Machine Learning Workshop 2006
Nanjing
Nov. 4-5, 2006

Beyond Binary Classification

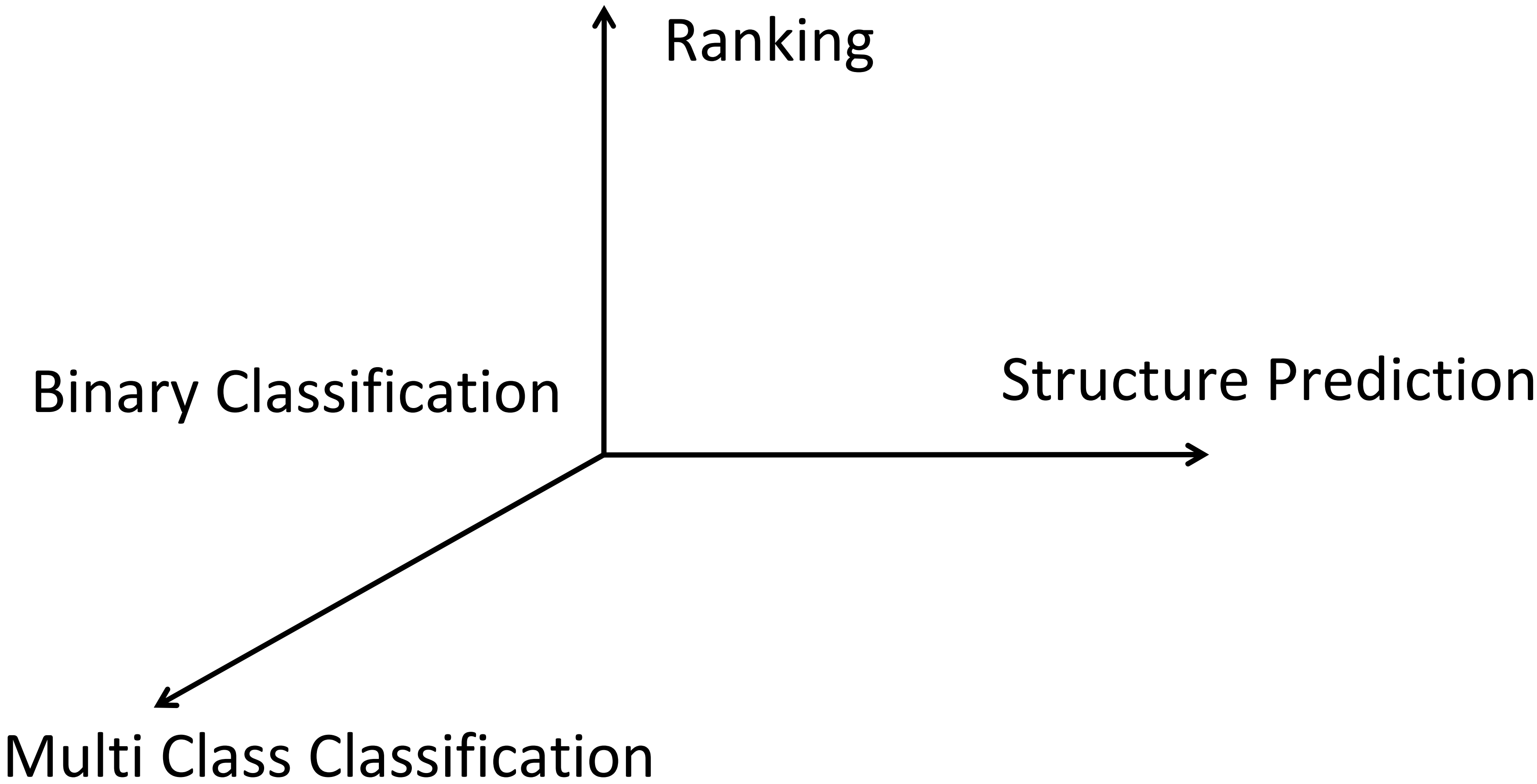
Hang Li
Microsoft Research Asia
hangli@microsoft.com

Introduction

Learning Methods for Classification

Method	Class
K Near Neighbor	Multi-Class
Naïve Bayes	Multi-Class
Decision Tree / Decision List	Multi-Class
Maximum Entropy / Logistic Regression	Multi-Class
Support Vector Machines	Binary Class
Ada Boost	Binary Class

From Binary Classification to More Complicated Predictions



This Talk: Survey on Learning Methods for Multi-Class Classification, Structure Prediction, and Ranking, Using SVM Approach

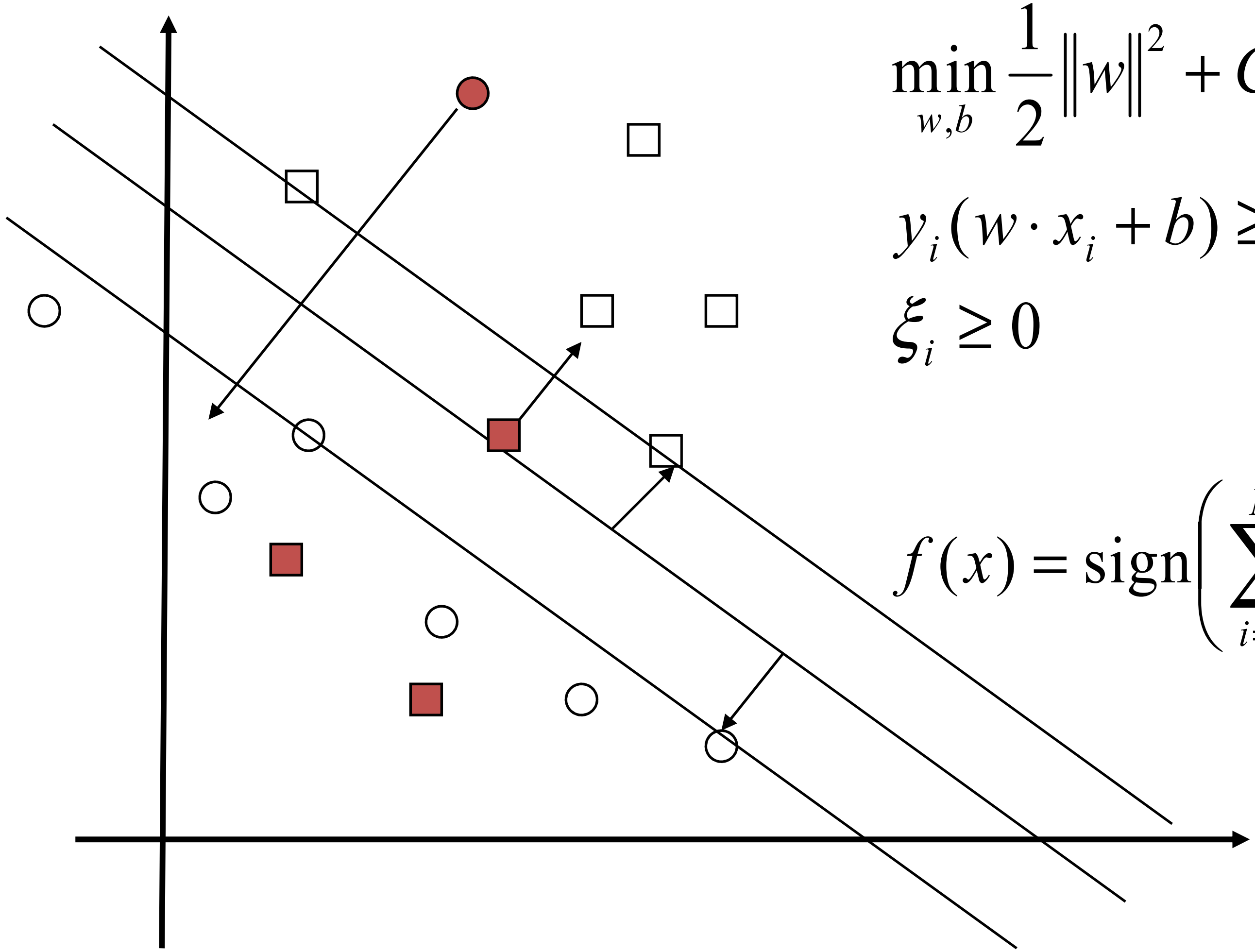
From Studies with Yunhua Hu, Yunbo Cao, Dijun Luo, Jun Xu, Tie-Yan Liu, and others

Talk Outline

- Introduction
- Multi-Class Classification
- Learning for structure Prediction
- Learning to Rank
- Summary

Multi-Class Classification

Linear SVM



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

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad i = 1, \cdots, N$$

$$\xi_i \geq 0$$

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i (x_i \cdot x) + b^*\right)$$

Multi Class Classification Problem

- Input space: $X \subseteq R^d$
- Output space: $Y = \{1, 2, \cdots, K\}$
- Prediction function: $f : X \to Y$
- Learning
  - Input: $S = \{(x_i, y_i), x_i \in X, y_i \in Y, i = 1, \cdots N\}$
  - Output: $f(x; \hat{w})$

Methods for Multi Class Classification

- Multi-Class
  - Crammer, K. & Singer. Y. (2001) On the algorithmic implementation of multiclass kernel-based machines. *Journal of Machine Learning Research*, 2(Dec):265--292.
- Error Correcting Output Code (ECOC)
  - Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263–286.
- Hierarchy

Multi-Class SVM
Cramer & Singer (2001)

- Model:

$$\forall k, \left\langle w_k, x \right\rangle$$

- Prediction:

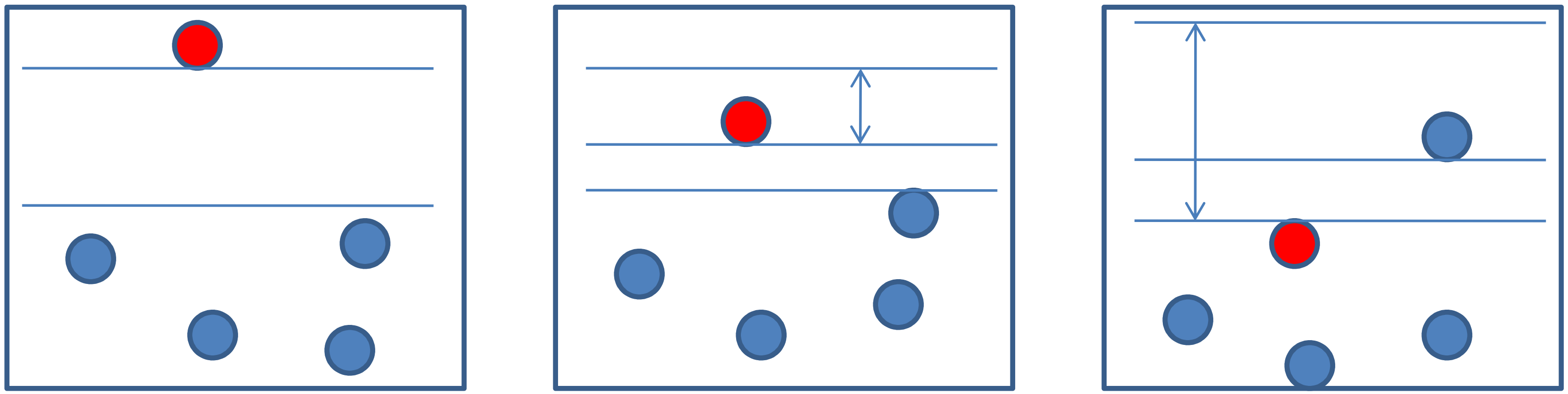
$$\arg\max_k \left\langle w_k, x \right\rangle$$

Multi-Class SVM

$$\min_{w,\xi} \frac{1}{2}\sum_{k=1}^{K} \| w_k \|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\forall i, \forall y \in Y \setminus y_i,\ \left\langle w_{y_i}, x_i \right\rangle - \left\langle w_y, x_i \right\rangle \geq 1 - \xi_i$$

$$\forall i, \xi_i \geq 0$$



Error Correcting Output Code
Dietterich & Bakiri (1995)

- Encoding

  $\mathbf{M}$ is matrix of size $K \times L$ over $\{-1, 0, +1\}$

- Base classifier construction

  L binary classifiers $h_1(x), \cdots, h_L(x)$

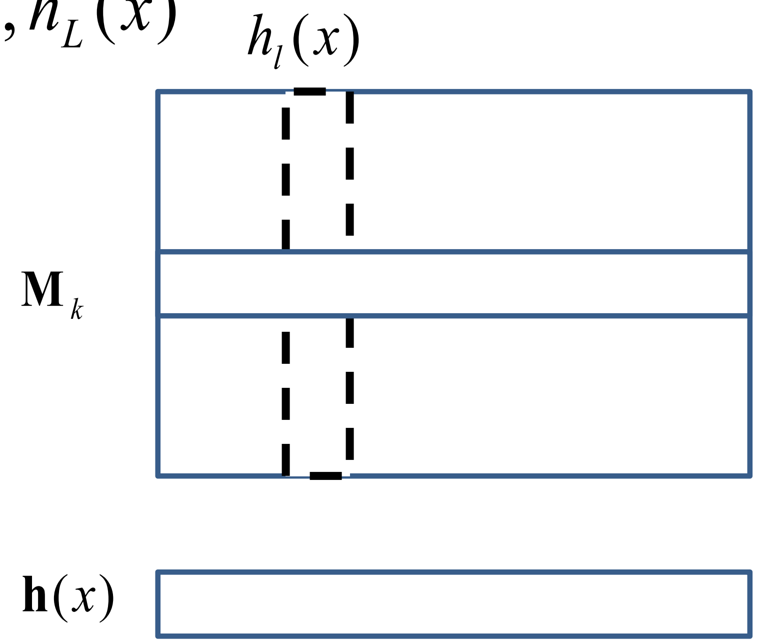
  $\mathbf{M}_{k,l} = +1$

  $\mathbf{M}_{k,l} = -1$

- Decoding

  $\mathbf{h}(x) = [h_1(x), \cdots, h_L(x),]$

  $\arg\min_k D(\mathbf{M}_k, \mathbf{h}(x))$



Encoding

- Coding Matrix
  - One vs rest
  - One vs one
  - Random code

$$\begin{pmatrix} +1 & -1 & -1 \\ -1 & +1 & -1 \\ -1 & -1 & +1 \end{pmatrix}$$

$$\begin{pmatrix} +1 & +1 & 0 \\ -1 & 0 & +1 \\ 0 & -1 & -1 \end{pmatrix}$$

Combining ECOC and Multi Class SVM in Single Framework

- Coding matrix (one vs rest) and decoding metric (dot product) are given
- Binary classifier: SVM
- Simultaneously training binary SVM classifiers is equivalent to Multi Class SVM

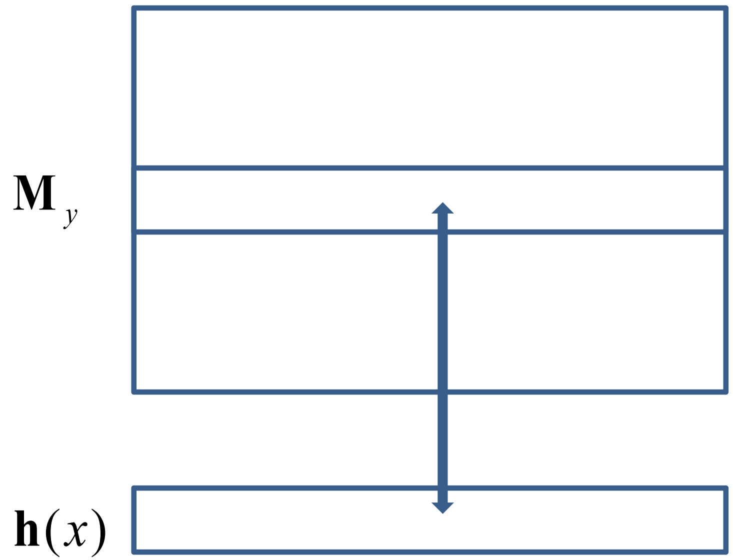
Using Framework of ECOC

- Classification

$$f(x) = \arg\max_{y} F(x, y)$$

$$F(x, y) = \left\langle \mathbf{M}_y, \vec{\mathbf{h}}(x) \right\rangle$$

$$\mathbf{h}(x) = \left[h_1(x), \cdots, h_L(x) \right]$$



- Loss Function

$$\left(1 - F(x, y; w) - \max_{y \neq y'} F(y, y'; w) \right)_+$$

Learning

- Training data

$$S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$$

- Regularized total loss

$$L(S;w) = \sum_{i=1}^{N} \left[1 - \left(\sum_{l=1}^{L} \mathbf{M}_{y_i,l} \langle w_l, x_i \rangle - \max_{y \neq y_i} \sum_{l=1}^{L} \mathbf{M}_{y,l} \langle w_l, x_i \rangle \right) \right]_+ + \lambda \sum_{l=1}^{L} \| w_l \|^2$$

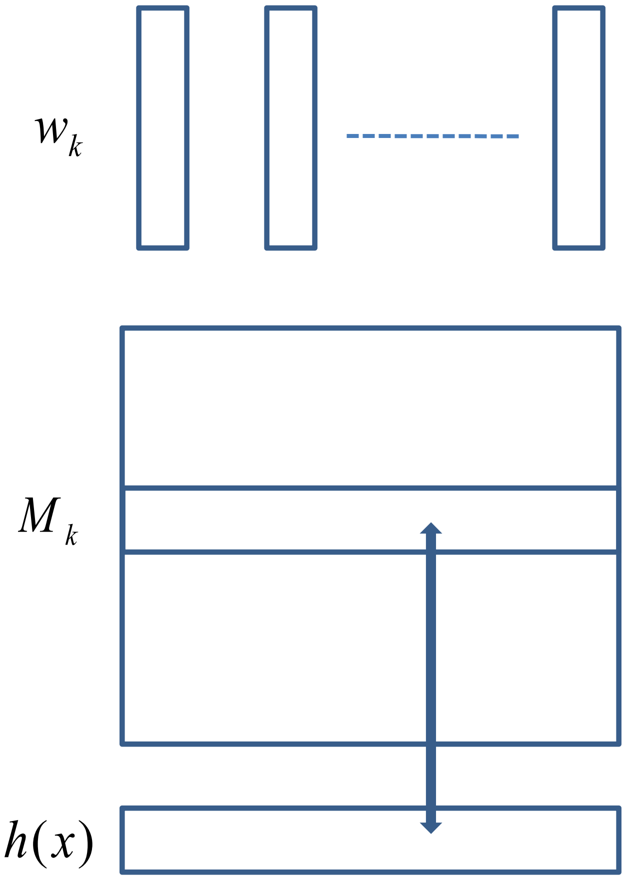
- SVM

$$\min_{w,\xi} \frac{1}{2} \sum_{l=1}^{L} \| w_l \|^2 + C \sum_{i=1}^{N} \xi_i$$

$$\forall i, \forall y \in Y \setminus y_i, \quad \sum_{l=1}^{L} \mathbf{M}_{y_i,l} \langle w_l, x_i \rangle - \sum_{l=1}^{L} \mathbf{M}_{y,l} \langle w_l, x_i \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

ECOC: One vs. Rest



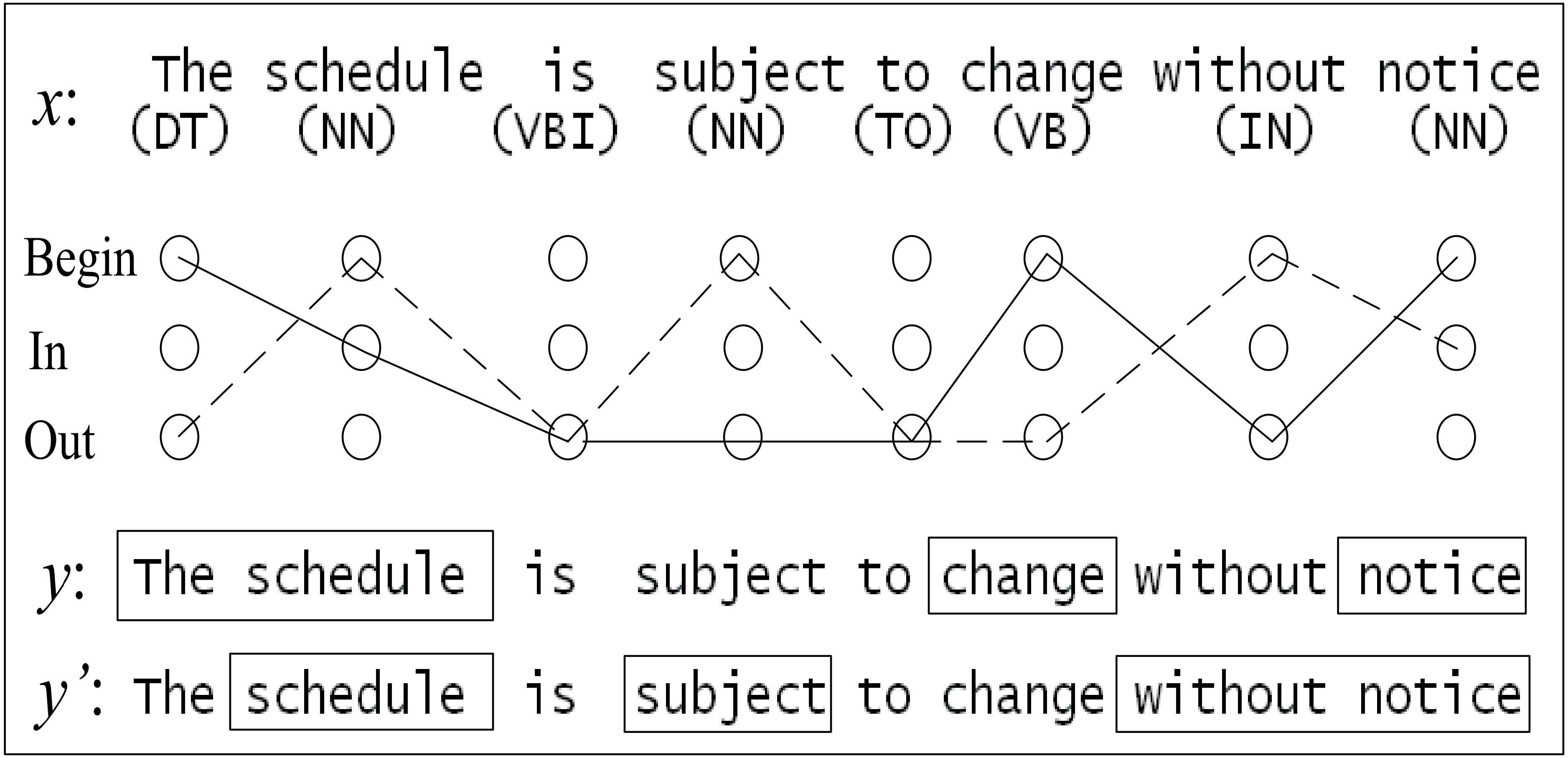
$$\min_{w,\xi} \frac{1}{2}\sum_{k=1}^{K} \| w_k \|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\forall i, \forall y \in Y \setminus y_i,\ \left\langle w_{y_i} - w_y, x_i \right\rangle \geq \frac{1}{2}(1 - \xi_i)$$

$$\xi_i \geq 0$$

Structure Prediction

Example of Structure Prediction
Term Extraction



Methods for Structure Prediction

- SVM
  - Taskar, B., Chatalbashev, V., Koller, D., & Guestrin, C. (2005). Learning structured prediction models: A large margin approach. *Twenty-Second International Conference on Machine Learning* (pp. 896-903). New York: ACM Press.
  - Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). *Large margin methods for structured and interdependent output variables.* The Journal of Machine Learning Research, 6, 1453-1484.
- Hidden Markov Model
- Conditional Random Fields

Discriminative Approach to Structure Prediction

- Input space: *X*, output space: *Y*
- Prediction function: $f : X \to Y$
- Discriminate function: $F : X \times Y \to R$
- Predict: $f(x;w) = \arg\max_{y \in Y} F(x, y; w)$
- Linear function: $F(x, y; w) = \left\langle w, \Psi(x, y) \right\rangle$

  $\Psi(x, y) \in R^d$

Discriminative Approach

- Prediction:
  - Input: (x, Y)
  - Output: $f(x; \hat{w}) = \arg\max_{y \in Y} F(x, y; \hat{w})$

- Learning
  - Input: $S = \{(x_i, y_i, Y_i), i = 1, \cdots N\}$
  - Output: $f(x; \hat{w})$

Notes

- Space Y is large
- F is function of both x and y
- Outputs y's are interdependent
- Number of outputs y's is exponential
- Dynamic programming algorithm must exist for computing $F(x,y;w)$

SVM Model
Tsochantaridis et al (2004)

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

$$\forall i, \forall y \in Y_i \setminus y_i : \langle w, \Psi(x_i, y_i) \rangle - \langle w, \Psi(x_i, y) \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

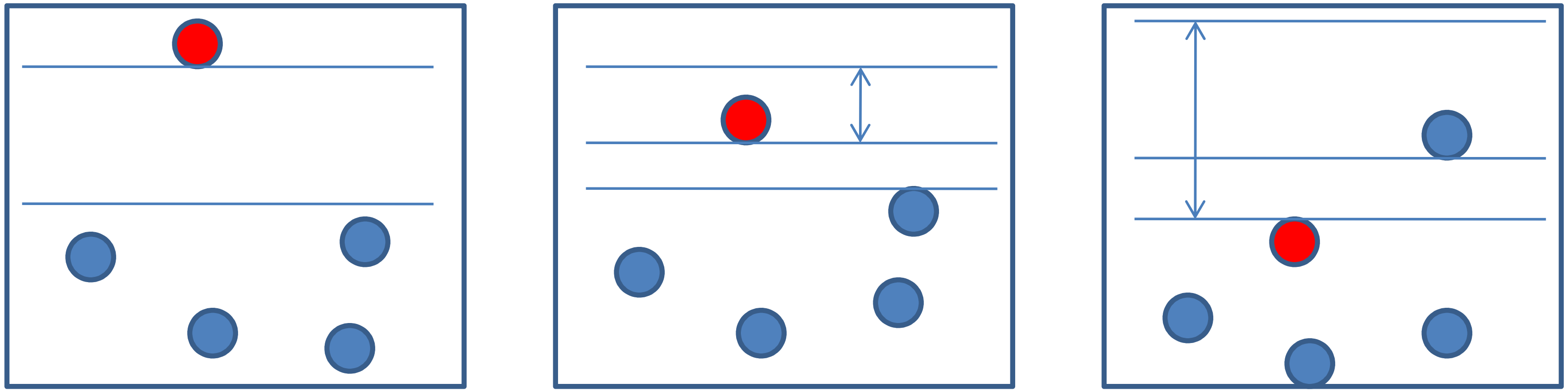
Loss Function

- General loss function

$$L\left(F(x, y; w) - \max_{y' \in Y \setminus y} F(x, y'; w) \right)$$

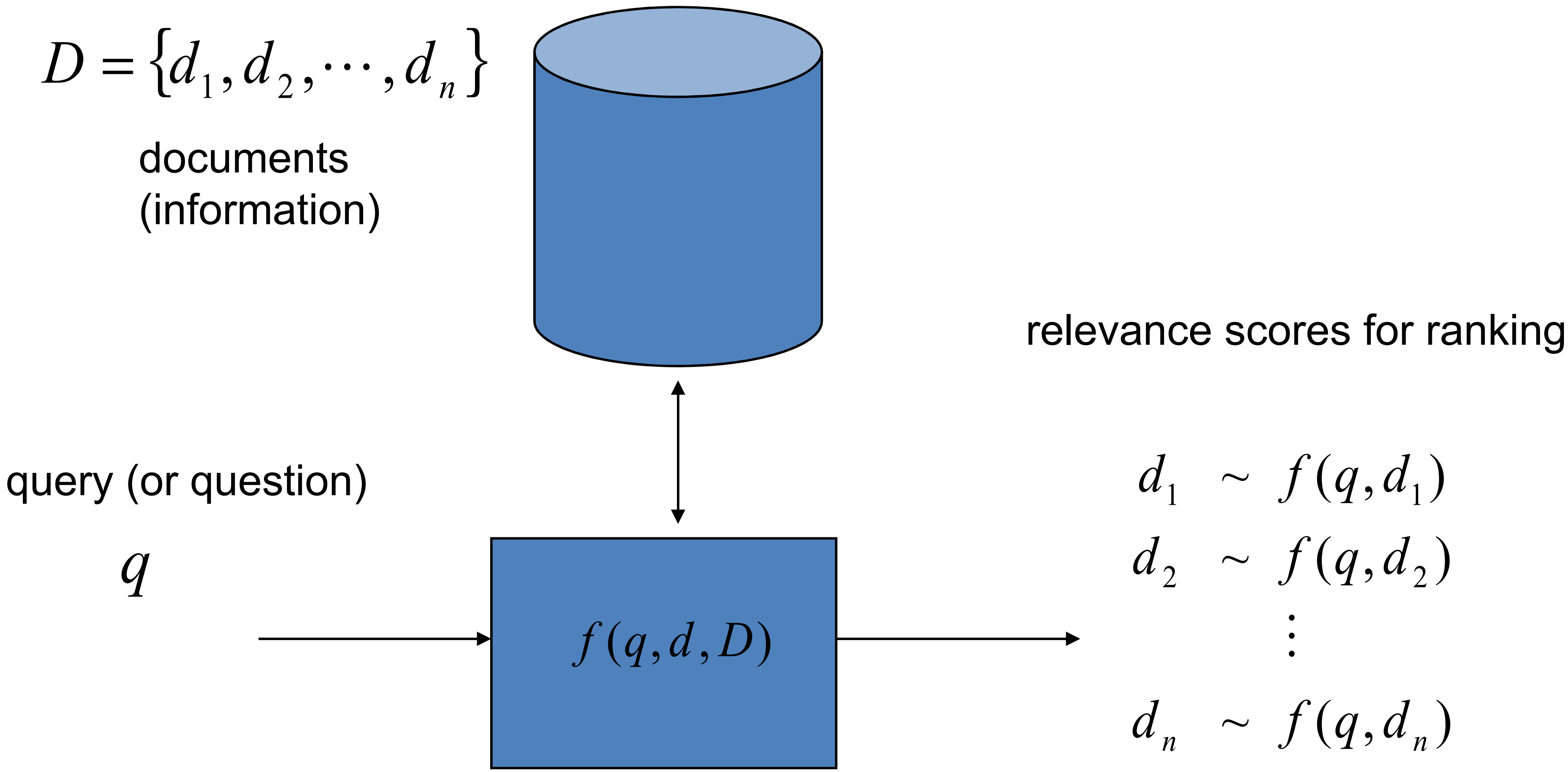
- Hinge loss function

$$L\left(1 - \left[F(x, y; w) - \max_{y' \in Y \setminus y} F(x, y'; w) \right]_{+} \right)$$



Ranking

Example of Ranking: Information Retrieval



Methods for Ranking

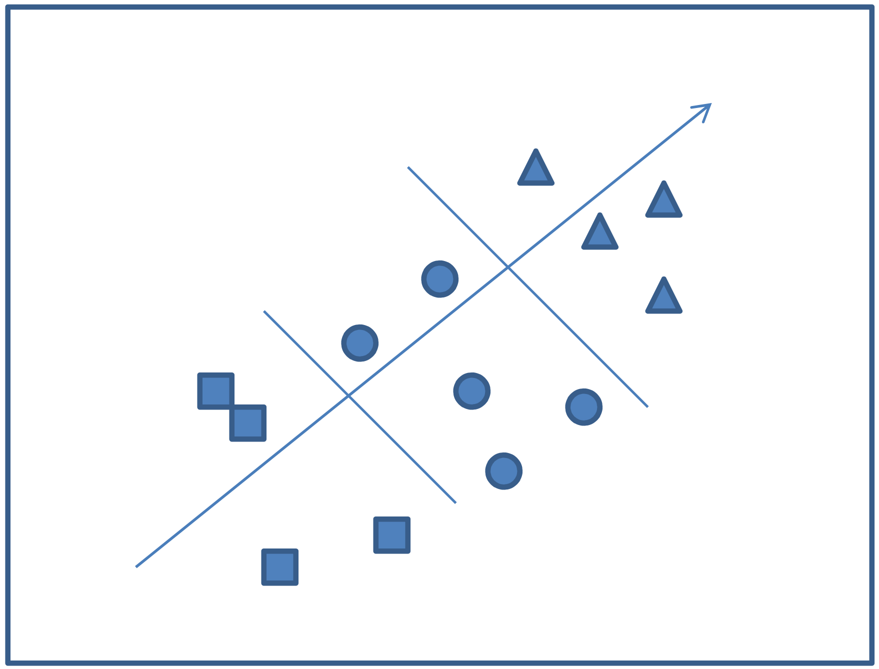
- Point-wise Ranking Methods
  - A. Sahshu, A. Levin, Ranking with Large Margin Principle: Two Approaches, NIPS’03
- Pair-wise Ranking Methods
  - R. Herbrich, T. Graepel, and K. Obermayer. Large Margin Rank Boundaries for Ordinal Regression. Advances in Large Margin Classifiers, pages 115-132, 2000.
  - Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, Hsiao-Wuen Hon, Adapting Ranking SVM to Document Retrieval, Proc. of SIGIR’06.

Point-wise Ranking SVM
Shashua and Levin (2003)

- Input space: *X*, output space: *Y with order*
- Ranking function $f : X \to Y$
- Ranking: by $f(x; w)$

$$f(x) = \arg\min_{k \in \{1, \cdots, K\}} \left(w \cdot x + b_k < 0 \right)$$

$$b_K = \infty$$



Point-wise Ranking SVM

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_i \sum_j (\xi_i^j + \xi_i^{*j+1})$$

$$\left\langle w, x_i^j \right\rangle - b_j \leq -1 + \xi_i^j$$

$$\left\langle w, x_i^{j+1} \right\rangle - b_j \geq 1 - \xi_i^{*j+1}$$

$$\xi_i^j \geq 0, \xi_i^{*j} \geq 0$$

$$j = 1, \cdots, k-1, \quad i = 1, \cdots, i_j$$

(Pair-wise) Ranking SVM Herbrich et al (2000)

- Input space: *X*
- Ranking function $f : X \to R$
- Ranking: $x_i \succ x_j \Leftrightarrow f(x_i; w) > f(x_j; w)$
- Linear ranking function: $f(x; w) = \left\langle w, x \right\rangle$

$$\left\langle w, x^{(1)} - x^{(2)} \right\rangle > 0 \quad \Leftrightarrow \quad f(x^{(1)}; w) > f(x^{(2)}; w)$$

- Transforming to binary classification:

$$(\vec{x}^{(1)} - \vec{x}^{(2)}, z), \;\; z = \begin{cases} +1 & y^{(1)} \succ y^{(2)} \\ -1 & y^{(2)} \succ y^{(1)} \end{cases}$$

Pair-wise Ranking SVM

$$\min_{w,\xi} \frac{1}{2} \| w \|^2 + C\xi_i$$

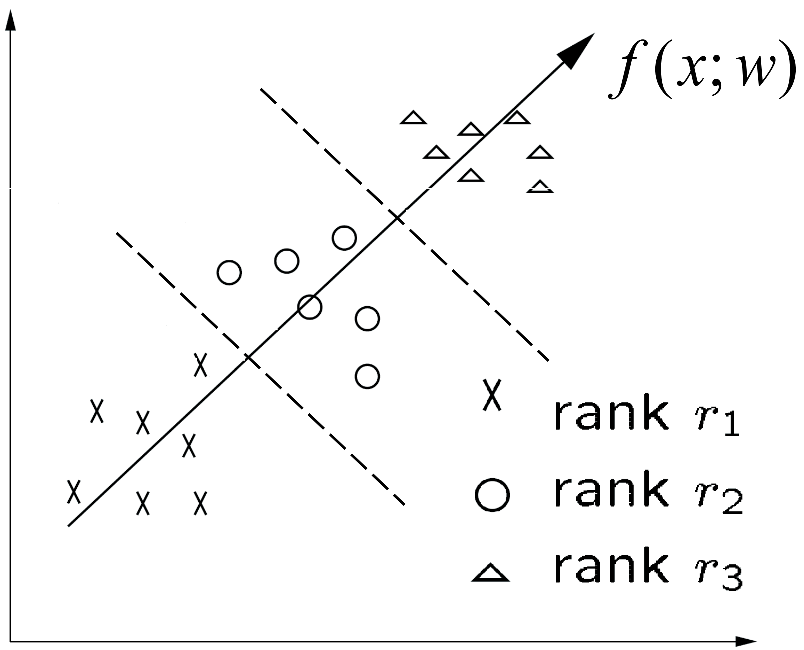
$$z_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$\min_{w} \sum_{i=1}^{l} \left[1 - z_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle \right]_+ + \lambda \| w \|^2$$

Pair-wise Ranking SVM

- Learning
  - Input: $S = \{(x_i^{(1)} > x_i^{(2)})\}_{i=1}^{m}$
  - Output: $f(x; \hat{w})$



Direct Application of Ranking SVM to Information Retrieval

- Query document pair → feature vector
- Combining instance pairs from all queries

Applying Ranking SVM to Document Retrieval

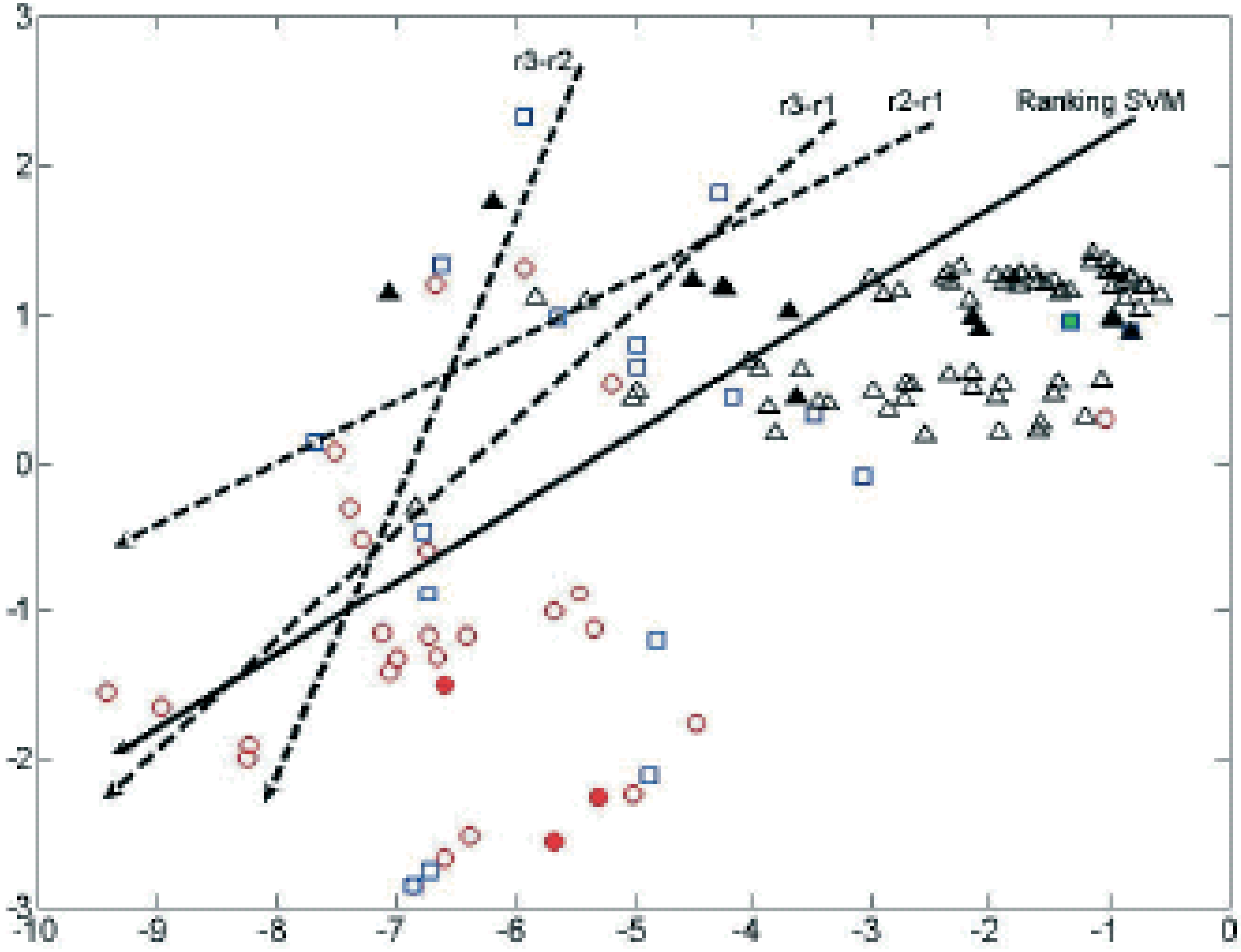
- Cost sensitiveness: negative effects of making errors on top

  d: *definitely relevant*, p: *partially relevant*, n: *not relevant*
  ranking 1: p d p n n n n
  ranking 2: d p n p n n n

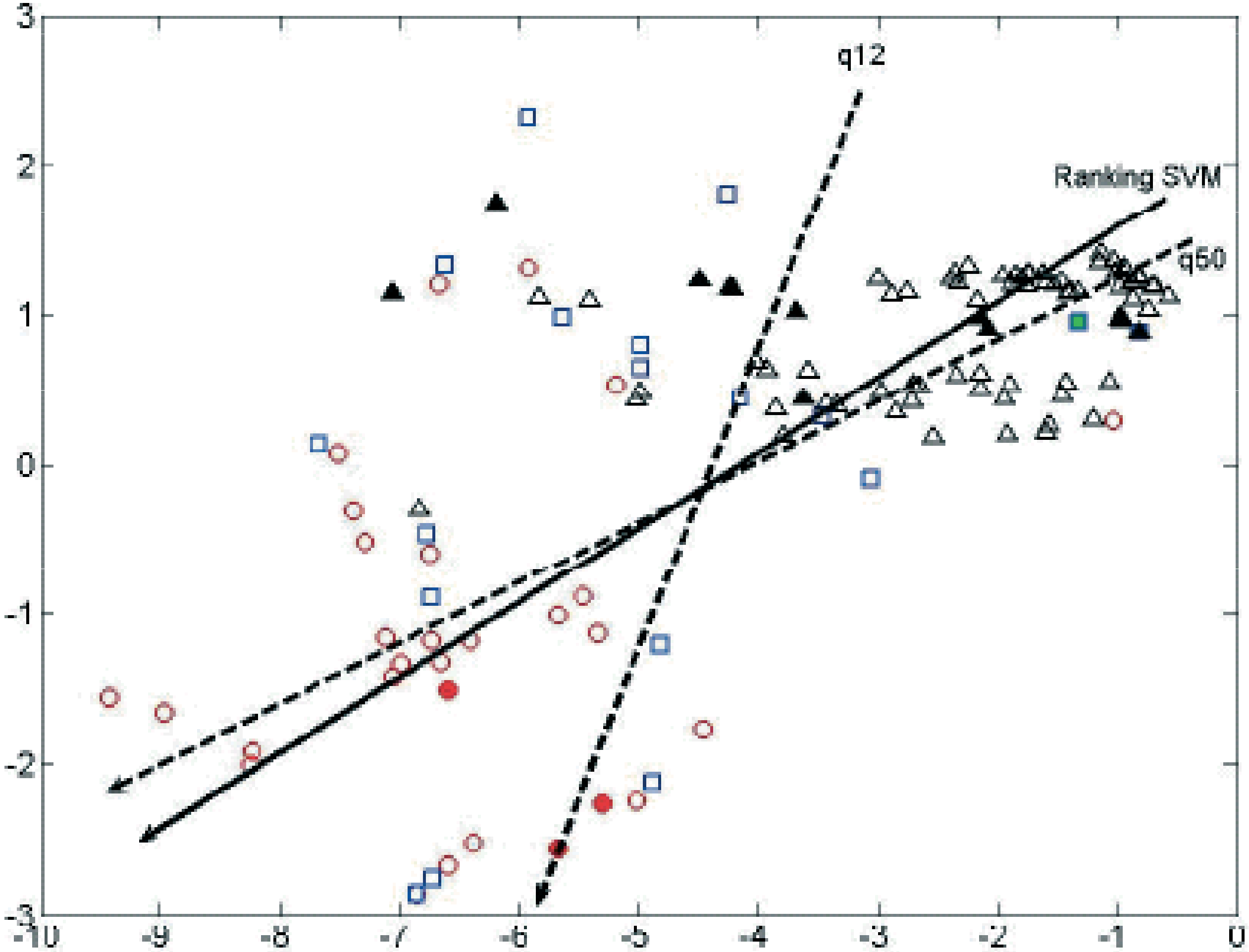
- Query normalization: number of instance pairs varies according to query

  q1: d p p n n n n
  q2: d d p p p n n n n n
  q1 pairs: 2*(d, p) + 4*(d, n) + 8*(p, n) = 14
  q2 pairs: 6*(d, p) + 10*(d, n) + 15*(p, n) = 31

Rank Pair Distinction



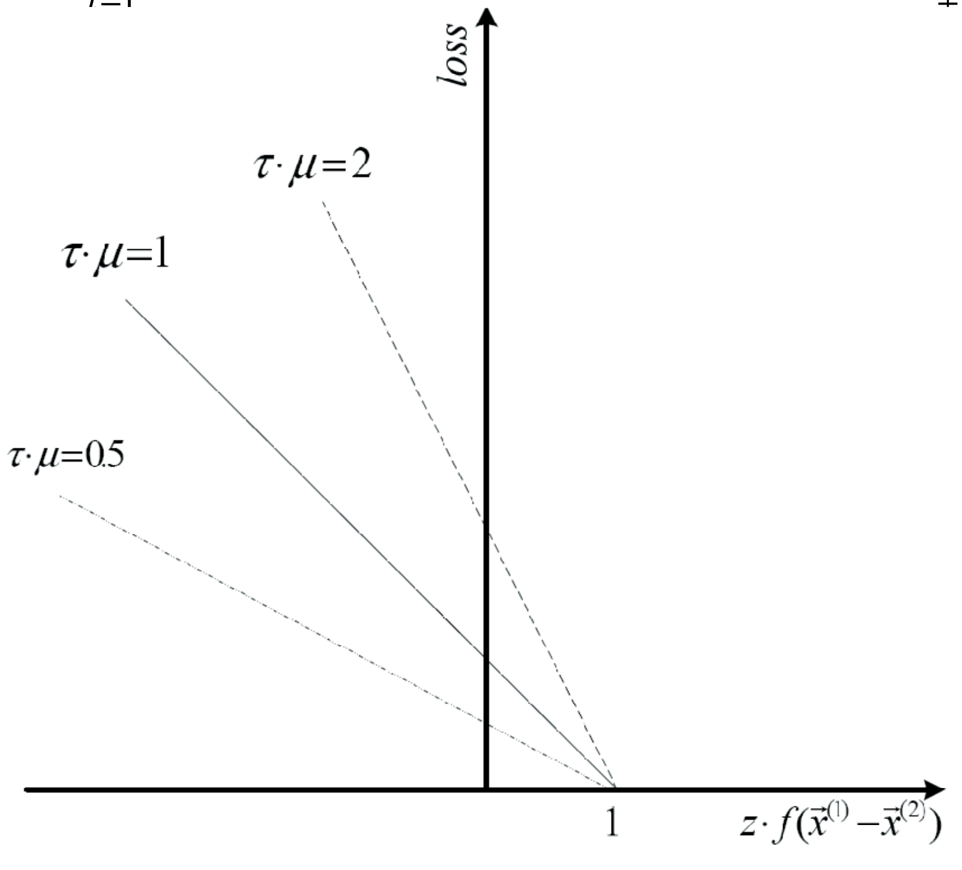
Query Normalization



Ranking SVM for IR
Cao et al (2006)

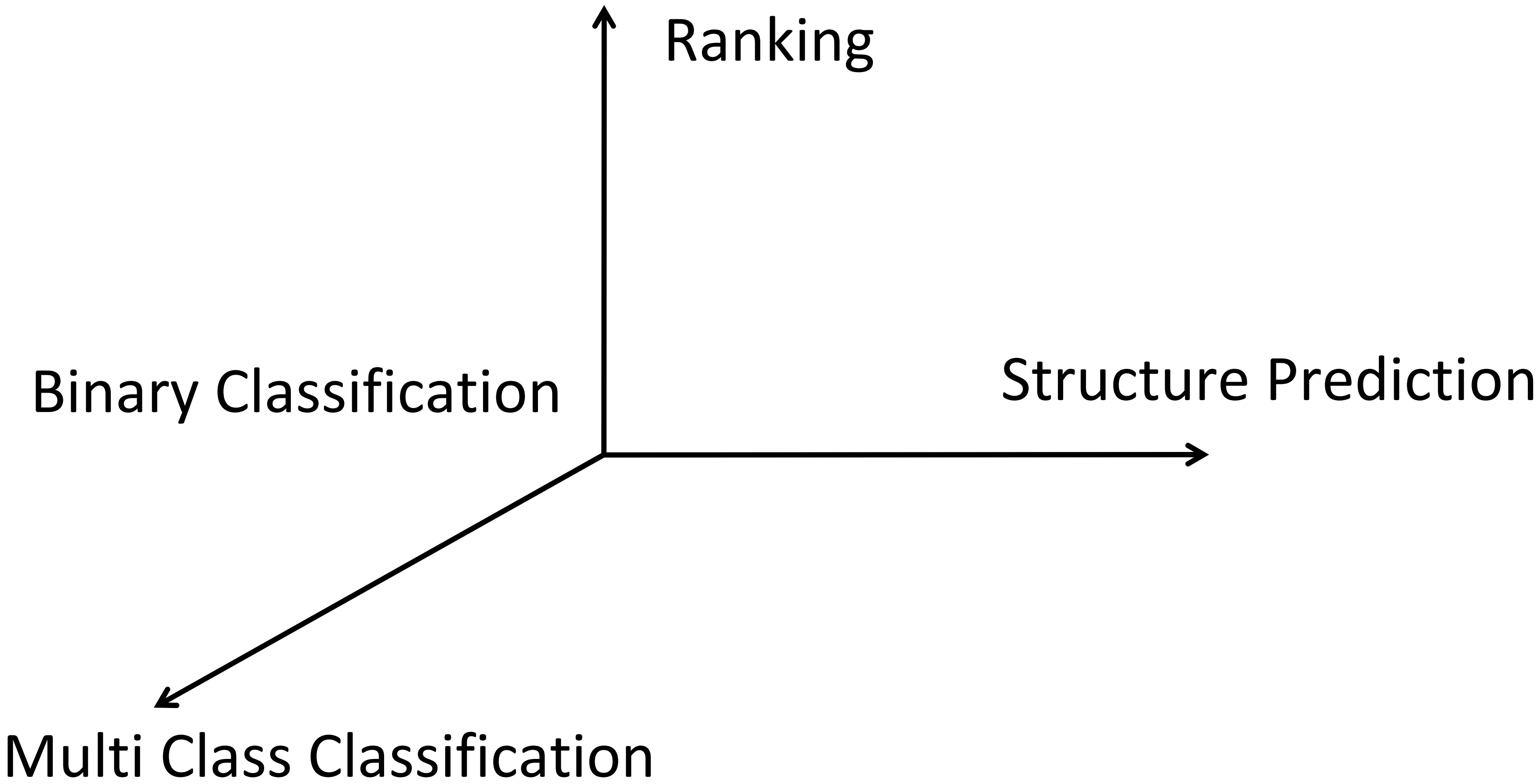
Loss Function

$$\min_{\vec{w}} L(\vec{w}) = \sum_{i=1}^{l} \tau_{k(i)} \mu_{q(i)} \left[1 - z_i \left\langle \vec{w}, \vec{x}_i^{(1)} - \vec{x}_i^{(2)} \right\rangle \right]_+ + \lambda \left\| \vec{w} \right\|^2$$



Summary

Summary


Ranking
Binary Classification
Structure Prediction
Multi Class Classification

Summary

- Multi-Class Classification

  Combining ECOC and Multi-Class SVM in Single Framework

- Structure Prediction

  Transforming into Binary Classification

- Ranking

  Transforming into Binary Classification

Thank You