Distance Metric Learning

Outlines
- I Introduction
- II Application of DML

Lei Wu
University of Pittsburgh

How to measure distance?

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$

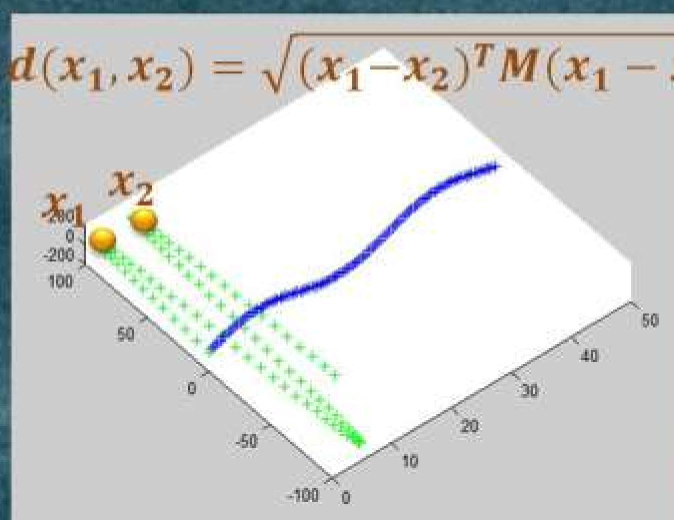
$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T (x_1 - x_2)} \longrightarrow M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$d(x_1, x_2) = |x_{11} - x_{21}| \longrightarrow M = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

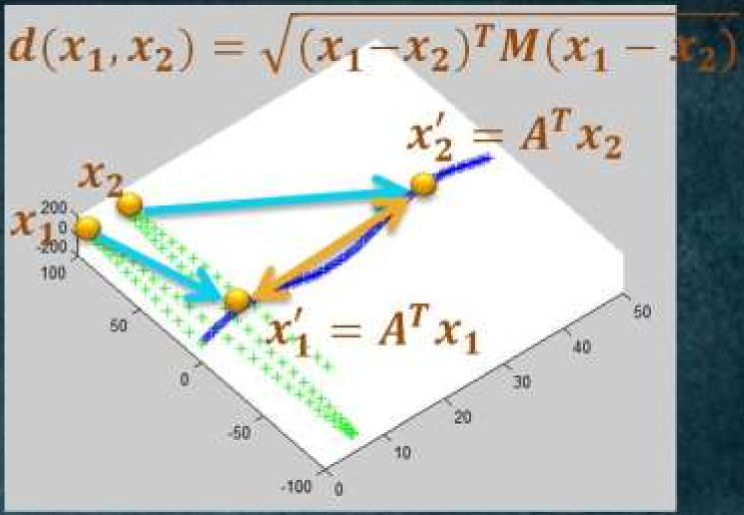
$$d(x_1, x_2) = |x_{12} - x_{22}| \longrightarrow M = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$d(x_1, x_2) = 0 \longrightarrow M = \begin{bmatrix} sin^2(\alpha) - cos^2(\alpha) & 2sin(\alpha)cos(\alpha) \\ 0 & -cos^2(\alpha) + sin^2(\alpha) \end{bmatrix}$$

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$



How to measure distance?



$$d(x_1, x_2) = \sqrt{(A^T x_1 - A^T x_2)^T (A^T x_1 - A^T x_2)}$$

learn A

generate M

apply M to new sample x

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T A A^T (x_1 - x_2)}$$

$$d(x_1, x) = \sqrt{(x_1 - x)^T M (x_1 - x)}$$

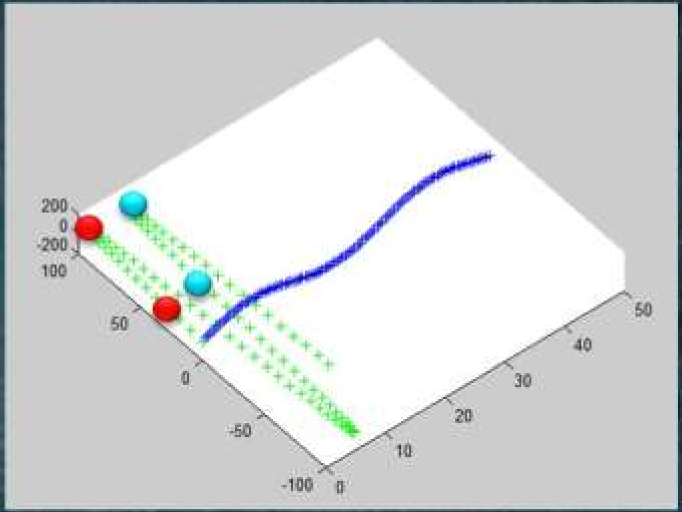
$$M = AA^T$$

Categorization

- **Unsupervised DML**
  - ***Linear Model***
    - *PCA,*
    - *MDS*
  - ***Nonlinear Model***
    - *LLE,*
    - *ISOMAP,*
    - *Laplacian Eigenmaps*
    - *Kernel PCA*
- **Supervised DML**
  - **Global Distance Metric Learning**
    - *Probabilistic Global Distance Metric Learning (PGDM)*
  - **Local Distance Metric Learning**
    - *Neighborhood Components Analysis (NCA)*
    - *Relevant Component Analysis (RCA)*
    - *Discriminative Component Analysis (DCA)*
    - *Probabilistic Relevant Component Analysis (pRCA)*
    - *Large Margin Nearest Neighbor (LMNN)*
    - *Information-Theoretic Metric Learning (ITML)*
    - *Bregman Distance Function Learning (BDFL)*
  - **Certain Constraint**
    - *NCA, RCA, DCA, LMNN,ITML,BDFL, etc*
  - **Uncertain Constraint**
    - *pRCA*
  - **Matrix Form**
    - *NCA, RCA, DCA, LMNN,ITML,pRCA, etc*
  - **Functional Form**
    - *BDFL*

Supervised metric learning?

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$



Object function

$$L(M)$$

Side information

Similar pairs

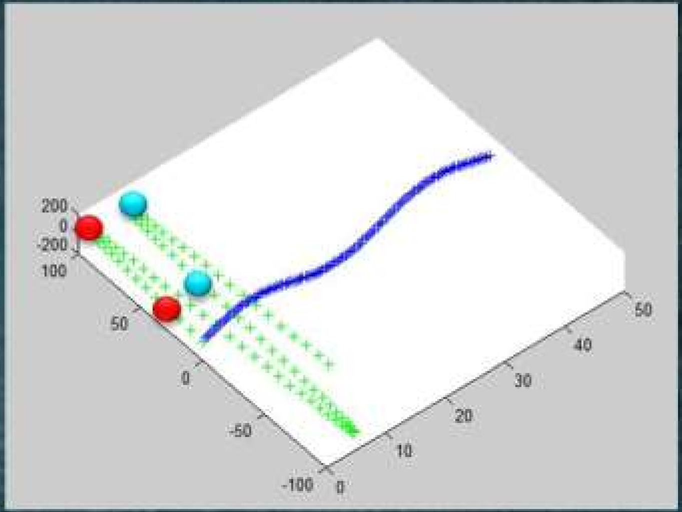
Dissimilar pairs

Regularization

$$R(M)$$

Supervised metric learning?

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$



Object function

$L(M)$

Side information

Similar pairs

Dissimilar pairs

Regularization

$R(M)$

Object function

Probabilistic Global Distance Metric Learning (PGDM)

Map similar points close to each other

$$\min_A \sum_{(x_i,x_j)\in S} \left\|x_i - x_j\right\|_A$$

$$s.t. \sum_{(x_i,x_j)\in D} \left\|x_i - x_j\right\|_A \geq 1$$

Minimize the distance between similar samples

Preserve certain distance between others

Extended Reading:

E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, Distance metric learning, with application to clustering with side-information, *Advances in Neural Information Processing Systems* 15, vol. 15, 2002, pp. 505-512.

Object function

Neighborhood Components Analysis (NCA)

Neighbors should gain probability to be in the same class

$$p_{ij} = \frac{exp(-||Ax_i - Ax_j||^2)}{\sum_{k \neq i} exp(-||Ax_i - Ax_k||^2)}$$

$$max_A \sum_{t=1}^{n} log(\sum_{j \in C_i} p_{ij})$$

Minimize the distance between neighbors

j is neighbor of i

Extended Reading:

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. NIPS*, 2005.

Object function

Relevant Component Analysis (RCA)

Weight the distance with the covariance

Covariance of the data

$$C = \frac{1}{p}\sum_j \sum_i (x_{ji} -$$

Weight the distance by covarience

$$d(x_i, x_j) = \sqrt{(x_i - x_j)C^{-1}(x_i - x_j}$$

Equivalent to max the mutual information

$$max_f I(X,Y)$$

$$\Leftrightarrow \quad s.t.\frac{1}{p}\sum_j \sum_i \left|\left|x_{ji} - m_j\right|\right|^2 \leq K, Y = f(X) = AX$$

Extended Reading:

N. Shental and D. Weinshall, Learning distance functions using equivalence relations, In Proceedings of the Twentieth International Conference on Machine Learning, vol. 21, 2003, pp. 11-18.

Object function

Information maximiz

Mutual information
between original data
and embedded data

Find a mapping that maximize the mutual info ***nal space and that in transformed space***

$$\max_A I(X,Y)$$

$$s.t.\ \frac{1}{p}\sum_j \sum_i \left\| y_{ji} - m_j \right\|^2 \le K \qquad Y = AX$$

$$I(X,Y) = H(Y) - H(Y|X) \Rightarrow \max I(X,Y) = \max H(Y) \Rightarrow \max|A| \Rightarrow \max|M|$$

$$p(y) = \frac{p(x)}{|J(x)|} \Rightarrow H(Y) = -\int p(y) \log p(y)\, dy = -\int p(x) \log \frac{p(x)}{|J(x)|}\, dx = H(X) + \log|J(X)|$$

$$\Longrightarrow \quad \max_M |M| \quad s.t.\ \frac{1}{p}\sum_j \sum_i \left\| x_{ji} - m_j \right\|_M^2 \le K, \quad M \succ 0$$

$$\Longrightarrow \quad M = \frac{K}{N} C^{-1}$$

Extended Reading:

A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. International Conference on Machine Learning*, 2003.

Object function

Discriminative Component Analysis (DCA)

Minimize the distance within each chunklet and maximize the distance between chunkle

Covariance between chunklets

$$C_b = \frac{1}{k}\sum_i \sum_j (m_i - m_j)(m_i - m_j)^T$$

Covariance within chunklet

$$C_w = \frac{1}{n}\sum_i \sum_j (x_{ji} - m_j)(x_{ji} - m_j)^T$$

$$J(A) = argmax_A \frac{A^T C_b A}{A^T C_w A}$$

Extended Reading:

Steven C. H. Hoi, Wei Liu, Michael R. Lyu, and Wei-Ying Ma. 2006. Learning Distance Metrics with Contextual Constraints for Image Retrieval. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (CVPR '06), 2006.

Object function

probabilistic Relevance Component Analysis (pRCA)

Learnin… …d on probabilistic side information. Minimize the distance betwee… …probability to belong to same chunklet and maximize the distanc… …vith low pr… …chunklets

Optimize the membership probability, as well as Metric

Probabilistic side information

$$\min_{M \succ 0, P, m} \sum_i \sum_j p_{ij} \|x_i - m_j\|_M^2 - \lambda \log |M|$$

$$s.t.\ \|P - P_0\| < \gamma,$$

$$\sum_i p_{ij} = 1, p_{ij} > 0$$

Extended Reading:

Lei Wu, Steven C.H. Hoi, Rong Jin, Jianke Zhu, Nenghai Yu, "Distance Metric Learning from Uncertain Side Information with Application to Automated Photo Tagging", *ACM International Conference on Multimedia (MM'09)*, 2009.

Object function

Large Margin Nearest Neighbor (LMNN)

Maximize the margin between the distance of similar samples and the distance of dissimilar samples

$$min_{M \succ 0} \sum \eta_{ij} \|x_i - x_j\|_M^2 + c \sum_{ijl} \eta_{ij}(1 - y_{il}) \left(1 + \|x_i - x_j\|_M^2 - \|x_i - x_l\|_{M}^2 \right)_+$$

Select neighbor
ij are the ne
neighbor in th
label

penalizes large distances between each input and its target neighbors

Select samples with different labels

penalizes small distances between each input and all other inputs that do not share the same label

Extended Reading:

Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.* 10 (June 2009), 207-244.

Object function

Information-Theoretic Metric Learning (ITML)

Regularize the Mahalanobis matrix M to be as close as possible to a given Mahalanobis distance function, parameterized by M_0

$$\min_{M \succ 0} KL(p(x, M_0) || p(x, M))$$
$$s.t.\ d_M(x_i, x_j) < u, (x_i, x_j) \in S$$
$$d_M(x_i, x_j) > l, (x_i, x_j) \in D$$
$$p(x, M) = \frac{1}{Z} exp(-\frac{1}{2} d_M(x, u))$$

Minimize the KL divergence to the predefined distance

Make the distance between Similar samples smaller than u

Make the distance between dissimilar samples larger than l

Extended Reading:

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning* (ICML '07),2007.

Object function

Bregman Distance Function Learning (BDFL)

Extend the Mahalanobis matrix M to Bregman functional form $\nabla^2\varphi(x)$

Use a function form rather than matrix form for distance metric

$$d(\boldsymbol{x_1},\boldsymbol{x_2}) = \boldsymbol{\varphi}(\boldsymbol{x_1}) - \boldsymbol{\varphi}(\boldsymbol{x_2}) - (\boldsymbol{x_1} - \boldsymbol{x_2})^\top \nabla\varphi(\boldsymbol{x_2})$$
$$\Rightarrow d(\boldsymbol{x_1},\boldsymbol{x_2}) = (\boldsymbol{x_1} - \boldsymbol{x_2})^\top \nabla^2\varphi(\widehat{\boldsymbol{x}})(\boldsymbol{x_1} - \boldsymbol{x_2})$$

$$\boldsymbol{min}_{\varphi\in\Omega(\mathcal{H}_k),b\in R^+}\ \frac{1}{2}|\varphi|^2_{\mathcal{H}_k} + C\sum_i \ell(y^i[d(x_1^i, x_2^i) - b^i])$$

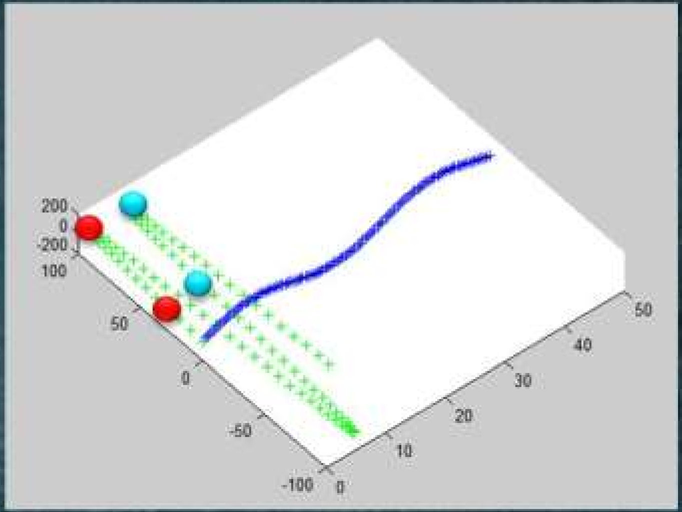
Search for an optimal convex function from a Reproducing Kernel Hilbert Space

Extended Reading:

Lei Wu, Rong Jin, Steven C.H. Hoi, Jianke Zhu, Nenghai Yu, "Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering", *Advances in Neural Information Processing Systems* ***(NIPS'09)***, *2009.*

Supervised metric learning?

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$



Object function

$$L(M)$$

Side information

Similar pairs

Dissimilar pairs

Regularization

$$R(M)$$

Side information

Probabilistic Global Distance Metric Learning (PGDM)

Map similar points close to each other

$$\min_A \sum_{(x_i, x_j) \in S} \left\| x_i - x_j \right\|_A^2$$

$$s.t. \sum_{(x_i, x_j) \in D} \left\| x_i - \cdots \right\| \geq 1$$

S, D: hard side info

Extended Reading:

E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, Distance metric learning, with application to clustering with side-information, *Advances in Neural Information Processing Systems* 15, vol. 15, 2002, pp. 505-512.

Side information

Neighborhood Components Analysis (NCA)

Pairwise hard constraint

$$p_{ij} = \frac{exp(-||Ax_i - Ax_j||^2)}{\sum_{k \neq i} exp(-||Ax_i - Ax_k||^2)}$$

$$max_A \sum_{t=1}^{n} log(\sum_{j \in C_i} p_{ij})$$

side info:
$j \in C_i$ *or* $j \notin C_i$

Extended Reading:

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. NIPS*, 2005.

Side information

Information maximization RCA

Find a mapping that maximize the mutual information of sample in original space and that in transformed space

$$\max_A I(X,Y)$$

$$s.t. \frac{1}{p}\sum_j \sum_i \left\| y_{ji} - m_j \right\|^2 \le K \qquad Y = AX$$

$$I(X,Y) = H(Y) - H(Y|X) \Rightarrow \max I(X,Y) = \max H(Y) \Rightarrow \max |A| \Rightarrow \max |M|$$

$$p(y) = \frac{p(x)}{|J(x)|} \Rightarrow H(Y) = -\int p(y) \log p(y) dy = -\int p(x) \log \frac{p(x)}{|J(x)|} dx = H(X) + \log|J(X)|$$

$$\Longrightarrow \quad \max_M |M| \quad s.t. \frac{1}{p}\sum_j \sum_i \left\| x_{ji} - m_j \right\|_M^2 \le K, \quad M \succ 0$$

$$\Longrightarrow \quad M = \frac{K}{N} C^{-1}$$

x_{ji} i-th sample in j-th chunklet
Hard constraint

Extended Reading:

A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. International Conference on Machine Learning*, 2003.

Side information

Discriminative Component Analysis (DCA)

Minimize the distance within each chunklet and maximize the distance between chunklets

$$C_b = \frac{1}{k}\sum_i \sum_j (m_i - m_j)(m_i - m_j)^T$$

$$C_w = \frac{1}{n}\sum_i \sum_j (x_{ji} - m_j)(x_{ji} - m_j)^T$$

$$J(A) = argmax_A \frac{}{A^T C}$$

x_{ji} i-th sample in j-th chunklet
Hard constraint

Extended Reading:

Steven C. H. Hoi, Wei Liu, Michael R. Lyu, and Wei-Ying Ma. 2006. Learning Distance Metrics with Contextual Constraints for Image Retrieval. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (CVPR '06), 2006.

Side information

Large Margin Nearest Neighbor (LMNN)

Maximize the margin between the distance of similar samples and the distance of dissimilar samples

$$min_{M\succ 0} \sum \eta_{ij} \|x_i - x_j\|_M^2 + c \sum_{ijl} \eta_{ij}(1 - y_{il}) \left(1 + \|x_i - x_j\|_M^2 - \|x_i - x_l\|_{M}{}^2\right)_+$$

Nearest neighbor +
hard constraints

Hard constraints

Extended Reading:

Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.* 10 (June 2009), 207-244.

Side information

Information-Theoretic Metric Learning (ITML)

Regularize the Mahalanobis matrix M to be as close as possible to a given Mahalanobis distance function, parameterized by M_0

$$
\begin{aligned}
& min_{M \succ 0} KL(p(x, M_0) || p(x, M)) \\
& s.t.\ d_M(x_i, x_j) < u, (x_i, x_j) \in S \\
& d_M(x_i, x_j) > l, (x_i, x_j) \in D \\
& p(x, M) = \frac{1}{Z} exp(-\frac{1}{2} d_M(x, u))
\end{aligned}
$$

S and D are hard constraints

Extended Reading:

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning* (ICML '07),2007.

Side information

Bregman Distance Function Learning (BDFL)

Pairwise hard constraints

$$d(x_1, x_2) = \varphi(x_1) - \varphi(x_2) - (x_1 - x_2)^\top \varphi^2(x_2)$$
$$\Rightarrow d(x_1, x_2) = (x_1 - x_2)^\top \nabla^2 \varphi(\hat{x})(x_1 - x_2)$$

$$min_{\varphi \in \Omega(\mathcal{H}_k), b \in R^+} \frac{1}{2} |\varphi|^2_{\mathcal{H}_k} + C \sum_i \ell(y^i [d(x_1^i, x_2^i) - b^i])$$

$Side\ info:$
$x_1\ x_2\ similar:\ y = 1$
$x_1\ x_2\ dissimilar: y = -1$

Extended Reading:

Lei Wu, Rong Jin, Steven C.H. Hoi, Jianke Zhu, Nenghai Yu, "Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering", *Advances in Neural Information Processing Systems* ***(NIPS'09)***, *2009.*

Side information

probabilistic Relevance Component Analysis (pRCA)

Probabilistic constraints

$$\min_{M \succ 0, P, m} \sum_i \sum_j p_{ij} \|x_i - m_j\|_M^2 - \lambda \log|M|$$

Probabilistic side information

$$\ldots \| < \gamma,$$

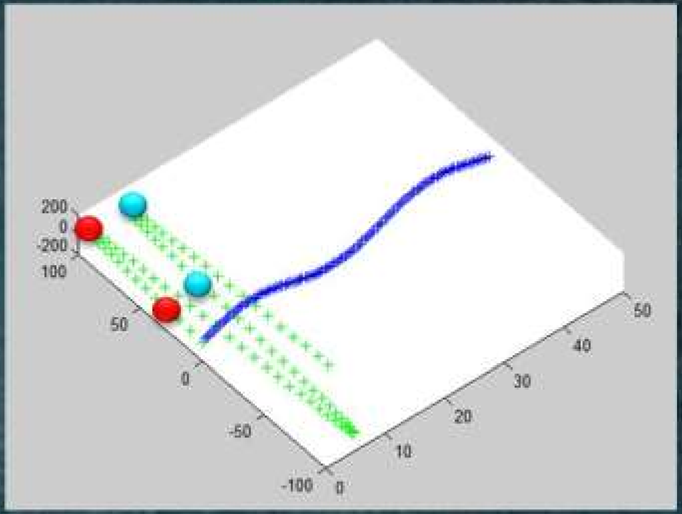
$$\sum_i \ldots_j > 0$$

Extended Reading:

Lei Wu, Steven C.H. Hoi, Rong Jin, Jianke Zhu, Nenghai Yu, "Distance Metric Learning from Uncertain Side Information with Application to Automated Photo Tagging", *ACM International Conference on Multimedia (MM'09)*, 2009.

Supervised metric learning?

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$



Object function

$L(M)$

Side information

Similar pairs

Dissimilar pairs

Regularization

$R(M)$

Regularization

Introducing additional information in order to solve an ill-posed problem or to prevent overfitting

Basic regularizers:

- **Bayesian information criterion $\|M\|_0$**
  - **Equal to minimum description length criterion (MDL)**
- **Least absolute shrinkage and selection operator (Lasso) $\|M\|_1$**
  - **Fundamental to compressed sensing**
- **Tikhonov regularization (ridge regression) $\|M\|_2$**
  - **Common method to handle ill posed problem**

Extended Reading:

A. Neumaier, Solving ill-conditioned and singular linear systems: A tutorial on regularization, SIAM Review 40 (1998), 636-666.

Regularization

Introducing additional information in order to handle specific requirements, such as sparsity

Some other regularizers:

- $\boldsymbol{tr}(\boldsymbol{M})$
  - **Regularize diagonals**
- $\boldsymbol{log\ det}(\boldsymbol{M})$
  - **Equal to $\boldsymbol{tr}(\exp(\boldsymbol{M}))$**
- $\|\boldsymbol{M}\|_{(2,1)} = \sum_i \left(\sum_j |m_{ij}|^2\right)^{\frac{1}{2}}$
  - **Regularize columns**

Extended Reading:
Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha and Tat-Seng Chua and Hong-Jiang Zhang, An Efficient Sparse Metric Learning in High-Dimensional Space via l1-Penalized Log-Determinant Regulaization, ICML 2009

Regularization

Introducing additional information in order to handle specific requirements, such as sparsity

Generalized regularizers for metric learning:

- $tr(LM)$
  - $If\ L = I \Rightarrow tr(M) \Rightarrow Generalized\ Sparese\ metric\ learning$
  - $If\ L = \sum \quad \Rightarrow \sum_{x_i x_j} (x_i - x_j)^{\top} M (x_i - x_j)$
  - $if\ L = I, M = v^{\top} v \quad \Rightarrow \|v\|_2 \Rightarrow D - ranking\ vector\ machine$
  - $if\ L = M, \Rightarrow tr(MM) \Rightarrow \|M\|_{Fro}^2 \Rightarrow Pair - wise\ SVM$
    - $\|M\|_{Fro} = \sqrt{\sum_i \sum_j |m_{ij}|^2} = \sqrt{tr(MM)}$

Extended Reading:

Kaizhu Huang, Yiming Ying, and Colin Campbell. 2009. GSML: A Unified Framework for Sparse Metric Learning. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining* (ICDM '09). IEEE Computer Society, Washington, DC, USA, 189-198.

Seminar on the property of Metric

What is the dimension of Metric M?

$$d \times d$$

Seminar on the property of Metric

Metric M must be symmetric or can be represented as a nonsymmetric form?

$$M = \frac{M + M^{\top}}{2} + \frac{M - M^{\top}}{2}$$

Seminar on the property of Metric

Metric M must positive semi-definite (PSD)?

$$if\ \ x^{\top}Mx \leq 0$$
$$x = y - z \Leftrightarrow ||y - z||_M \leq ||y - y||_M$$

Seminar on the property of Metric

Metric M must meet the triangular inequality?

$$\|x-y\|_M \leq \|x-z\|_M + \|z-y\|_M$$
$$If\ \ y = x?$$

What property should distance preserve?

$$d(x_1, x_2) = d(x_2, x_1)$$

$$d(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2$$

$$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$$

What is the relation between Metric learning and support vector machine?

DML

$$min_M \sum_i y_i(\|(x_{i1} - x_{i2})A\|^2 - b_i) + log|M|$$

SVM

$$min_{w,b,\alpha} \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \, [y_i(w_i x_i - b_i) - 1]$$

$$\Leftrightarrow \quad \max_{\alpha} -\frac{1}{2}\sum_{i,j} y_i y_j K(x_i, x_j)\alpha_i \alpha_j + \sum_j \alpha_j$$

$$s.t. \sum_i \alpha_i y_i = 0, \qquad 0 \le \alpha_i \le C, \forall i$$

$$K(x_i, x_j) = \left|\left|x_i - x_j\right|\right|^2 - c$$

What is the relation between Metric learning and Embedding?

DML: pair-wise constraints, preserve the supervised side info

$$M^* = argmin_M \sum_{(x_{i1}, x_{i2}) \in S} \|Ax_{i1} - Ax_{i2}\|^2$$

Embeding: preserve the geodesic distance between samples

$$W^* = argmin_W \sum_i \left\| x_i - \sum_{x_j \in N(x_i)} W_j x_j \right\|^2$$

Exploration of Distance Function Learning (BDFL)
(Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering (NIPS09))

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$

Drawbacks: M is a dxd matrix

$$d(x_1, x_2) = \varphi(x_1) - \varphi(x_2) - (x_1 - x_2)^\top \nabla\varphi(x_2) \quad \textbf{(Original)}$$

$$d(x_1, x_2) = \left(\nabla\varphi(x_1) - \nabla\varphi(x_2)\right)^\top (x_1 - x_2) \quad \textbf{(Modified)}$$

$$d(x_1, x_2) = (x_1 - x_2)^\top \nabla^2\varphi(\hat{x})(x_1 - x_2) \quad \textbf{(Intermediate Value Theorem)}$$

$$A \sim \nabla^2\varphi(\hat{x})$$

Advantage:

1. $\varphi(\hat{x})$ is a function $\mathbb{R}^d \to \mathbb{R}$, and $\nabla\varphi(x_1)$ is a vector rather than a matrix

$$\mathcal{O}(dxd) \to \mathcal{O}(d)$$

2. $\varphi(\hat{x})$ is local sensitive. Hessian matrix of convex function $\nabla^2\varphi(\hat{x})$ depends on the location of x_1 and x_2

Any Problem?

Exploration of Distance Function Learning

(Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering (NIPS09))

$$d(x_1,x_2)=\sqrt{(x_1-x_2)^T M(x_1-x_2)} \leftarrow d(x_1,x_2)=(x_1-x_2)^{\top}\nabla^2\varphi(\hat{x})(x_1-x_2)$$

Property of Mahalanobis distance

$$d(x_1,x_2)=d(x_2,x_1) \quad \surd$$

$$d(x_1,x_2)=0 \Leftrightarrow x_1=x_2 \quad \surd$$

$$d(x_1,x_3)\le d(x_1,x_2)+d(x_2,x_3) \quad \surd$$

Property of Bregman distance function $\nabla^2\varphi(\hat{x})$

$$d(x_1,x_2)=d(x_2,x_1) \quad \surd$$

$$d(x_1,x_2)=0 \Leftrightarrow x_1=x_2 \quad \surd$$

$$d(x_1,x_3)\le d(x_1,x_2)+d(x_2,x_3) \quad ?$$

Let Ω be the closed domain for x. If $\exists m, M \in \mathbb{R}, M > m > 0$, and

$$mI \preccurlyeq min_{x\in\Omega}\nabla^2\varphi(x) \preccurlyeq max_{x\in\Omega}\nabla^2\varphi(x) \preccurlyeq MI$$

where I is identity matrix, we have the following inequality

$$\sqrt{d(x_a,x_b)} \le \sqrt{d(x_a,x_c)}+\sqrt{d(x_c,x_b)}+\left(\sqrt{M}-\sqrt{m}\right)\left(d(x_a,x_c)d(x_c,x_b)\right)^{\frac{1}{4}}$$

Let Ω be the closed domain for x. If $\exists m, M \in \mathbb{R}, M > m > 0$, and
$$mI \preccurlyeq min_{x\in\Omega}\nabla^2\varphi(x) \preccurlyeq max_{x\in\Omega}\nabla^2\varphi(x) \preccurlyeq MI$$
where I is identity matrix, we have the following inequality
$$\sqrt{d(x_a,x_b)} \leq \sqrt{d(x_a,x_c)} + \sqrt{d(x_c,x_b)} + (\sqrt{M}-\sqrt{m})\left(d(x_a,x_c)d(x_c,x_b)\right)^{\frac{1}{4}}$$

Proof. First, let us denote by f as follows:
$$f = (\sqrt{M}-\sqrt{m})[d(x_a,x_c)d(x_c,x_b)]^{1/4}$$
The square of the right side of Eq. (2) is
$$(\sqrt{d(x_a,x_c)} + \sqrt{d(x_c,x_b)} + f^{1/4})^2 = d(x_a,x_b) - \eta(x_a,x_b,x_c) + \delta(x_a,x_b,x_c)$$
where
$$\begin{aligned}\delta(x_a,x_b,x_c) &= f^2 + 2f\sqrt{d(x_a,x_c)} + 2f\sqrt{d(x_c,x_b)} + 2\sqrt{d(x_a,x_c)d(x_c,x_b)}\\ \eta(x_a,x_b,x_c) &= (\nabla\varphi(x_a) - \nabla\varphi(x_c))(x_c - x_b) + (\nabla\varphi(x_c) - \nabla\varphi(x_b))(x_a - x_c).\end{aligned}$$
From this above equation, the proposition holds if and only if $\delta(x_a,x_b,x_c) - \eta(x_a,x_b,x_c) \geq 0$.
From the fact that
$$\begin{aligned}&\delta(x_a,x_b,x_c) - \eta(x_a,x_b,x_c)\\ = &\frac{(\sqrt{M}-\sqrt{m})^2 + 2(\sqrt{M}-\sqrt{m})\left(d(x_a,x_c)^{\frac{3}{4}}d(x_c,x_b)^{\frac{1}{4}} + d(x_c,x_b)^{\frac{3}{4}}d(x_a,x_c)^{\frac{1}{4}}\right) + 2d(x_a,x_c)d(x_c,x_b)}{\sqrt{d(x_a,x_c)d(x_c,x_b)}}\end{aligned}$$
since $\sqrt{M} > \sqrt{m}$ and the distance function $d(\cdot) \geq 0$, we get $\delta(x_a,x_b,x_c) - \eta(x_a,x_b,x_c) \geq 0$. $\square$

Any other problem?

Exploration of Distance Function Learning
(Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering (NIPS09))

$$\boldsymbol{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \nabla^2 \boldsymbol{\varphi}(\widehat{\boldsymbol{x}})(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

$$\min_{\varphi, b} \frac{1}{2} |\varphi|^2_{\mathcal{H}_k} + C \sum_{i=1}^{n} \ell(y_i[d(x_i^1, x_i^2) - b])$$

How to solve?

$$\varphi(x) = \int dy \kappa(x, y) q(y) = \int dy\ exp(x^\top y) q(y)$$

$$= \boxed{\int du\ exp(x^\top X u) q(u)}$$

$$Assume\ q(y) = \sum_i \alpha_i \delta(y - x_i)$$

$$Min_{\alpha \in \mathbb{R}^N_+, b} \mathcal{L} = \frac{1}{2} \alpha^\top K \alpha + C \sum_i \ell(y_i[z_i^\top \alpha - b])$$

$$z_i = [exp(x_i^1) - exp(x_i^2)] \circ [X^\top (x_i^1 - x_i^2)]$$

Extended Reading:

Lei Wu, Rong Jin, Steven C.H. Hoi, Jianke Zhu, Nenghai Yu, "Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering", *Advances in Neural Information Processing Systems* ***(NIPS'09)***, *2009*.

The function that minimizes

$$\min_{\varphi,b} \frac{1}{2}|\varphi|^2_{\mathcal{H}_k} + C\sum_{i=1}^{n} \ell(y_i[d(x_i^1, x_i^2) - b])$$

Admits the following form:

$$\varphi(x) \in \mathcal{H}_{\parallel} = \int_{y\in A} dy q(y) h(x^\top y) = \int du\, h(x^\top X u) q(u)$$

Where $u \in R^N$, and $X = (x_1, \cdots, x_N)$.

Proof. We write $\varphi(x) = \varphi_{\parallel}(x) + \varphi_{\perp}(x)$ where

$$\varphi_{\parallel}(x) \in \mathcal{H}_{\parallel} = \int_{y\in\mathcal{A}} dyq(y)h(x^\top y),\ \varphi_{\perp}(x) \in \mathcal{H}_{\perp} = \int_{y\in\bar{\mathcal{A}}} dyq(y)h(x^\top y)$$

Thus, the distance function defined in (1) is then expressed as

$$\begin{aligned} d(x_a, x_b) &= (x_a - x_b)^\top \left(\nabla\varphi_{\parallel}(x_a) - \nabla\varphi_{\parallel}(x_b)\right) + (x_a - x_b)^\top \left(\nabla\varphi_{\perp}(x_a) - \nabla\varphi_{\perp}(x_b)\right) \\ &= \int_{y\in\mathcal{A}} q(y)(h'(x_a^\top y) - h'(x_b^\top y))y^\top (x_a - x_b) + \int_{y\in\bar{\mathcal{A}}} q(y)(h'(x_a^\top y) - h'(x_b^\top y))y^\top (x_a - x_b) \\ &= \int_{y\in\mathcal{A}} q(y)(h'(x_a^\top y) - h'(x_b^\top y))y^\top (x_a - x_b) = (x_a - x_b)^\top \left(\nabla\varphi_{\parallel}(x_a) - \nabla\varphi_{\parallel}(x_b)\right) \end{aligned}$$

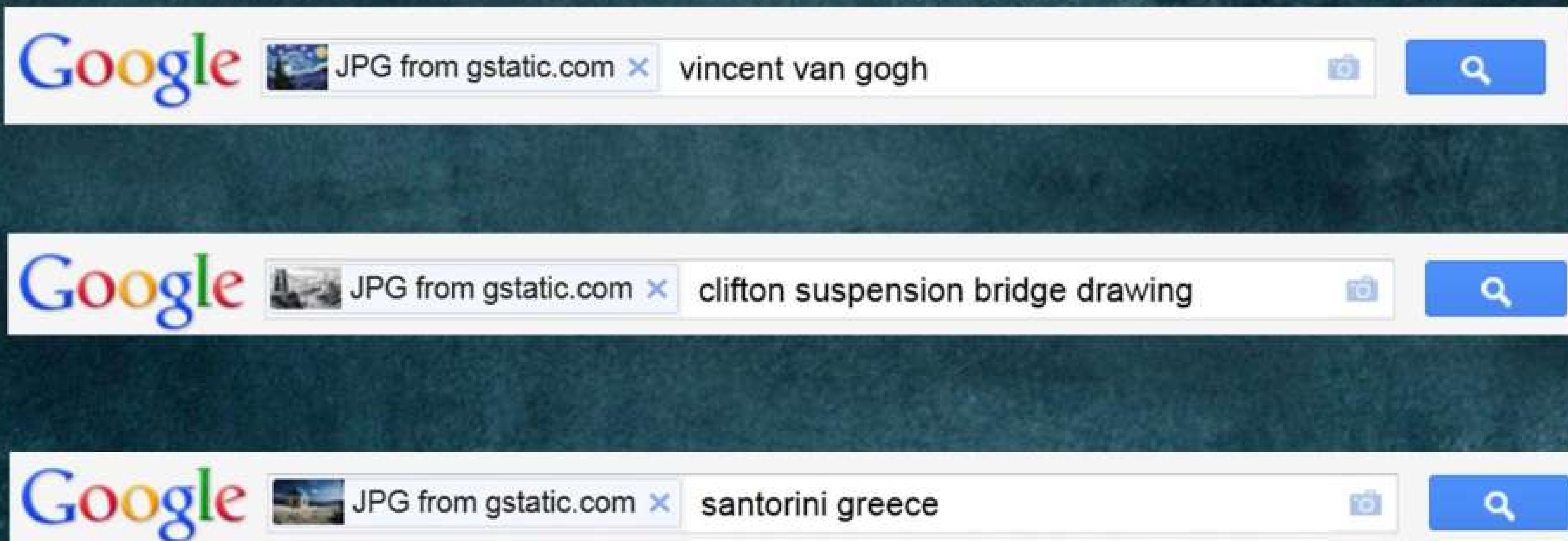
Since $|\varphi(x)|^2_{\mathcal{H}_\kappa} = |\varphi_{\parallel}(x)|^2_{\mathcal{H}_\kappa} + |\varphi_{\perp}(x)|^2_{\mathcal{H}_\kappa}$, the minimizer of (1) should have $|\varphi_{\perp}(x)|^2_{\mathcal{H}_\kappa} = 0$. Since $|\varphi_{\perp}(x)| = \langle\varphi_{\perp}(\cdot), \kappa(x,\cdot)\rangle_{\mathcal{H}_\kappa} \le |\kappa(x,\cdot)|_{\mathcal{H}_\kappa}|\varphi_{\perp}|_{\mathcal{H}_\kappa} = 0$,, we have $\varphi_{\perp}(x) = 0$ for any x. We thus have $\varphi(x) = \varphi_{\parallel}(x)$, which leads to the result in the theorem. □

Distance Metric Learning II

Application of DML

Lei Wu
University of Pittsburgh

Distance measurement is important.

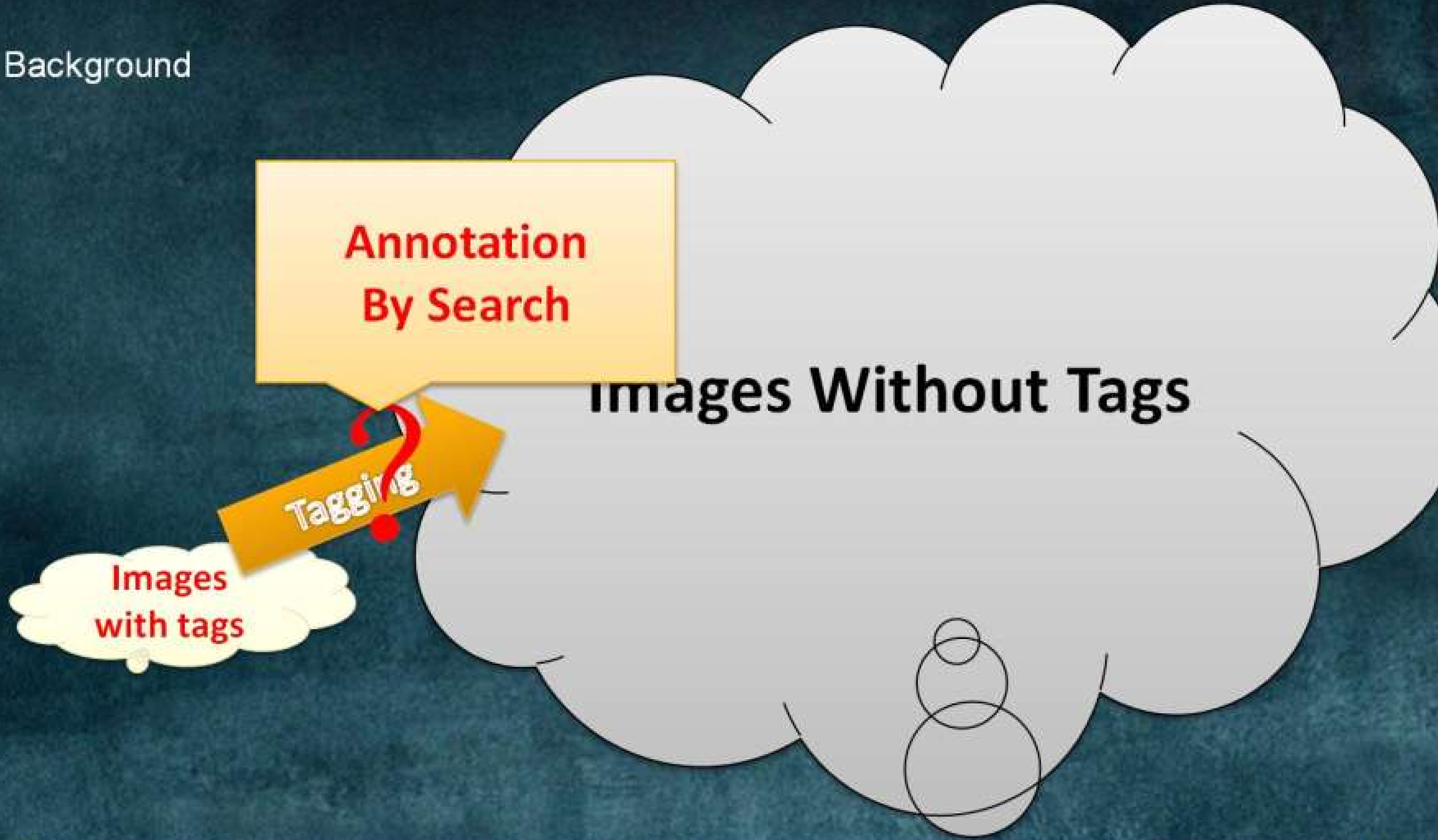


Background

- Annotation/tagging is essential to making images accessible to Web users
- Social media data in social websites enjoy rich tagging information provided by Web users

flickr facebook Picasa YouTube Broadcast Yourself

Background



Taken the tagged images as knowledge, is it possible to automatically tag the billions of images?

Motivation

- Annotation by Search (Wang et al. 2006)
  - resolve the challenge of ***auto-photo annotation*** by leveraging the emerging huge amount of **rich image surrounding text**

Main problems which limit the Annotation by Search
- **Web noise**
- **Semantic gap**

Sun
Bird
Sky
Blue
…

Bird
Fly
White
Cloud
…

Sun
Cloud
Hawk
Fly
…

Eagle
…

Motivation

$$d_M(x_i, x_j) = \sqrt{((x_i - x_j)^\top \boxed{M} (x_i - x_j))}$$

- Distance Metric Learning
  - Learning to optimize the metric ***M***
  - Side Information (a.k.s. "Pairwise Constraints")
    - Similar pairs $S(x_1, x_2)$: x_1 and x_2 belong to the same category
    - Dissimilar pairs $D(x_1, x_2)$: x_1 and x_2 belong to different categories

Motivation

- Certain side information
  - Generated by humans
  - Noise free
  - Hard constraints: similar=1; dissimilar=0
-
  How about learning a better soft constraints automatically from uncertain info of the Web?
  - Small-scale
  - Inaccurate

Motivation

Certain Side Info

Pros:
- Simple
- Easy to Adopt

Cons:
- Manual
- Expensive

Uncertain Side Info

Pros:
- Learn from Web
- Large amount

Cons:
- Complicated
- Noisy

Author： Lei
Tags:
Sun, Bird
Sky, Blue
…

Author: Lei
Tags:
Bird， Fly
White， Cloud
…

Motivation

- Annotation by Search from Social Media
  - NO *explicit* **pairwise side information** available
  - But rich information is available with social images
- Ideas of our research
  - To discover *implicit pairwise relationship* between social images via a probabilistic approach
  - To learn effective **distance metrics from uncertain side information** that is discovered from social images implicitly

Probabilistic Relevance Component Analysis (pRCA)

- The objective function of pRCA:

Minimize Sum of square distances of examples from their chunklet's centers

$$\min_{M\succeq 0,\mu,P} \quad \boxed{\sum_{i=1}^{n}\sum_{k=1}^{m} p_i^{(k)}\|x_i-\mu_k\|_M^2} - \boxed{\lambda\log|M|}$$

$$s.t. \quad \|P-P_0\|_F^2 \le \gamma,$$

$$\sum_{k} p_i^{(k)} = 1,$$

regularization preventing the trivial solution

Corollary 1. *When fixing the means of chunklets* $\boldsymbol{\mu}$ *and the matrix of probability assignments* $\boldsymbol{P}$ *(assuming with hard assignments of 0 and 1), the Probabilistic Relevance Component Analyasis (**pRCA**) formulation reduces to the regular RCA learning.*

Time Cost for Metric Learning

- To evaluate the time efficiency performance of the proposed DML algorithm on the same dataset

Table 1: Time cost of different DML methods.

(s)	baseline	RCA	DCA	ITML
Time	N/A	731.63	865.58	1185.27
(s)	LMNN	NCA	RDML	pRCA
Time	1673.23	28989.78	824.81	891.15

- Findings
  - *The most efficient method is the regular RCA approach*
  - *The most time-consuming one is NCA*
  - *pRCA is quite competitive, which is worse than RCA,DCA, and RDML, but is considerably better than ITML, LMNN, and NCA*

Some Good Examples

Query Photo	Top Recommended Tags
	autumn, fall, forest , trees, nature , tree wood, germany , path , creative
	sunset, clouds, sky, sea, beach, abigfave, sun, water, landscape, ocean
	tiger, zoo , specanimal, impressedbeauty, abigfave, nature, animal , cat, animals, aplusphoto
	garden, flowers, yellow, nature, hdr, nikon, spring, festival, impressed beauty

Some Poor Examples

Query Photo	Top Recommended Tags
	macro, nikon, bokeh, nature, flower, canon, storm, eos, plane, flickrsbest
	nikon, street, water, sport, blue, bike, lebanon, kids, eric mckenna, krissy mckenna
	winter, photography, art , beach usa, fashion, portrait , travel, party, snow
	park, river, travel, trees, lake, hiking, winter, green, vacation, water

Conclusions

- Distance metric learning (DML) is very useful tool in solving distance based applications
- There are quite a lot of interesting research problems in DML
- A little step on DML will make great impact to many applications
- This lecture is only an introduction. More details please refer to the reference papers
- Think hard and maybe a great idea will come out to change the world

References

- J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. NIPS*, 2005.
- E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In NIPS2002, 2002
- A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. International Conference on Machine Learning*, 2003.
- Steven C. H. Hoi, Wei Liu, Michael R. Lyu, and Wei-Ying Ma. 2006. Learning Distance Metrics with Contextual Constraints for Image Retrieval. (CVPR '06), 2006.
- Lei Wu**,** Steven C.H. Hoi, Rong Jin, Jianke Zhu, Nenghai Yu, "Distance Metric Learning from Uncertain Side Information with Application to Automated Photo Tagging", *ACM International Conference on Multimedia (MM'09)*, 2009.
- Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res*. 10 (June 2009), 207-244.
- Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning* (ICML '07),2007.
- Lei Wu, Rong Jin, Steven C.H. Hoi, Jianke Zhu, Nenghai Yu, "Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering", *Advances in Neural Information Processing Systems* ***(NIPS'09),*** *2009.*

Thanks

leiwu@cs.pitt.edu