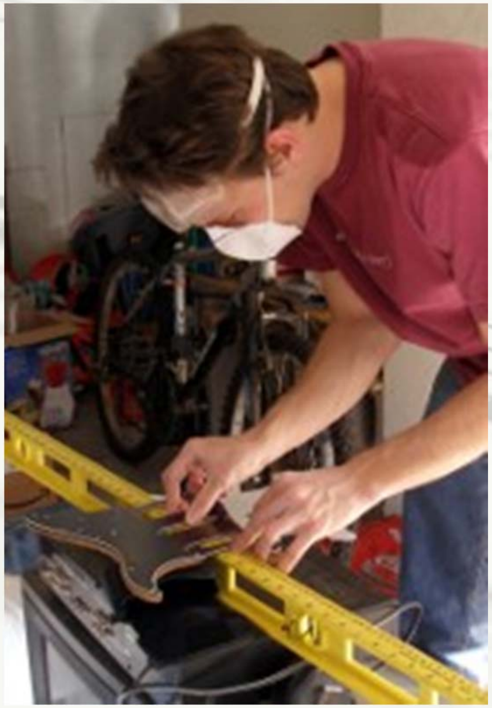
Metric Learning to Rank

Brian McFee, Gert Lanckriet

ICML 2010

presenter: Yin-Tzu Lin (阿孜孜^.^)　　2011/03

The Authors



Biran McFee

Prof. Gert Lanckriet

Computer **A**udition **L**aboratory,
Department of Electrical and Computer Engineering (**ECE**)
University of California, San Diego (**UCSD**)

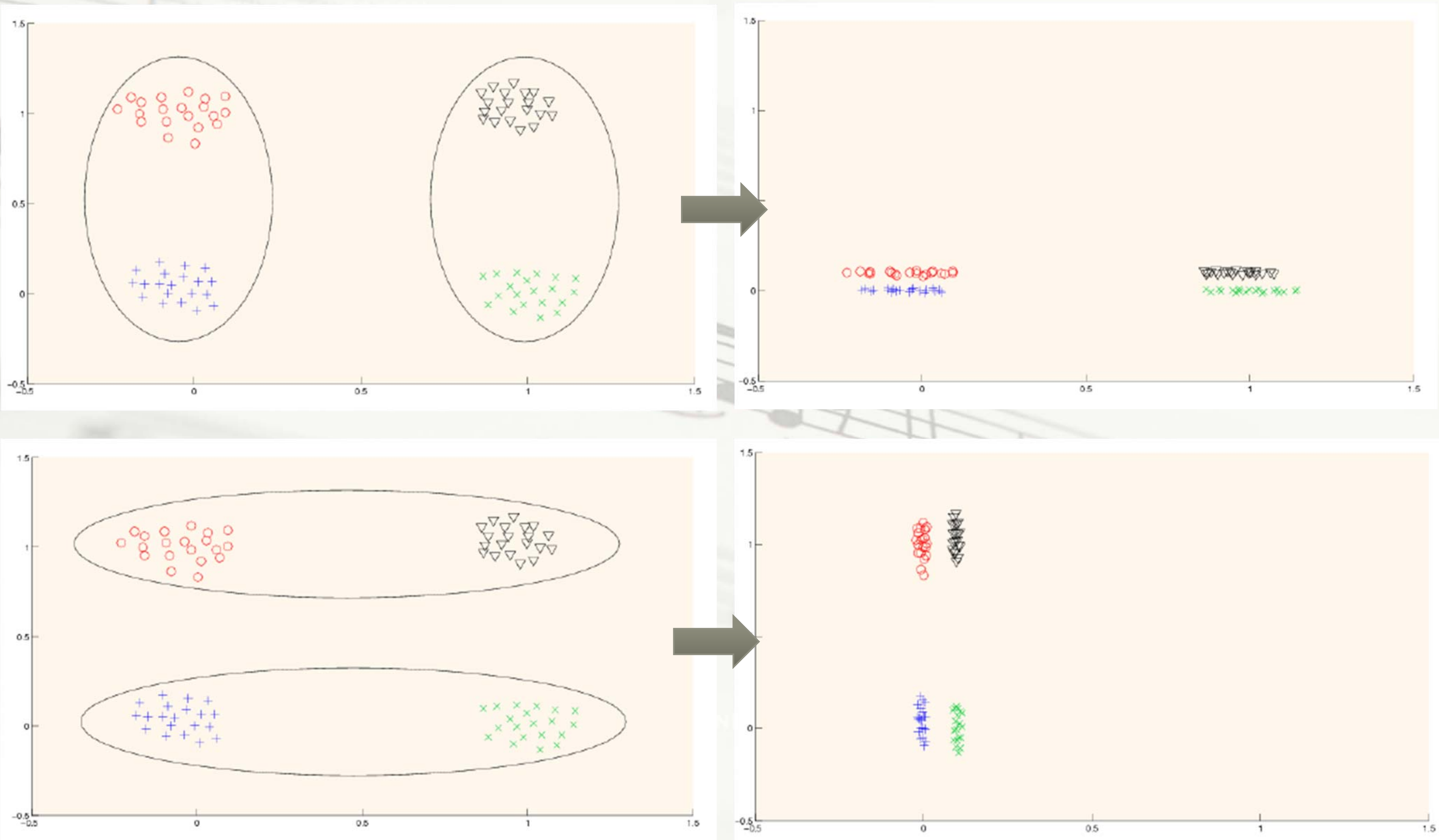
顧名思義

Metric Learning to Rank

= Metric Learning + Learning to Rank

Image from ICML 2010 Tutorial

Metric Learning



Metric Learning

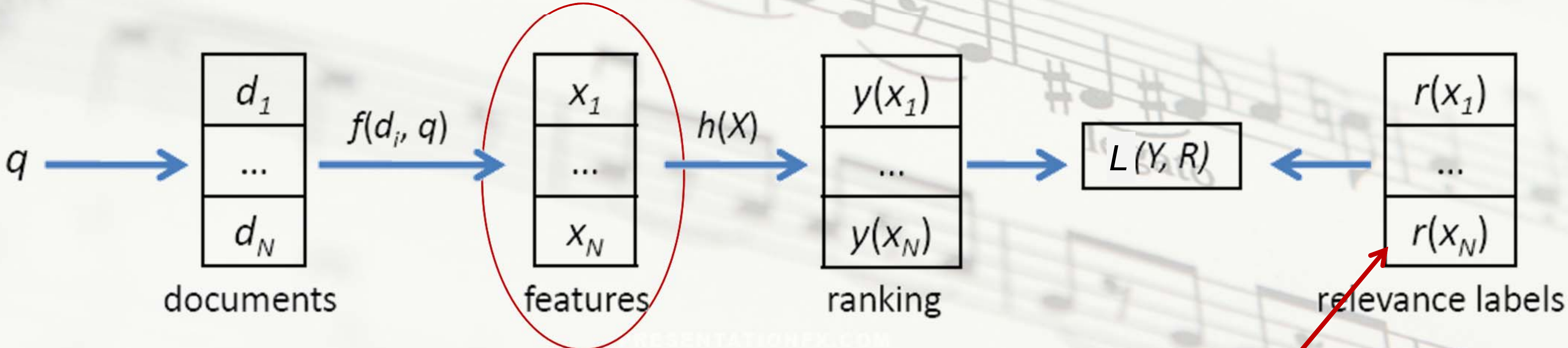
- Aims to learn a distance/similarity function for a given problem

$$
\begin{aligned}
d(\mathbf{x}_1, \mathbf{x}_2) &= ||\mathbf{x}_1 - \mathbf{x}_2||_W^2 \\
&= (\mathbf{x}_1 - \mathbf{x}_2)^T W (\mathbf{x}_1 - \mathbf{x}_2) \\
&= (\mathbf{x}_1 - \mathbf{x}_2)^T L^T L (\mathbf{x}_1 - \mathbf{x}_2) \\
&= ||L\mathbf{x}_1 - L\mathbf{x}_2||^2
\end{aligned}
$$

- Common methods
  - Unsupervised Methods:
    - PCA, Kernel PCA, MDS, *ISOMap*, *Laplacian Eigenmap*(LE), *Locally Linear Embedding*(LLE)
  - Supervised Methods :
    - LDA, Neighborhood Component Analysis (NCA), Large Margin NN Classifier (LMNN), Relevant Components Analysis (RCA), DistBoost
- Cons
  - Previous works only focus on classification problem
  - The same class lie closely

Learning to Rank

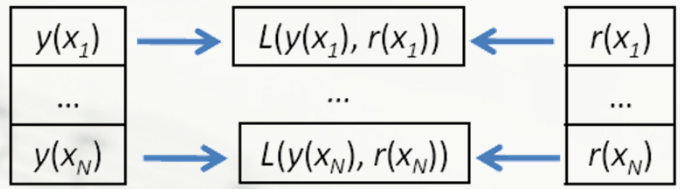
- Given query q and document collection $\{d_1, \ldots, d_N\}$
  - **Input**: query-document instances $X=\{x_1, \ldots, x_N\}$, $x_i = f(d_i,q)$, $x_i \in \mathbb{R}^d$
  - **Output**: ranking $Y=\{y(x_1), \ldots, y(x_N)\}$: permutation of X by ranker $h(x)$
  - **Evaluation (loss) function**: $L(Y, R)$, $R=\{r(x_1), \ldots, r(x_N)\}$: true relevance of x_i



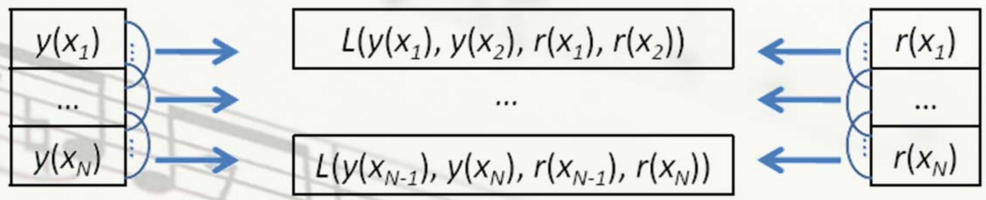
$r(x_i) \rightarrow (d_i, q)$ relevant or not, or level of relevant

Learning to Rank

- Common Approach
  - Point-wise



  - Pair-wise



  - List-wise (Structural)

- Cons
  - No parameterization on the distance metric during optimization

Goal of this work

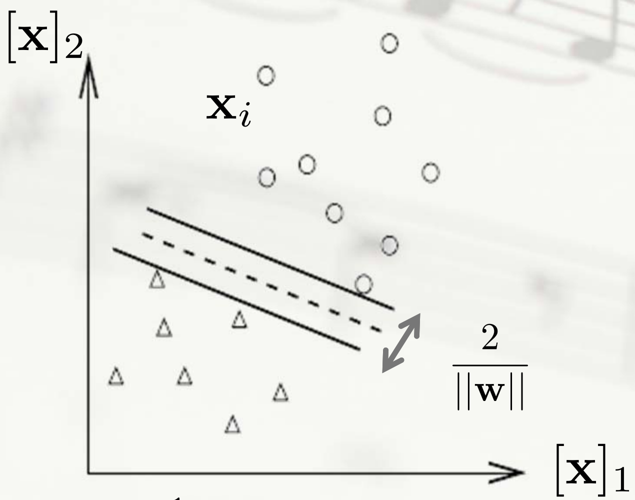
- Bridge the gap between metric learning and ranking
- Learning a distance function that optimize for true quantity of interest: the ranking
- Provide parameterization of ranking function by distance metric
  - Natural for information retrieval application

$[\mathbf{x}]_d$: d^{th} dimension of $\mathbf{x}$

Structured SVM

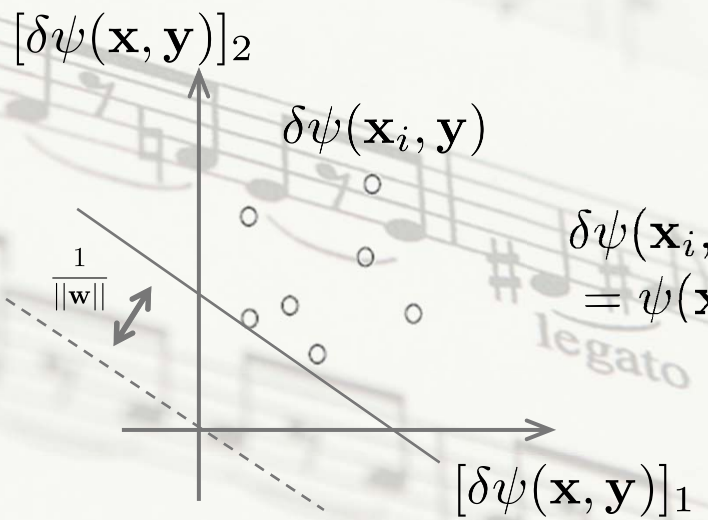
- Similar to SVM, but the ground truth is in complex structure

SVM $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$



$$\min_{\mathbf{w}}\left(\frac{1}{2}\|\mathbf{w}\|^2\right)$$
$$\text{s.t. } y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1$$
$$\forall \mathbf{x}_i \in \mathcal{X}, y_i \in \{0,1\}$$

Structured SVM $f(\mathbf{x}) = \arg\max_{\mathbf{y}}\langle(\mathbf{w}, \psi(\mathbf{x},\mathbf{y}))\rangle$



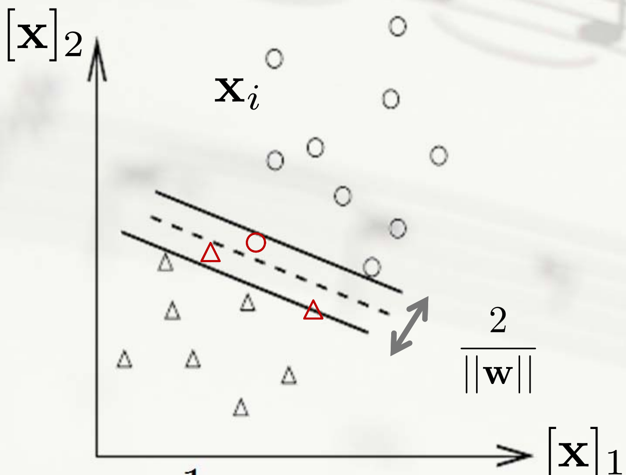
$$\delta\psi(\mathbf{x}_i,\mathbf{y}) = \psi(\mathbf{x}_i,\mathbf{y}_{\mathbf{x}_i}) - \psi(\mathbf{x}_i,\mathbf{y})$$

$$\min_{\mathbf{w}}\left(\frac{1}{2}\|\mathbf{w}\|^2\right)$$
$$\text{s.t. } \langle\mathbf{w}, \delta\psi(\mathbf{x}_i,\mathbf{y})\rangle \geq 1$$
$$\forall \mathbf{x}_i \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y}\backslash \mathbf{y}_{\mathbf{x}_i},\ \mathbf{y} \text{ is complex structure}$$

Soft Margin

- Add ξ_i to allow some outliers, avoiding over-fitting

SVM $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$

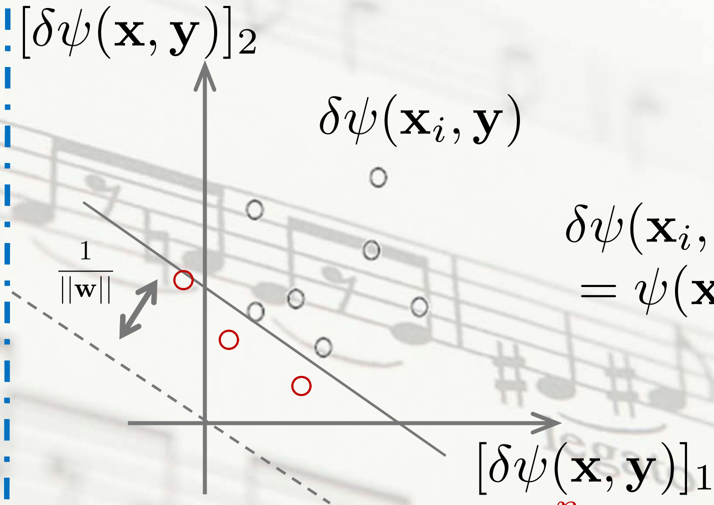


$$\min_{\mathbf{w},\xi}\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i\right)$$

s.t. $y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$

$\xi_i \geq 0$, $\forall \mathbf{x}_i \in \mathcal{X}, y_i \in \{0,1\}$

Structured SVM $f(\mathbf{x}) = \arg\max_{\mathbf{y}} \langle(\mathbf{w}, \psi(\mathbf{x},\mathbf{y}))\rangle$



$$\delta\psi(\mathbf{x}_i,\mathbf{y}) = \psi(\mathbf{x}_i,\mathbf{y}_{\mathbf{x}_i}) - \psi(\mathbf{x}_i,\mathbf{y})$$

$$\min_{\mathbf{w},\xi}\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\cdot\frac{1}{n}\sum_{i=1}^{n}\xi_i\right)$$

s.t. $\langle\mathbf{w}, \delta\psi(\mathbf{x}_i,\mathbf{y})\rangle \geq 1 - \xi_i, \quad \xi_i \geq 0$

s.t. $\langle\mathbf{w}, \delta\psi(\mathbf{x}_i,\mathbf{y})\rangle \geq \triangle(\mathbf{y}_{\mathbf{x}_i},\mathbf{y}) - \xi_i, \quad \xi_i \geq 0$

$\forall \mathbf{x}_i \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y}\backslash\mathbf{y}_{\mathbf{x}_i}$, $\mathbf{y}$ is complex structure

Notation

$\mathcal{X} \subset \mathbb{R}^d$	Input: the training set of n points in $\mathbb{R}^d$
$\mathcal{Y}$	Output: the set of permutations over $\mathcal{X}$
y_q^*	The true ranking for point q
$\Delta(y_q^*, y)$	The loss incurred by predicting y instead of y_q^*
$W \succeq 0$	The learned (positive semidefinite) metric $W = L^{\mathsf{T}} L$
$\|a - b\|_W$	The learned distance between a and b

Apply to ranking

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

$$\text{s.t. } \langle \mathbf{w}, \delta\psi(\mathbf{x}_i,\mathbf{y})\rangle \geq \Delta(\mathbf{y}_{\mathbf{x}_i},\mathbf{y}) - \xi_i$$

$$\xi_i \geq 0,\ \forall \mathbf{x}_i \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\backslash\mathbf{y}_{\mathbf{x}_i}$$

$$\delta\psi(\mathbf{x}_i,\mathbf{y}) = \psi(\mathbf{x}_i,\mathbf{y}_{\mathbf{x}_i}) - \psi(\mathbf{x}_i,\mathbf{y})$$

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\cdot\frac{1}{|\mathcal{X}|}\sum_{q\in\mathcal{X}}\xi_q \qquad , \forall q \in \mathcal{X}, y \in \mathcal{Y}\backslash\mathbf{y}_q^*$$

$$\text{s.t. } \langle w, \psi(q, y_q^*)\rangle - \langle w, \psi(q, y)\rangle \geq \Delta(y_q^*, y) - \xi_q, \quad \xi_q \geq 0$$

Score(good ranking)-Score(bad ranking) ≥ Loss(bad ranking)

Key Problem:

1. The definition of y and feature map $\psi(q,y)$
   (We often only know weather (q,d_i) is rel or not)
2. The definition of loss function $\Delta(y_q^*, y)$
3. Efficient algorithm

Partial order feature map

$$\psi_{po}(q,y) = \sum_{i \in \mathcal{X}_q^+} \sum_{j \in \mathcal{X}_q^-} y_{ij} \frac{\phi(q,i) - \phi(q,j)}{|\mathcal{X}_q^+| \cdot |\mathcal{X}_q^-|}$$

$\mathcal{X}_q^+$: relevant docs set of q (ground truth)

$\mathcal{X}_q^-$: irrelevant docs set of q (ground truth)

$$i \in \mathcal{X}_q^+,\ j \in \mathcal{X}_q^-,\ y_{ij} = \begin{cases} +1 & i \text{ before } j \\ -1 & i \text{ after } j \end{cases}$$

$\forall i, j,\ y_{ij}^* = 1$

$\phi(q,i)$: retrival model results vector (page rank, TF-IDF, etc.)

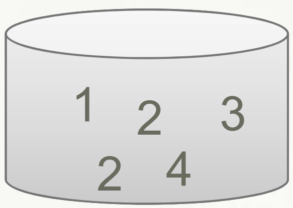
Note:

At testing stage, the predicted $\hat{y}$ is sorting $\langle w, \phi(q,i) \rangle$ in descending order

$$\begin{aligned} \hat{y} &= \arg\max_y \langle w, \psi(q,y) \rangle \\ &= \langle w, \phi(q,i) \rangle \searrow i \in \mathcal{X} \end{aligned}$$

---

q=2



Good rankings:

y_{21} = 1
y_{23} = 1
y_{24} = 1
…

2	2
2	2
1	3
4	1
3	4

. . .

Bad rankings:

y_{21} = -1
y_{23} = -1
y_{24} = 1

2	4
3	3
1	**2**
2	1
4	2

y_{21} = 1
y_{23} = -1
y_{24} = -1

. . .

Link to metric learning

Making the learned metric in terms of Frobenius Inner Product

$$\begin{aligned}\|q-i\|_W^2 &= (q-i)^T W (q-i) \\ &= tr(W(q-i)(q-i)^T) \\ &= tr(W^T(q-i)(q-i)^T) \\ &= \langle W, (q-i)(q-i)^T \rangle_F\end{aligned}$$

Leads to a nature choice of ϕ:

$$\phi(q,i) = -(q-i)(q-i)^T$$

Note:

$$\begin{aligned}d(\mathbf{x}_1,\mathbf{x}_2) &= ||\mathbf{x}_1-\mathbf{x}_2||_W^2 \\ &= (\mathbf{x}_1-\mathbf{x}_2)^T W(\mathbf{x}_1-\mathbf{x}_2) \\ &= (\mathbf{x}_1-\mathbf{x}_2)^T L^T L(\mathbf{x}_1-\mathbf{x}_2) \\ &= ||L\mathbf{x}_1-L\mathbf{x}_2||^2\end{aligned}$$

Frobenius Inner Product:

$$\begin{aligned}\langle A, B\rangle_F &= \textstyle\sum_i \sum_j A_{ij}B_{ij} \\ &= trace(A^T B)\end{aligned}$$

$$trace(A) = \sum_i A_{ii}$$

Note:

$$\langle w, \psi(q, y_q^*)\rangle - \langle w, \psi(q,y)\rangle \geq \triangle(y_q^*, y) - \xi_q$$

$$\psi_{po}(q,y) = \sum_{i\in\mathcal{X}_q^+}\sum_{j\in\mathcal{X}_q^-} y_{ij}\frac{\phi(q,i)-\phi(q,j)}{|\mathcal{X}_q^+|\cdot|\mathcal{X}_q^-|}$$

➡ Sorting ascending $\|q-i\|_W^2 \equiv$ sorting desc $\langle W, \phi(q,i)\rangle_F$.
The predicted order ($\hat{y}$) will maximize $\langle W, \psi_{po}(q,\hat{y})\rangle_F$

Loss Function $\Delta(y_q^*, y)$

- $\Delta(y_q^*, y) \leftarrow \text{score}(y_q^*) - \text{score}(y) = 1 - \text{score}(y)$

$\text{score} \in \{\text{AUC, Pre@k, MAP, MRR, NDCG}\}$

  - Area under ROC Curve (AUC)
  - Precision@k
  - Mean Average Precision (MAP)

$$AP(q) = \frac{1}{|\mathcal{X}_q^+|} \sum_{k=1}^{|\mathcal{X}_q^+|+|\mathcal{X}_q^-|} \text{Prec@k} \cdot \mathbb{1}[k \in \mathcal{X}_q^+]\ ,\ MAP = \sum_{q \in \mathcal{Q}} AP(q)$$

  - Mean Reciprocal Rank (MRR)

$$MRR(q) = \frac{1}{|\mathcal{X}_q^+|} \sum_{k=1}^{|\mathcal{X}_q^+|} \frac{1}{rank(k)}$$

  - Normalized Discounted Cumulative Gain (NDCG)

$$NDCG(q; y; k) = \frac{\sum_{i=1}^{k} D(i) \mathbb{1}[i \in \mathcal{X}_q^+]}{\sum_{i=1}^{k} D(i)},\ D(i) = \begin{cases} 1 & i = 1 \\ \frac{1}{\log_2(i)} & 2 \le i \le k \end{cases}$$

$\frac{1}{2}\|w\|^2 = \frac{1}{2}w^T w$

$\rightarrow \frac{1}{2}\langle W, W\rangle_F = \frac{1}{2}tr(W^T W)$

Summary

$$\min_{W,\xi} \frac{1}{2}tr(W^T W) + C \cdot \frac{1}{|\mathcal{X}|}\sum_{q\in\mathcal{X}} \xi_q \ , \ \forall q \in \mathcal{X}, y \in \mathcal{Y}\backslash \mathbf{y}_q^*$$

$$\text{s.t. } \langle W, \psi_{po}(q, y_q^*)\rangle_F - \langle W, \psi_{po}(q, y)\rangle_F \geq \triangle(y_q^*, y) - \xi_q, \quad \xi_q \geq 0$$

$$\psi_{po}(q,y) = \sum_{i\in\mathcal{X}_q^+}\sum_{j\in\mathcal{X}_q^-} y_{ij}\frac{\phi(q,i)-\phi(q,j)}{|\mathcal{X}_q^+|\cdot|\mathcal{X}_q^-|}$$

$$\phi(q,i) = -(q-i)(q-i)^T$$

$$\begin{aligned}\triangle(y_q^*, y) &= \text{score}(y_q^*) - \text{score}(y)\\ &= 1 - \text{score}(y)\end{aligned}$$

score $\in$ {AUC, Pre@k, MAP, MRR, NDCG}

Solving structured SVM

- $|\mathcal{Y}|$ is super-exponential
- Cutting-plane algorithm
  - 1-slack scaling $\xi_q, \forall q \in \mathcal{X} \rightarrow \xi$
  - alternates between updating the set of constraint and finding W and ξ
  - Until the loss of new constraint $< \epsilon$
- Replace 2-norm to 1-norm for sparsity

$$\frac{1}{2}tr(W^TW) \rightarrow tr(W)$$

Input: data $\mathcal{X}$, rankings $y_1^*, \ldots, y_n^*$, slack trade-off $C > 0$, accuracy threshold $\epsilon > 0$

Output: metric $W \succeq 0$, slack variable $\xi \geq 0$

$\boxed{\mathcal{C} \leftarrow \emptyset}$ Set of constraints

repeat

Solve for the optimal metric and slack:

$$(W, \xi) \leftarrow \mathrm{argmin}_{W,\xi}\, f(W, \xi) = \mathrm{tr}(W) + C\xi$$
$$\text{s.t. } W \succeq 0$$
$$\xi \geq 0$$
$$\forall (y_1, y_2, \ldots, y_n) \in \mathcal{C} :$$
$$\frac{1}{n}\sum_{i=1}^{n} \langle W, \delta\psi_{po}(q_i, y_i^*, y_i)\rangle_F \geq \frac{1}{n}\sum_{i=1}^{n} \Delta(y_i^*, y_i) - \xi$$

Find W and ξ

for $i = 1$ **to** n **do**

$y_i \leftarrow \mathrm{argmax}_{y \in \mathcal{Y}}\, \Delta(y_i^*, y) + \langle W, \psi_{po}(q_i, y)\rangle_F$

end for

$\mathcal{C} \leftarrow \mathcal{C} \cup \{(y_1, \ldots, y_n)\}$

until

$$\frac{1}{n}\sum_{i=1}^{n} \Delta(y_i^*, y_i) - \langle W, \delta\psi_{po}(q_i, y_i^*, y_i)\rangle_F \leq \xi + \epsilon$$

Find rankings y that most violate

$$\langle W, \psi_{po}(q, y_q^*)\rangle_F - \langle W, \psi_{po}(q, y)\rangle_F \geq \triangle(y_q^*, y) - \xi_q$$

→ add to $\mathcal{C}$

Can reduce to

sort $\forall i \in \mathcal{X}_q^+$ by desc $\langle W, \phi(q, i)\rangle_F$

sort $\forall j \in \mathcal{X}_q^-$ by desc $\langle W, \phi(q, i)\rangle_F$

Find a interleaving of the above

Terminate if error < ε

Experiment

- Classification on UCI Data
- Ranking on eHarmony Data
- Apply to Music Similarity [McFee et al. ISMIR 2010]

Classification Result

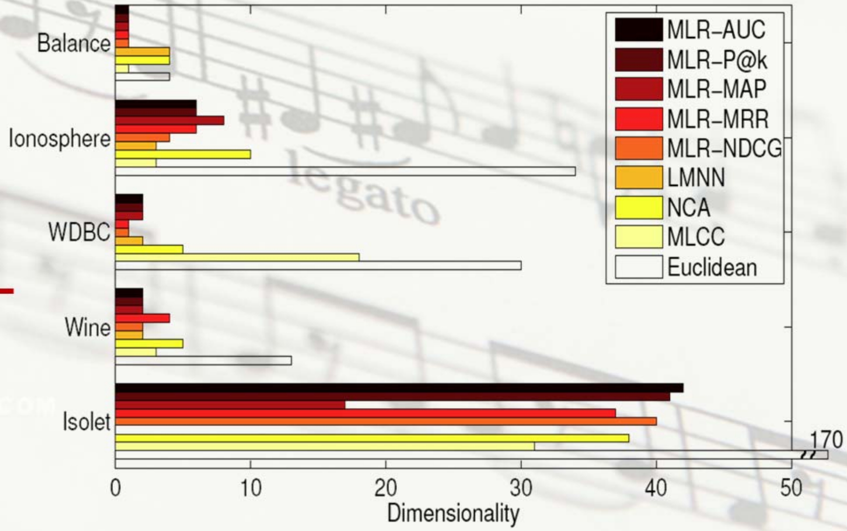
- UCI Dataset

	d	# Train	# Test	# Classes
Balance	4	500	125	3
Ionosphere	34	281	70	2
WDBC	30	456	113	2
Wine	13	143	35	3
IsoLet	170	6238	1559	26

KNN Classification error(%)

Algorithm	Bal.	Ion.	Wdbc	Wine	Isolet
MLR-AUC	7.9	12.3	2.7	1.4	4.5
MLR-P@k	8.2	12.3	2.9	1.5	4.5
MLR-MAP	6.9	12.3	2.6	1.0	5.5
MLR-MRR	8.2	12.1	2.6	1.5	4.5
MLR-NDCG	8.2	11.9	2.9	1.6	4.4
LMNN	8.8	11.7	2.4	1.7	4.7
NCA	4.6	11.7	2.6	2.7	10.8
MLCC	5.5	12.6	2.1	1.1	4.4
Euclidean	10.3	15.3	3.1	3.1	8.1

Dimension Reduction



Ranking Results

- eHarmony: a online dating service witch matching users by personality traits

	Matchings	Unique users	Queries
Training	506688	294832	22391
Test	439161	247420	36037

- Results

Algorithm	AUC	MAP	MRR	Time	$\lvert\mathcal{C}\rvert$
MLR-AUC	0.612	0.445	0.466	232	7
MLR-MAP	0.624	0.453	0.474	2053	23
MLR-MRR	0.616	0.448	0.469	809	17
SVM-MAP	0.614	0.447	0.467	4968	36
Euclidean	0.522	0.394	0.414		

Apply to Music Similarity

- Swat10k Dataset
  - 10,870 songs
  - 3,748 unique artists

	Training	Validation	Test	Discard
# Artists	746	700	700	1602
# Songs	1842	1819	1862	5347
# Relevant	39.5	37.7	36.4	

- Ground truth source
  - Collaborative filtering (from last.fm) on artist similarity

$$F_{ui} = \begin{cases} 1 & \text{user } u \text{ listened to artist } i \\ 0 & \text{otherwise,} \end{cases}$$

$$S_{ij} = \frac{F_i^\top F_j}{\|F_i\| \cdot \|F_j\|}$$

    - Discard artist that fewer than 100 user
    - set the top 10 in the collaborative score(S_{ij}) list as relevant
  - Transfer to song-level similarity (songs of the same artist share the same relevant scores)

Apply to Music Similarity(2)

$$k(u,v) = \exp\left(-\chi^2(u,v)\right)$$

$$\chi^2(u,v) = \sum_{i=1}^{5000} \frac{(u_i - v_i)^2}{u_i + v_i}.$$

- Features
  - △MFCC (39 dim MFCCs)
    - Random pick 1000 songs, pick 1000 MFCCs for each → 5000 cluster (codewords)
    - Represent a song as the histogram of the 5000 codewords
    - Further represent a song with chi-square distance to each song in the training set (PCA to 39 dim)
  - Auto Tagging
    - Auto tagger of [D. Turnbull et al. TASLP Feb. 2008]
    - 149 dim vector: the i^{th} dim ← the prob(i^{th} tag applies to the song), given the observed △MFCCs
  - Human Tagging
    - Tags from Pandora Music Genome Project
    - 1053 dim 0101 weakly-label vector

Apply to Music Similarity(3)

- Results

Data source	AUC	MAP	MRR
MFCC	0.630	0.057	0.249
Optimized MFCC	0.719	0.081	0.275
Auto-tags	0.726	0.090	0.330
Optimized auto-tags	0.776	0.116	0.327
Human tags	0.770	0.187	0.540
Optimized human tags	0.939	0.420	0.636

Conclusion

- Proposed a metric learning algorithm optimize for rank-based loss function
- MLR improves over baseline Euclidean distance
  - But linear model may not suffice to capture ranking structure
  - Future direction: incorporate non-linear transformations

THANK YOU^__^