电子科技大学
University of Electronic Science and Technology of China

Supervised Metric Learning

Liu

Data Mining Lab, Big Data Research Center, UESTC
Email： junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Overview

➢1. Introduction

- Definition
- Application

➢2.Comparing

- Similarity measure
- Dimensionality reduction

➢Supervised metric learning algorithms

- LDA
- MMDA
- ITML
- RCA
- NCA

What is metric learning

A **Distance Metric if it satisfies the following four properties:**

- Nonnegativity: $D(x, y) \geq 0$
- Coincidence: $D(x, y) = 0$ *if and only if* $x = y$
- Symmetry: $D(x, y) = D(y, x)$
- Subadditivity: $D(x, y) + D(y, z) \geq D(x, z)$

Where $D : X \times X \rightarrow R$ *and* X *represents a set of data points*

What is metric learning

数据挖掘实验室 Data Mining Lab

- Two important distance metric:

1. *Euclidean distance, which measures the distance between* ***x and y by***

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top (\mathbf{x}-\mathbf{y})}$$

2. *Mahalanobis distance, which measures the distance between* ***x and y by***

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{S} (\mathbf{x}-\mathbf{y})}$$

where S is the inverse of the data covariance matrix

What is metric learning

Generalized Mahalanobis distance

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{M} (\mathbf{x}-\mathbf{y})}$$

where M is some arbitrary Symmetric Positive Semi-Definite (SPSD) matrix.

We can decompose M as M = UΛU′ and let W = UΛ^1/2

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{W}\mathbf{W}^\top (\mathbf{x}-\mathbf{y})} = \sqrt{(\mathbf{W}^\top(\mathbf{x}-\mathbf{y}))^\top (\mathbf{W}^\top(\mathbf{x}-\mathbf{y}))}$$
$$= \sqrt{(\tilde{\mathbf{x}}-\tilde{\mathbf{y}})^\top (\tilde{\mathbf{x}}-\tilde{\mathbf{y}})}$$

where $\boldsymbol{x} = \boldsymbol{Wx}$

What is metric learning

Distance Metric Learning:

The problem of learning a mapping function f (projection matrix ***W****), such that $f(x)$ and $f(y)$ will be in the Euclidean space and $D(x, y) = \|f(x) - f(y)\|$, where $\|\cdot\|$ is the l2 norm.*

Application

- Clustering
  - Similarity measure in K-means
- Classification
  - KNN
- Image retrieval

Comparing

- Similarity measure

  -- similarity measures are in some sense the inverse of distance measure

$$s(x, y) = -||x - y||_2^2$$

- Dimensionality reduction

  -- most of the existing metric learning approaches can be viewed as a standard Euclidean distance in some embedding space.

Supervised distance learning algorithms

Local	Global
NCA Goldberger et al. (2004), ANMM Wang and Zhang (2007), LMNN Weinberger et al. (2005)	LDA Fukunaga (1990), LSI Xing etal. (2002), ITML Davis et al.(2007), MMDA Kocsor et al.(2004), RCA Shental et al. (2002)

LDA defines the compactness matrix and scatterness matrix as

$$\mathbf{\Sigma}_C = \frac{1}{C}\sum_{c}\frac{1}{n_c}\sum_{\mathbf{x}_i \in c}(\mathbf{x}_i - \bar{\mathbf{x}}_c)(\mathbf{x}_i - \bar{\mathbf{x}}_c)^\top$$

$$\mathbf{\Sigma}_S = \frac{1}{C}\sum_{c}(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^\top$$

The goal of LDA is to find a W which can be obtained ***by solving the following*** *optimization problem*

$$\min_{\mathbf{W}^\top\mathbf{W}=\mathbf{I}} \frac{tr(\mathbf{W}^\top \mathbf{\Sigma}_C \mathbf{W})}{tr(\mathbf{W}^\top \mathbf{\Sigma}_S \mathbf{W})}$$

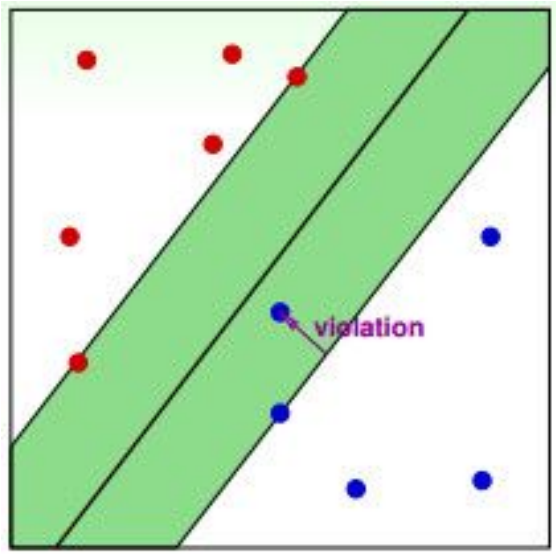
The goal of LDA is to find a W which can be obtained ***by solving the following*** *optimization problem*

$$\min_{\mathbf{W}^\top \mathbf{W}=\mathbf{I}} \frac{tr(\mathbf{W}^\top \mathbf{\Sigma}_{\mathcal{C}} \mathbf{W})}{tr(\mathbf{W}^\top \mathbf{\Sigma}_{\mathcal{S}} \mathbf{W})}$$

The learned distance between ***x_i and x_j*** *is the Euclidean distance between* ***Wx_i and Wx_j,*** *and the computational technique involved is eigenvalue decomposition.*

Motivation:

Intuitively what MMDA does is to find k ***orthogonal*** *projection hyperplanes such that on each projection direction the two data clusters are separated as well as possible,where* **W = [*w*1,*w*2, · · · ,*wk*].**



Imagine k orthogonal hyperolanes using soft SVM

Margin maximizing discriminant analysis

The goal of MMDA is to find a W which can be obtained ***by solving the following*** *optimization problem*

$$\min_{\mathbf{W},\mathbf{b},\xi_r\geq 0} \frac{1}{2}\sum_{r=1}^{d}\|\mathbf{w}_r\|^2 + \frac{C}{n}\sum_{r=1}^{d}\sum_{i=1}^{n}\xi_{ri}$$

$$s.t.\ \forall i = 1,\ldots,n, \quad r = 1,\ldots,d$$

$$l_i\left((\mathbf{w}^r)^T\mathbf{x}_i + b\right) \geq 1 - \xi_{ri},$$

$$\mathbf{W}^T\mathbf{W} = \mathbf{I}$$

The learned distance between $\mathbf{x}_i$ ***and*** $\mathbf{x}_j$ *is the Euclidean distance between* $\mathbf{W}\mathbf{x}_i$ ***and*** $\mathbf{W}\mathbf{x}_j$, *and the computational technique involved is* ***eigenvalue decomposition*** *and* ***quadratic programming***.

Information theoretic metric learning

- Assumption
  - 1.Distance between point pairs in must-link set is less than u
  - 2.Distance between point pairs in cannot-link set is larger than l
  - 3.There exists priori metric matrix M_0 (if sample set satisfies Gaussian distribution,using covariance matrix parameterize M_0).

Information theoretic metric learning

ITML solves the following optimization problem

$$\min_{M \succeq 0} KL(p(\boldsymbol{x}; M_0) \| p(\boldsymbol{x}; M))$$

$$\text{s.t.} \quad d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) \le u, \quad (\boldsymbol{x}_i, \boldsymbol{x}_j) \in S$$

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) \ge l, \quad (\boldsymbol{x}_i, \boldsymbol{x}_j) \in D$$

The learned distance metric is the Mahalanobis distance with precision matrix **M.**

Information theoretic metric learning

数据挖掘实验室

Data Mining Lab

If M's and M0's distribution have the same mean value

$$KL(p(\boldsymbol{x}; M_0)\|p(\boldsymbol{x}; M)) = \frac{1}{2} D_{ld}(M_0^{-1}, M^{-1})$$

$$D_{ld}(M, M_0) = \mathrm{tr}(MM_0^{-1}) - \log\det(MM_0^{-1}) - d$$

ITML solves the following optimization problem

$$\begin{aligned} \min_{M \succeq 0} \; & D_{ld}(M, M_0) \\ \text{s.t.} \; & \mathrm{tr}(M(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}) \leq u, \; (\boldsymbol{x}_i, \boldsymbol{x}_j) \in S \\ & \mathrm{tr}(M(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}) \geq l, \; (\boldsymbol{x}_i, \boldsymbol{x}_j) \in D \end{aligned}$$

Information theoretic metric learning

ITML solves the following optimization problem

$$\begin{aligned}\min_{M \succeq 0} \; & D_{ld}(M, M_0) \\ \text{s.t.} \; & \operatorname{tr}(M(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}) \leq u, \; (\boldsymbol{x}_i, \boldsymbol{x}_j) \in S \\ & \operatorname{tr}(M(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}}) \geq l, \; (\boldsymbol{x}_i, \boldsymbol{x}_j) \in D\end{aligned}$$

Using an efficient Bregman projection approach to solve problem

$$M_{t+1} = M_t + \beta M_t(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} M_t$$

Relevant component analysis

*The goal of RCA is to find a transformation that amplifies **relevant variability** and suppresses **irrelevant variability**.*

Relevant component analysis

Three steps of RCA:

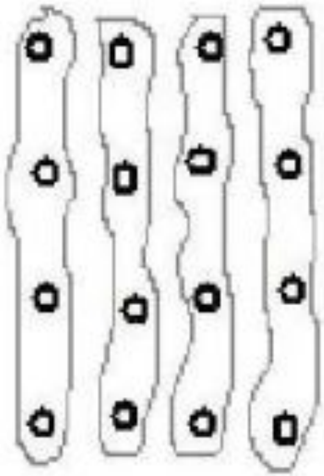
– Construct chunklets according to equivalence (must-link) constraints,

–RCA computes the following weighted within-chunklet covariance matrix:

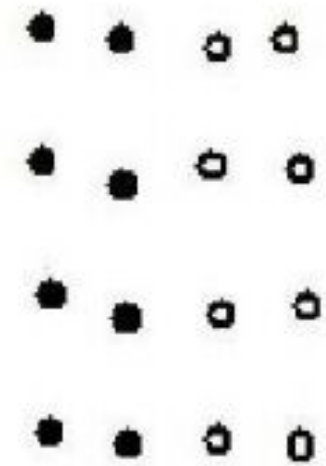
$$\mathbf{C} = \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{x}_{ji} - \bar{\mathbf{m}}_j)(\mathbf{x}_{ji} - \bar{\mathbf{m}}_j)^{\top}$$

– Compute the whitening transformation ***W = C^1/2****, and apply it to the original data points: ˜****x*** **=** ***Wx****.*

Difference between labeled data and chunklet data



Chunklet Data



Labeled Data:
Supervised Learing

Neighborhood component analysis

数据挖掘实验室
Data Mining Lab

- each point ***xi selects another point xj as its neighbor*** with some probability *pi j*

$$p_{ij} = \frac{\exp\left(-\|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_k\|^2\right)}$$

- NCA computes the probability that point *i will* be correctly classified

$$p_i = \sum p_{ij}$$

where $\mathcal{L}_i = \{j | l_i = l_j\})$

Neighborhood component analysis

数据挖掘实验室
Data Mining Lab

- The objective NCA maximizes is the expected number of points correctly classified

$$\mathcal{J}(\mathrm{W}) = \sum_i p_i = \sum_i \sum_{j \in \mathcal{L}_i} p_{ij}$$

The learned distance between ***xi and xj*** is the Euclidean distance between**W**x*i* ***and Wxj.****The* computational technique involved is **eigenvalue decomposition**.

Thanks