Text Categorization: Discriminative Classifiers

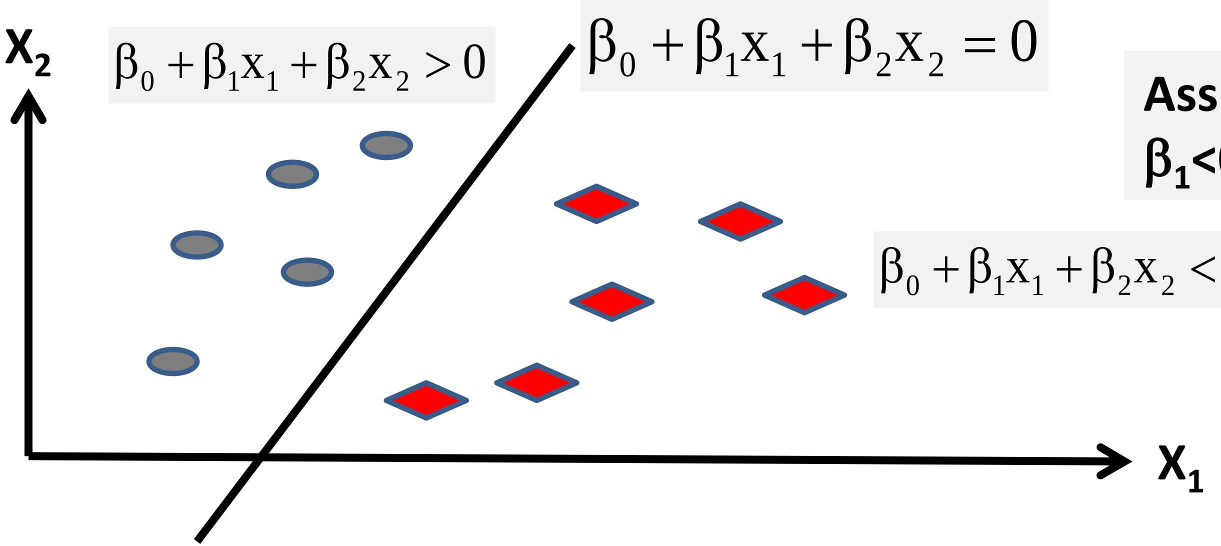
Part 2

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

Discriminative Classifier 3: Support Vector Machine (SVM)

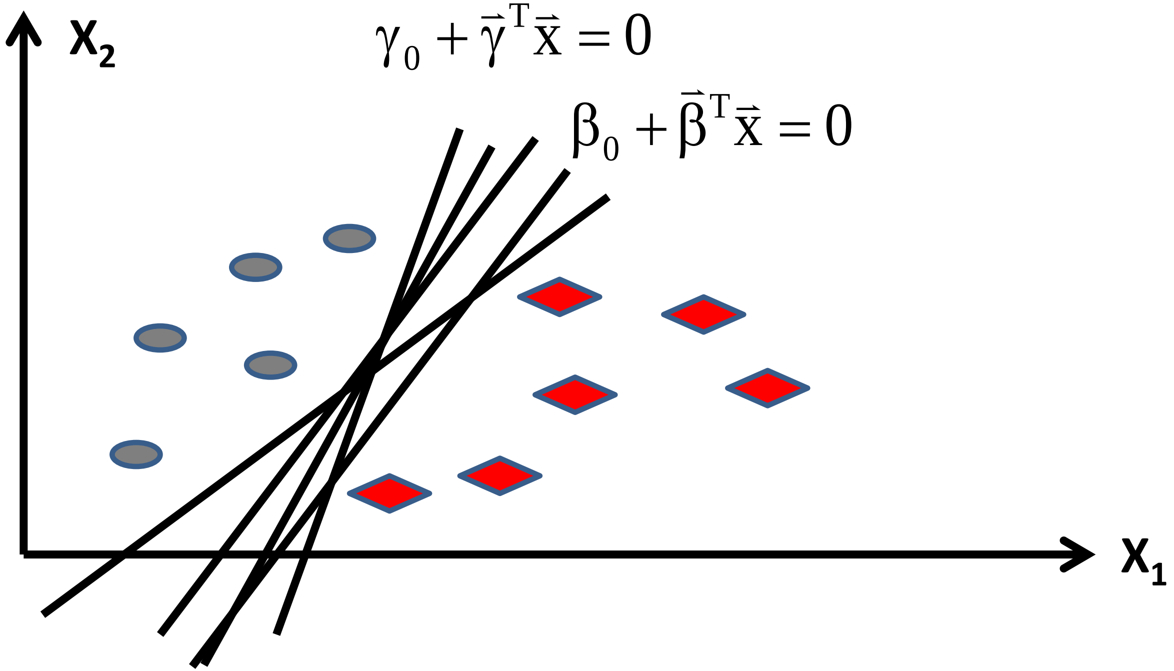
- Consider two categories: $\{\theta_1, \theta_2\}$
- Use a linear separator $f(X) = \beta_0 + \sum_{i=1}^{M} x_i \beta_i \quad \beta_i \in \Re$

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



Assume
$\beta_1 < 0$, $\beta_2 > 0$

Which Linear Separator Is the Best?

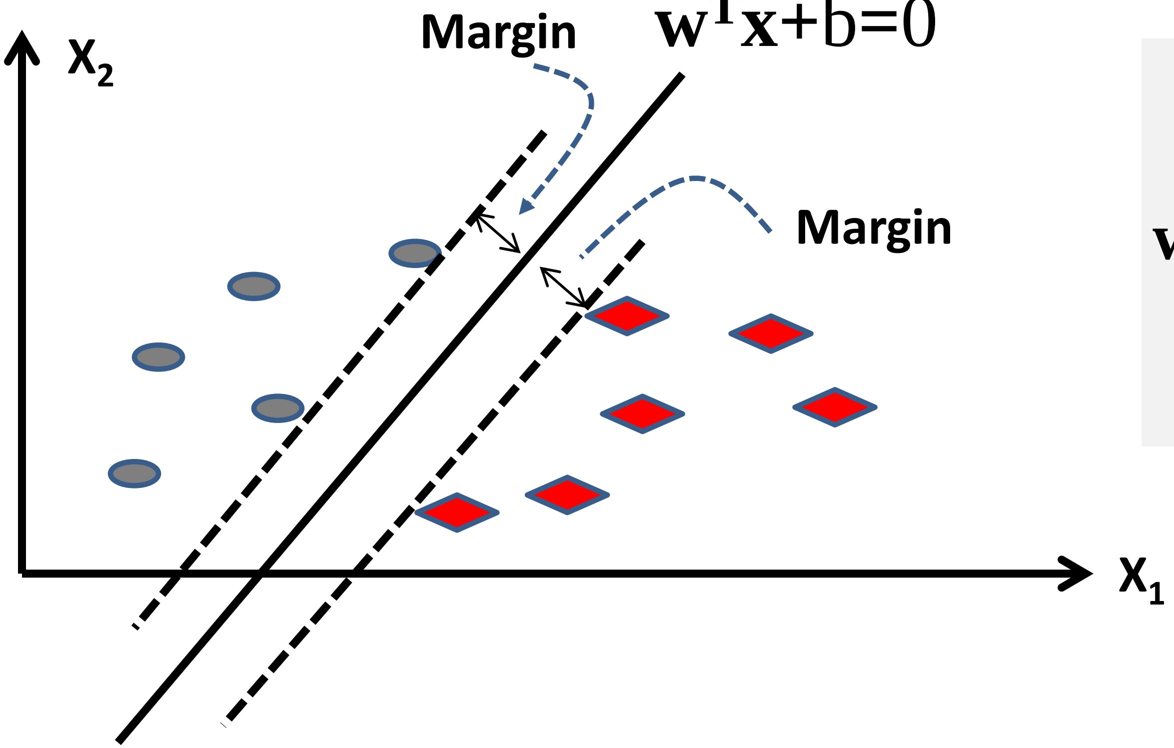


Best Separator = Maximize the Margin

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

Bias constant

$$\mathbf{w}^T\mathbf{x}+b=0$$



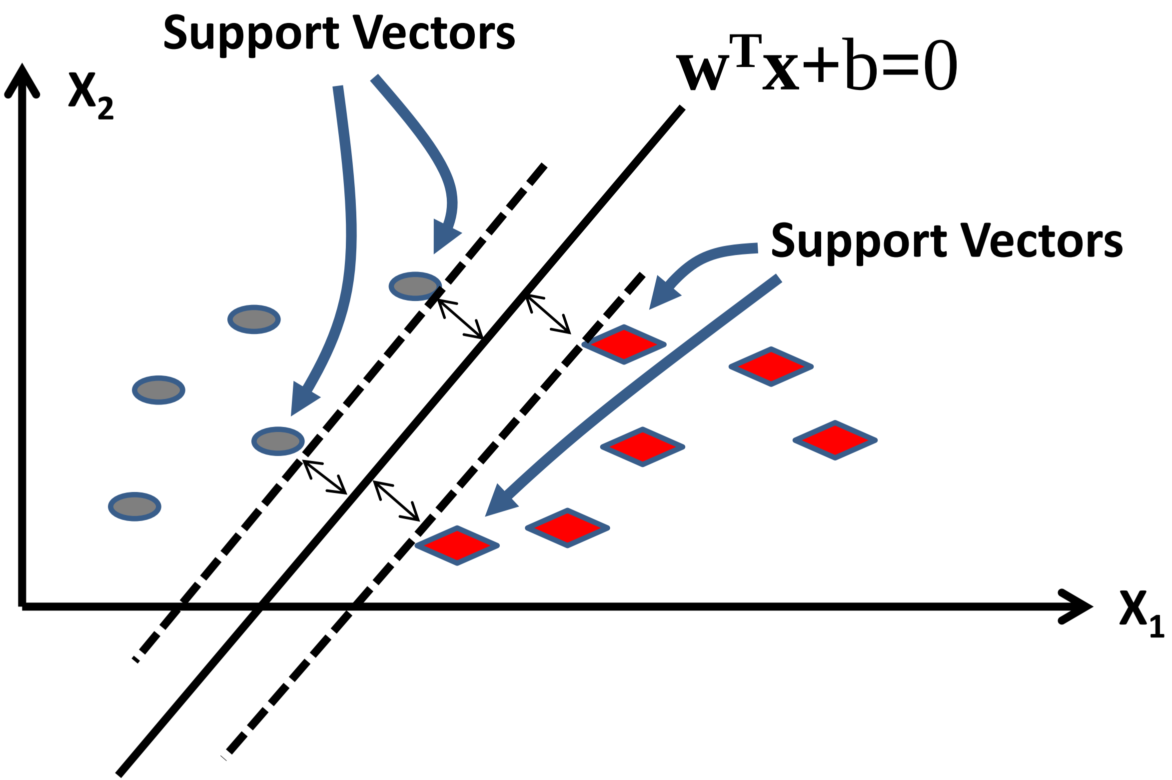
Feature Weights

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

Feature Vector (e.g., word counts)

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

Only the Support Vectors Matter



Linear SVM

Classifier: $\mathbf{f(x)=w^Tx}+b$

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

Parameters: $\mathbf{w}$, b

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

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

Goal 1: Correct labeling on training data:

If $\mathbf{y_i}=1 \rightarrow \mathbf{w^Tx_i}+b \geq 1$

If $\mathbf{y_i}=-1 \rightarrow \mathbf{w^Tx_i}+b \leq -1$

Constraint

$$\boxed{\forall \mathbf{i},\ \mathbf{y_i}(\mathbf{w^Tx_i}+b)\geq 1}$$

Objective

Goal 2: Maximize margin

Large margin $\Leftrightarrow$ **Small** $\mathbf{w^Tw}$

$$\boxed{\textbf{Minimize } \Phi(\mathbf{w})=\mathbf{w^Tw}}$$

The optimization problem is quadratic programming with linear constraints

Linear SVM with Soft Margin

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

Parameters: $\mathbf{w}$, b

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

Added to allow training errors

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

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

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)

Suggested Reading

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