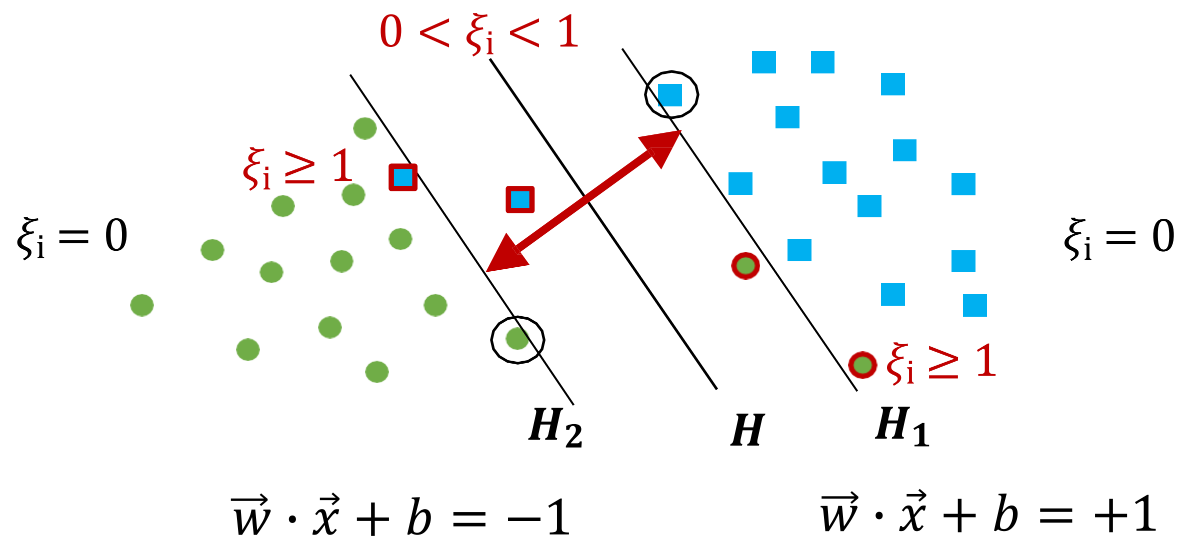


Soft Margin



Problem Definition: Soft-margin Maximization

- Given a set of training data $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big), y_i \in \{+1, -1\}$
- Goal: The soft-margin SVM algorithm aims to find a linear classifier that
  1. Maximizes (γ) the margin on S and
  2. Minimize the misclassification error ($C \sum_{i=1}^{n} \xi_i$)



Soft-margin Maximization

- Soft-margin SVMs introduce a misclassification penalty (C) controls the trade-off between
    1. Maximizing the margin (same as hard-margins)
    2. Minimizing the loss

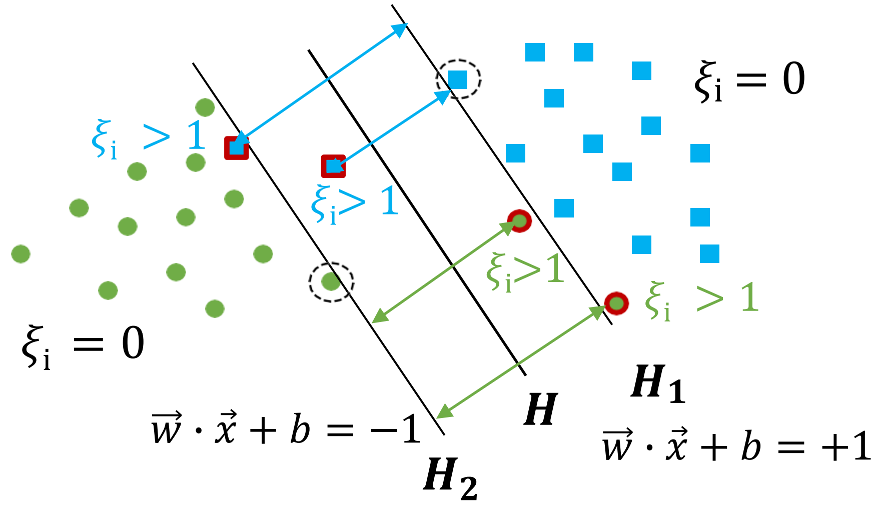
- Primal formulation for the soft-margin

hard-margin

$$\min_{w,b} \frac{\|\vec{w}\|}{2} + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_{\mathrm{i}}(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_{\mathrm{i}}, \quad i = 1, 2, \dots, n, \xi_{\mathrm{i}} \geq 0$$

slack variables (ξ_i)



Slack Variables

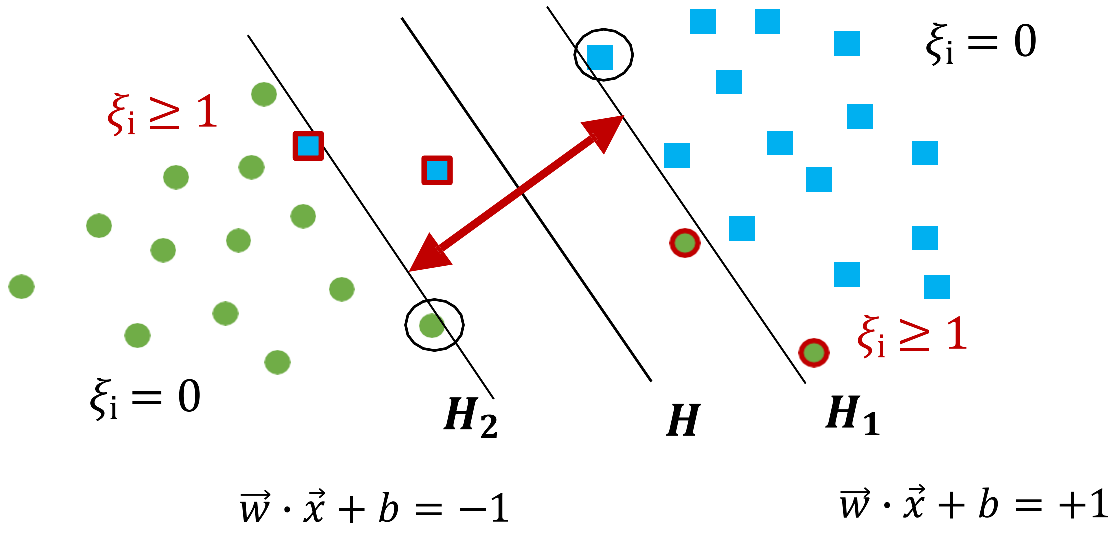
- Value of slack variables $\boldsymbol{\xi_i}$:

$$y_{\mathrm{i}}(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_{\mathrm{i}}, \quad i = 1, 2, \ldots, n, \ \xi_{\mathrm{i}} \geq 0$$

→ If $y_{\mathrm{i}}(\vec{w} \cdot \vec{x_i} + b) \geq 1, \quad \xi_i = 0$ (no penalty)

→ If $y_{\mathrm{i}}(\vec{w} \cdot \vec{x_i} + b) < 1, \quad \xi_i = 1 - y_{\mathrm{i}}(\vec{w} \cdot \vec{x_i} + b)$

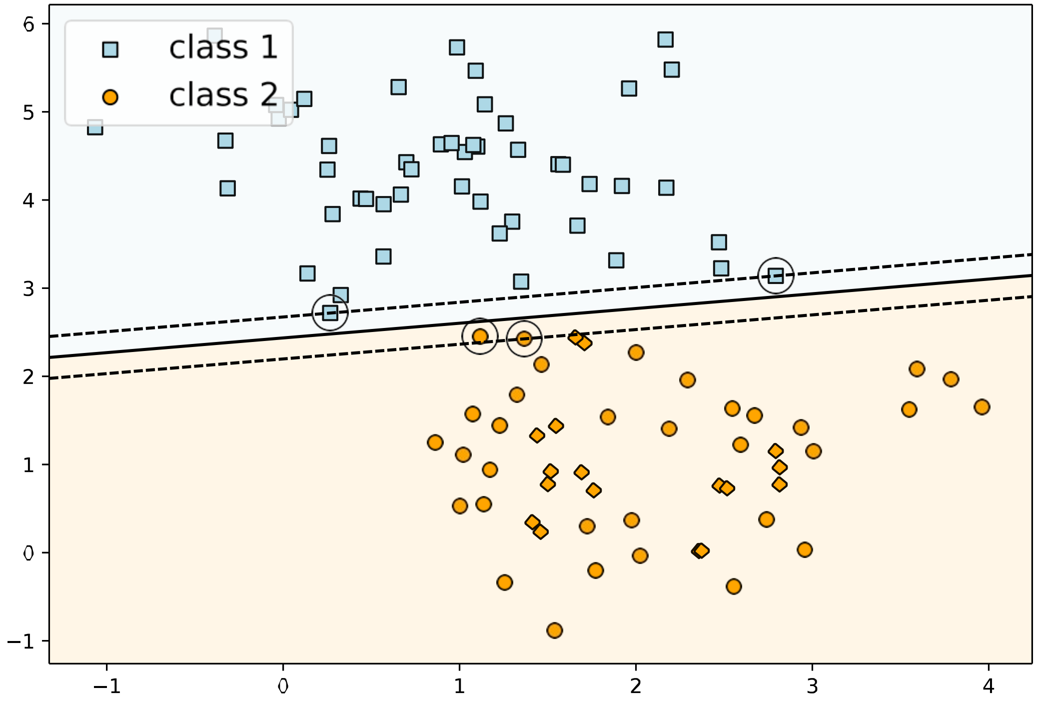
$$\xi_i = \max(0, 1 - y_{\mathrm{i}}(\vec{w} \cdot \vec{x_i} + b))$$



Misclassified data point: $\xi_{\mathrm{i}} > 1$
Data point close to the decision boundary: $0 < \xi_{\mathrm{i}} < 1$

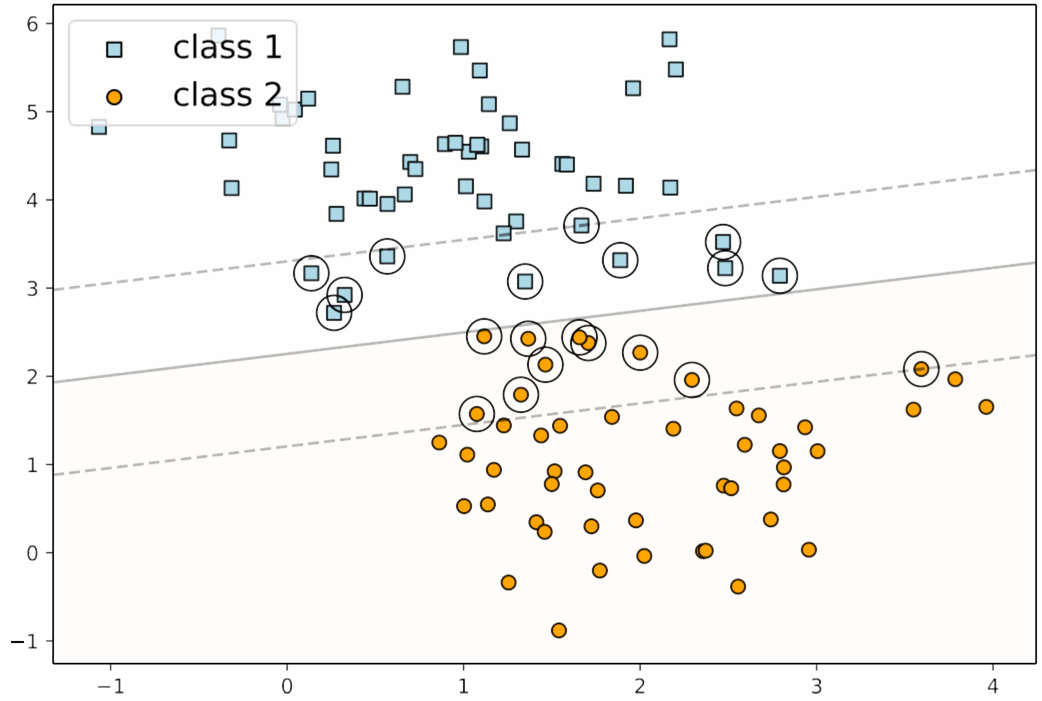
Example: Misclassification Penalty C

C = 10.0



Lower Tolerance to Misclassification Error
Smaller Margin

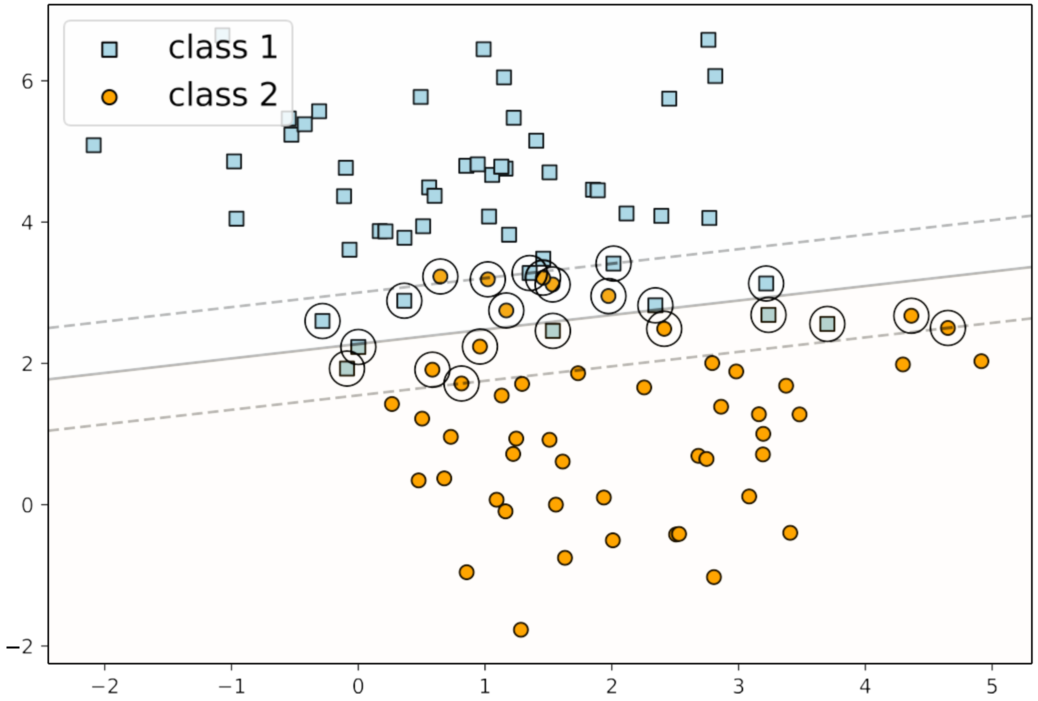
C = 0.1



Higher Tolerance to Misclassification Error
Larger Margin

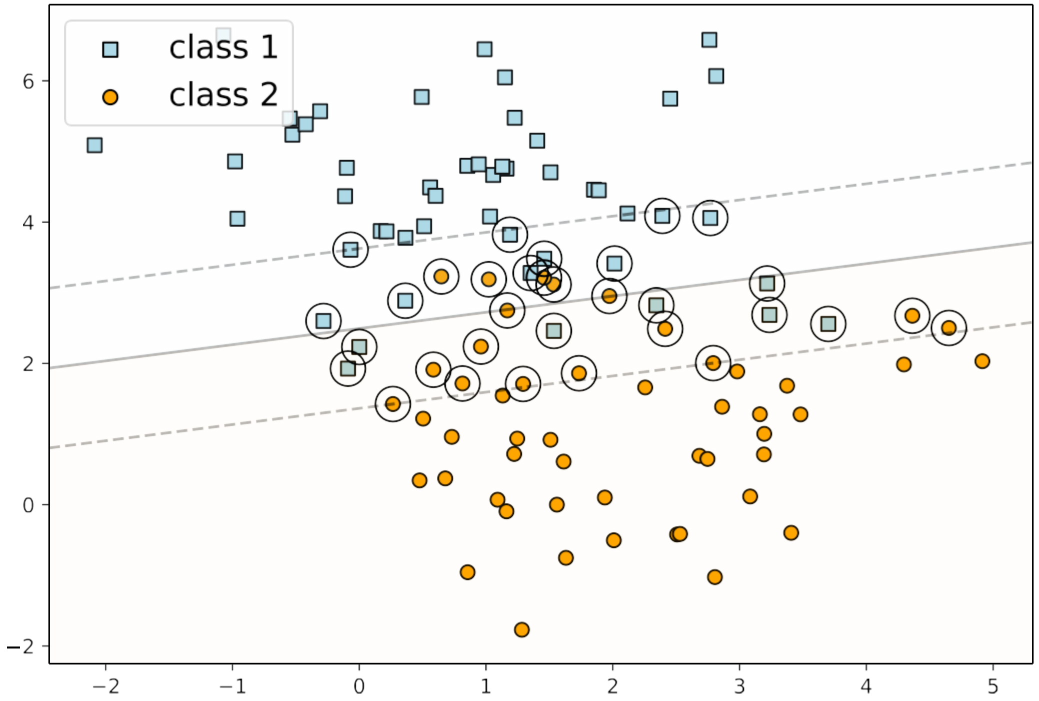
Example: Misclassification Penalty C

C = 10.0



Lower Tolerance to Misclassification Error
Smaller Margin

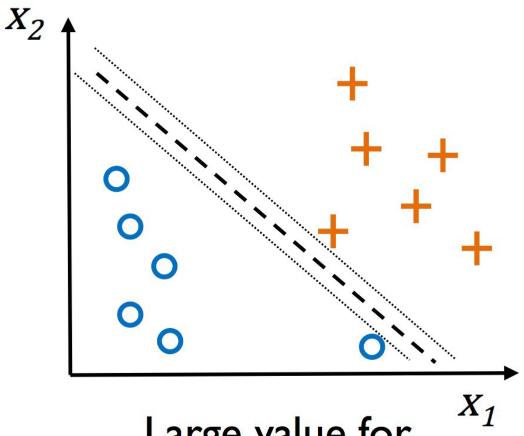
C = 0.1



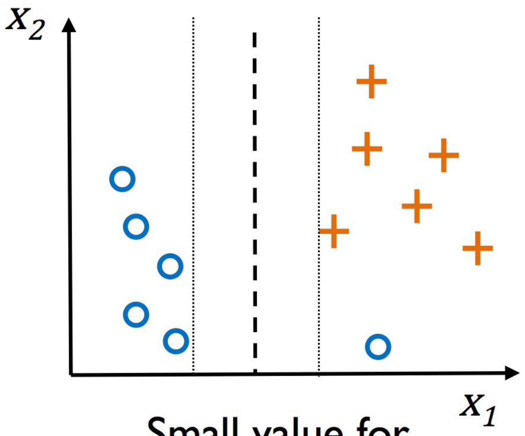
Higher Tolerance to Misclassification Error
Larger Margin

Bias and Variance

- A large value of C keeps the errors small at the cost of a reduced margin (can lead to *Overfitting*, low bias, high variance)
- A small value of C allows more misclassification while increasing the margin on the remaining examples (can lead to *Underfitting*, large bias, low variance)



Large value for parameter C

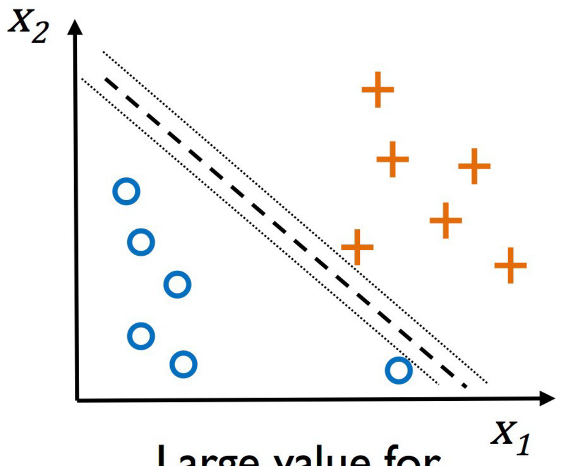


Small value for parameter C

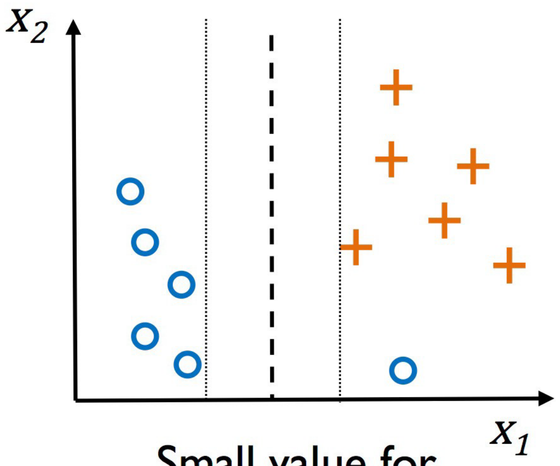
Quiz: Hard-margin v.s. Soft-margin

- Question: How do we recover hard-margin SVMs from soft-margin SVMs?

- Set misclassification penalty $C = \infty$



Large value for parameter C



Small value for parameter C

Quiz: Hard-margin v.s. Soft-margin

- Question: Given linearly separable dataset, we train a hard-margin SVM and find out that the margin is so small that the model becomes prone to overfitting. How would you resolve the overfitting issue?

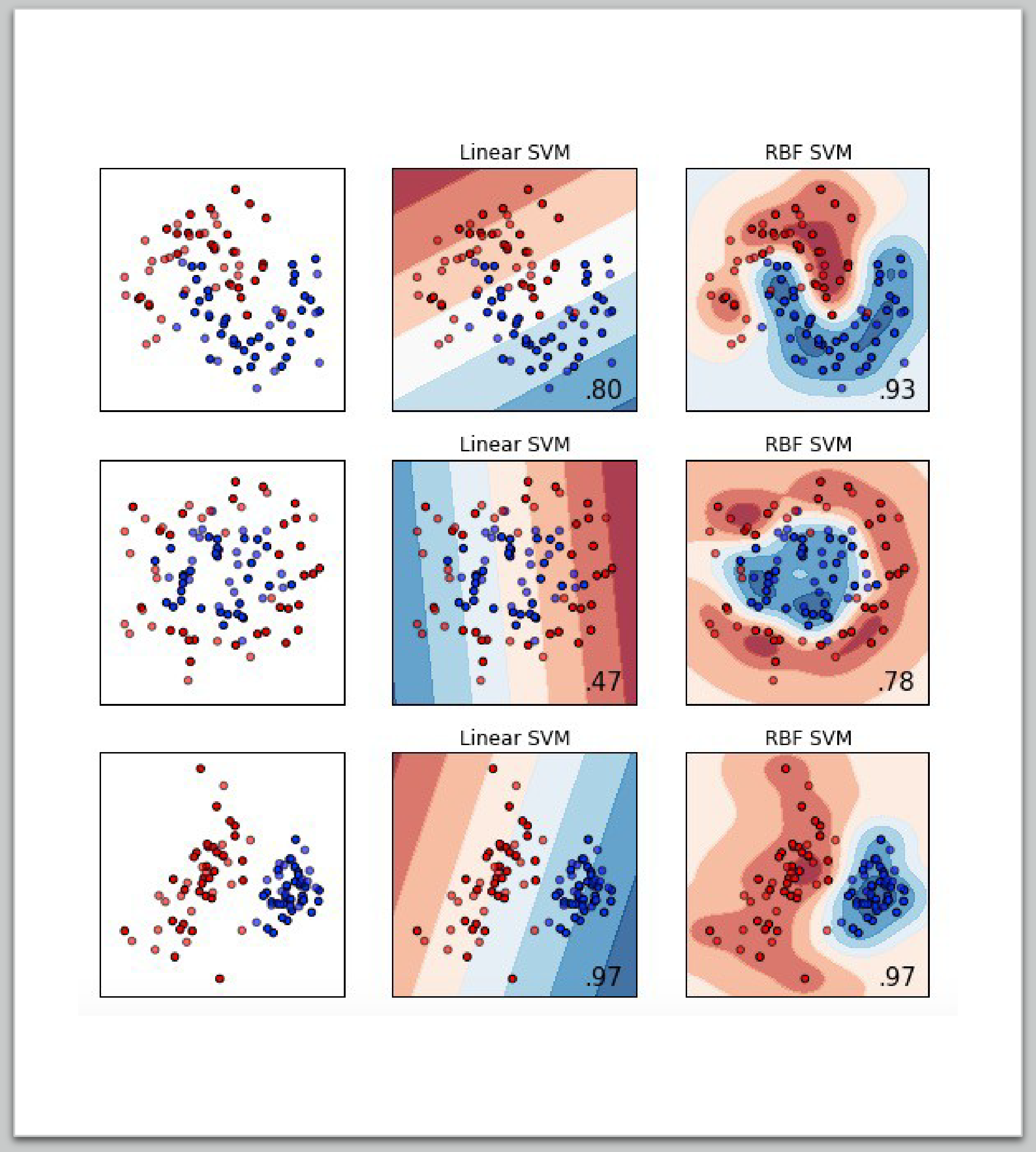
- We can opt for a larger margin by using soft-margin SVM in order to help the model generalize better.

Jupyter Notebook

Soft-margin SVMs Coding Example

SUMMARY

- Soft-margin SVMs
  - Non-Linearly Separable Data
  - Margin Maximization Hyperplane (MMH)
  - Misclassification Minimization
  - Primal Form Optimization
  - Misclassification Penalty Parameter (C)
  - Slack Variables ($\xi_{\%}$)



Resources

- SVM Website: http://www.kernel-machines.org/

- Representative Implementation
  - **LIBSVM**: an efficient implementation of SVM, multi-class classifications, nu-SVM, one-class SVM, including also various interfaces with java, python, etc.
  - **SVM-light**: simpler but performance is not better than LIBSVM, support only binary classification and only in C
  - **SVM-torch**: another recent implementation also written in C
  - **Scikit-Learn**: a set of supervised learning methods used for classification, regression and outliers detection. [link]

Resources (Contd.)

- Book Chapters: Christopher Bishop, "Pattern Recognition and Machine Learning" (PDF)
  - Sec 7.1.1-7.1.3
  - Sec 4.1.1, 4.1.2
  - Sec 6.1, 6.2
  - Appendix E

- Literatures
  - C.J.C. Burges, Chris J.C. Burges "A Tutorial on Support Vector Machines for Pattern Recognition." Data Mining and Knowledge Discovery, 1998 (PDF)