Text Categorization: Discriminative Classifiers

Part 2

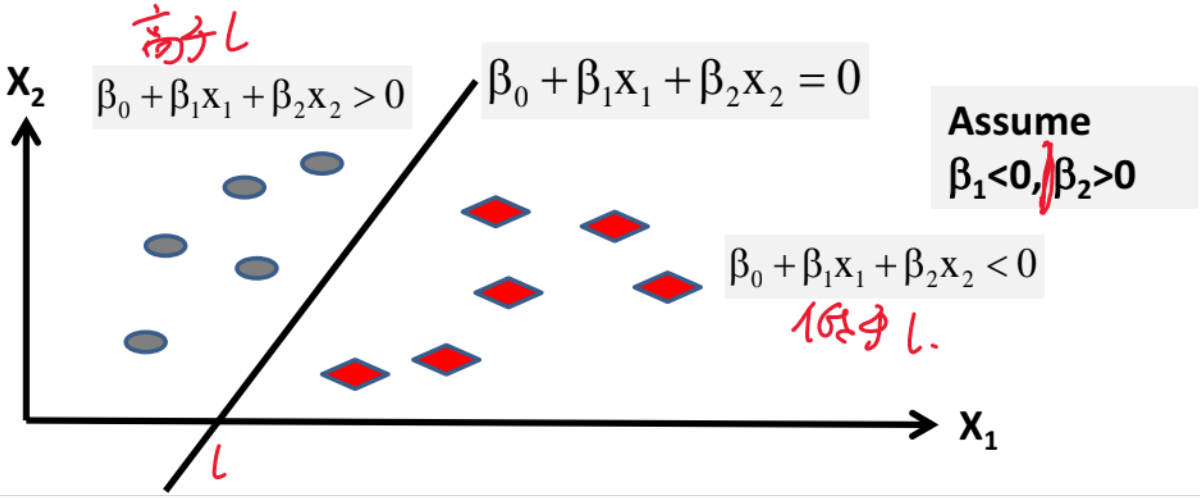
ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Discriminative Classifier 3: Support Vector Machine (SVM)

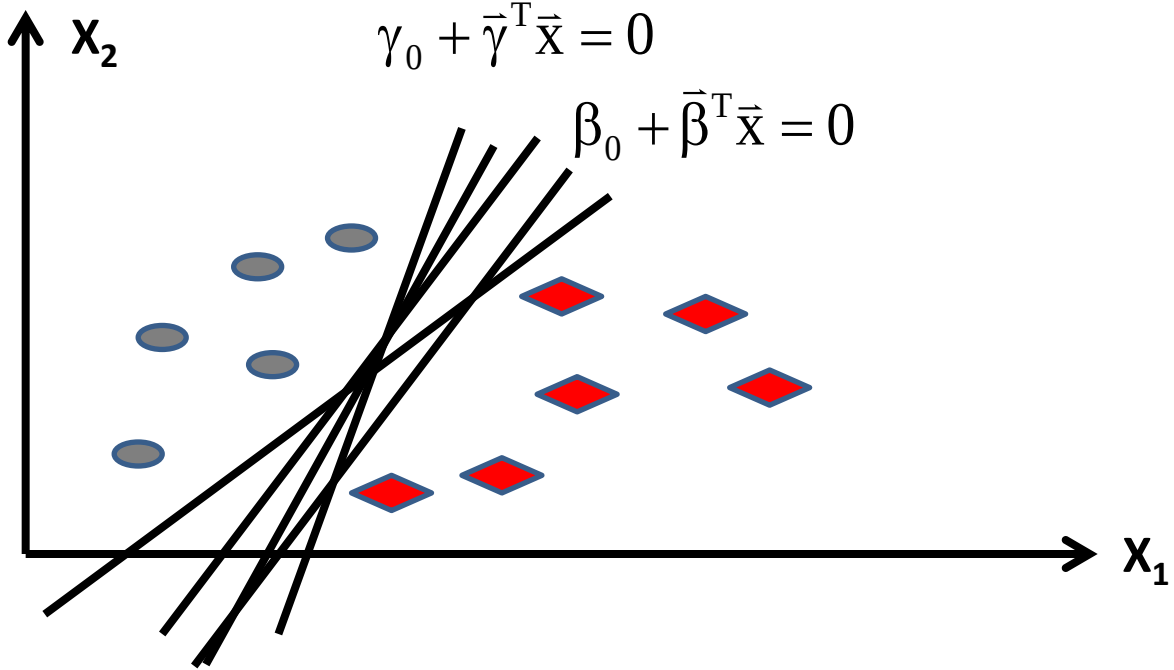
- Consider two categories: $\{\theta_1, \theta_2\}$

$$f(X) \geq 0 \Rightarrow X \text{ is in category } \theta_1$$
$$f(X) < 0 \Rightarrow X \text{ is in category } \theta_2$$

- Use a linear separator $f(X) = \beta_0 + \sum_{i=1}^{M} x_i \beta_i \quad \beta_i \in \Re$

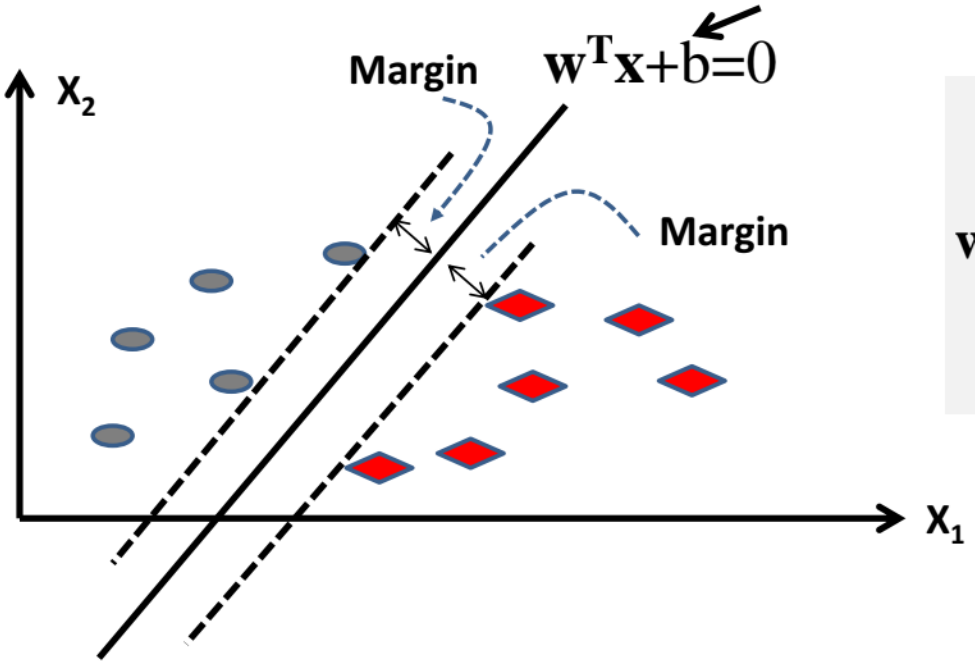


Which Linear Separator Is the Best?



Best Separator = Maximize the Margin

Notation Change: $\beta \rightarrow w$; $\beta_0 \rightarrow b$



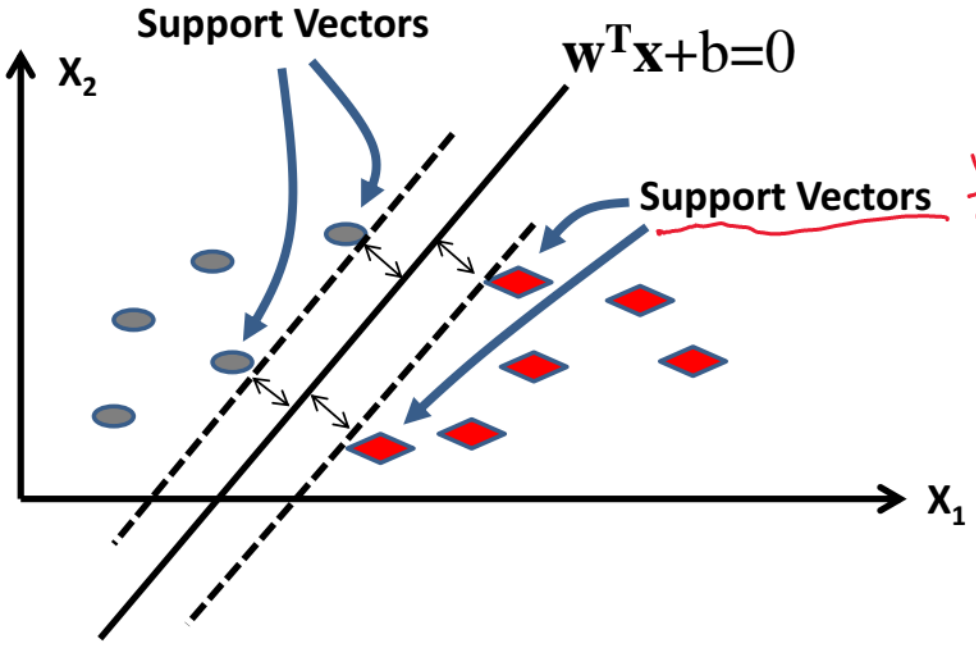
Feature Weights

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_M \end{pmatrix}$$

Feature Vector (e.g., word counts)

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_M \end{pmatrix}$$

Only the Support Vectors Matter



影响间隔的边界
最大的点

Linear SVM

Classifier: $f(x)=w^Tx+b$

Parameters: w, b

$f(X) \geq 0 \Rightarrow X$ is in category θ_1

$f(X) < 0 \Rightarrow X$ is in category θ_2

Training Data: $T=\{(x_i, y_i)\}$, $i=1, \ldots, |T|$. x_i is a feature vector; $y_i \in \{-1, 1\}$

Goal 1: Correct labeling on training data:

If $y_i=1$ ➔ $w^Tx_i+b \geq 1$

If $y_i=-1$ ➔ $w^Tx_i+b \leq -1$

结合成为一个式子

Constraint

$$\forall i,\ y_i(w^Tx_i+b) \geq 1$$

Goal 2: Maximize margin

所有权重的平方和

Large margin ⇔ Small w^Tw

Objective

$$\text{Minimize } \Phi(w)=w^Tw$$

The optimization problem is quadratic programming with linear constraints

Linear SVM with Soft Margin

Classifier: $f(x)=w^Tx+b>0?$

Parameters: w, b

Training Data: $T=\{(x_i, y_i)\}$, $i=1, \ldots, |T|$.

Added to allow training errors

Find w, b, and ξ_i to minimize $\Phi(w)=w^Tw+C\Sigma_{i\in[1,|T|]}\xi_i$

Subject to $\forall i\in[1,|T|],\ y_i(w^Tx_i+b)\geq 1-\xi_i, \quad \xi_i\geq 0$

C>0 is a parameter to control the trade-off between minimizing the errors and maximizing the margin

The optimization problem is still quadratic programming with linear constraints

Summary of Text Categorization Methods

- Many methods are available, but no clear winner
  - All require effective feature representation (need domain knowledge)
  - It is useful to compare/combine multiple methods for a particular problem
- Most techniques rely on supervised machine learning and thus can be applied to **any** text categorization problem!
  - Humans annotate training data and design features
  - Computer optimizes the combination of features
  - Good performance requires 1) effective features and 2) plenty of training data
  - Performance is generally (much) more affected by the effectiveness of features than by the choice of a specific classifier

Summary of Text Categorization Methods (cont.)

- How to design effective features? (application-specific)
  - Analyze the categorization problem and exploit domain knowledge
  - Perform error analysis to obtain insights
  - Leverage machine learning techniques (e.g., feature selection, dimension reduction, deep learning)
- How to obtain "enough" training examples?
  - Low-quality ("pseudo") training examples may be leveraged
  - Exploit unlabeled data (using semi-supervised learning techniques)
  - Domain adaptation/transfer learning ("borrow" training examples from a related domain/problem)

半监督学习

Suggested Reading

Manning, Chris D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007. (Chapters 13-15)