实验1. 度量学习实验报告

MG1733092，张则君，826320663@qq.com

2017年11月5日

综述

距离度量作为一种求解向量之间距离的方法,在分类、聚类等问题上应用广泛。传统的距离度量（如欧式距离）模式固定,无法通过不断学习得到一个合适的距离度量,使得传统距离度量往往不能揭示样本间属性的深层关系,造成其应用局限性很大。为了改进这一问题,涌现出了很多距离度量的新方法。根据对样本标记信息的利用程度不同,距离度量可以分为全局距离度量和局部距离度量。全局距离度量需要满足全部样本的成对约束,而局部距离度量仅仅在局部满足这些成对约束。

本文在有监督学习的实验下,对传统方法采用欧式距离,对全局距离度量选取了Relevant Component Analysis(RCA)算法,对局部距离度量选取了Neighbourhood Components Analysis(NCA)算法。实验首先在moons 数据集上测试了这三种算法的基本功能,然后通过字母识别实验对比了这三种算法。

任务1

设：在R^D中有n个向量$x_1, ..., x_n$,其对应标记为$y_1, ..., y_n$.

对于任意样本x_i, x_j,其距离公式为

$$d(x_i, x_j) = (x_i - x_j)M(x_i - x_j)^T$$

其中M为半正定对称矩阵,存在正交基P使$M = PP^T$.

RCA

RCA通过对数据全局的线性转换给相关特征大的权重，给不相关特征低的权重,使得数据的内在结构更容易的获得。每个chunklet是若干向量已知属于同一类的集合，通过chunklets可以学习到相关的特征。

本实验由于是有监督学习，协方差矩阵是已知变量，免去了调参；同时本实验通过RCA学习得到全秩马氏距离阵，因此度量学习目标可以直接学习得到，使得在优化算法中无需使用拉格朗日乘子法。

度量函数学习目标

设m_j为第j个chunklet的向量均值，C为类内协方差矩阵的和

$$m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji} \tag{1}$$

$$C = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - m_j)(x_{ji} - m_j)^T \tag{2}$$

则RCA度量学习目标为：

$$M = C^{-1} \tag{3}$$

NCA

NCA近邻成分分析通过随机选择近邻，通过优化留一法（LLO）的交叉检验结果求得马氏距离的度量矩阵。

度量函数学习目标

在NCA中，以留一法（LLO）正确率为最大化目标，设p_{ij}为样本x_j对x_i的分类影响概率，A_i为所有样本对x_i的分类影响概率，整个样本集的留一法正确率为$f(P)$

$$p_{ij} = \frac{exp(-||x_i - x_j||_M^2)}{\sum_l exp(-||x_i - x_l||_M^2)}, \quad p_{ii} = 0 \tag{4}$$

$$p_i = \sum_{j\in\Omega_i} p_{ij} \tag{5}$$

则NCA度量函数学习目标为

$$max\ f(P) = \sum_{i=1}^{n}\sum_{j\in\Omega_i} p_{ij} = \sum_{i=1}^{n} p_i \tag{6}$$

优化算法

采用随机梯度下降法求解，对f(P)中P求导得

$$\frac{\partial f}{\partial P} = 2P\sum_i(p_i\sum_k p_{ik}x_{ik}x_{ik}^T - \sum_{j\in\Omega_i} p_{ij}x_{ij}x_{ij}^T) \tag{7}$$

为了优化公式(7) 改进度量函数学习目标为

$$max\ g(P) = \sum_{i=1}^{n} log(\sum_{j\in\Omega_i} p_{ij}) = \sum_{i=1}^{n} log(p_i) \tag{8}$$

对g(P)中P求导得

$$\frac{\partial g}{\partial P} = 2P\sum_i(\sum_k p_{ik}x_{ik}x_{ik}^T - \frac{\sum_{j\in\Omega_i} p_{ij}x_{ij}x_{ij}^T}{\sum_{j\in\Omega_i} p_{ij}}) \tag{9}$$

其中$x_{ij} = x_i - x_j$,$x_{ik} = x_i - x_k$

根据公式(9)对每一个样本进行更新，更新公式为：

$$P = P + lr \cdot \frac{\partial g}{\partial P}^{i} = P + lr \cdot 2P(\sum_{k} p_{ik} x_{ik} x_{ik}^{T} - \frac{\sum_{j \in \Omega_i} p_{ij} x_{ij} x_{ij}^{T}}{\sum_{j \in \Omega_i} p_{ij}}) \tag{10}$$

其中lr为学习率

任务2

实验方法

在字母识别的实验中，采用任务1的NCA算法与RCA算法.

RCA

RCA中chuklets数为数据样本类别的数目，每个chuklet的大小为对应已知类别的集合大小.

NCA

NCA中采用随机梯度算法，需要调节学习率。这里我们对训练集采用10折交叉验证法，对学习率在{0.1,0.01,0.001,0.0001,0.00001}之间选取得到最合适的学习率0.0001.

实验结果

表一反映了三种算法的实验结果。实验通过重复30次取错误率的平均值作为评估算法的指标，并且计算了错误率的标准差。从实验结果来看NCA的算法效果优于欧式距离和RCA算法，NCA的错误率不仅低而且结果也更加稳定。

表 1: 三种距离度量算法结果对比

	KNN(1)	KNN(3)	KNN(5)
欧式距离	0.166880 ± 0.012082	0.206282 ± 0.013218	0.223248 ± 0.013466
RCA	0.132692 ± 0.011003	0.157137 ± 0.009946	0.167735 ± 0.00993
NCA	0.099658 ± 0.008061	0.121026 ± 0.009147	0.135812 ± 0.009203

代码说明

在不修改test_myDML.py前提下为了实现RCA和NCA两种算法，在my_DML.py代码增加了两个类RCA_Supervised和NCA，同时增加train_nca,train_rca函数作为nca,rca的训练函数；为了调节NCA的学习率增加了tune_nca函数，采用10折交叉验证法选择合适的学习率。本实验参考了metric-learn的库包。

对于RCA算法,借鉴了其良好的编程结构，对rca代码进行了大面积删除冗余。

对于NCA算法,借鉴了其良好的编程结构，并针对公式(10)对nca代码进行修改，使之与公式一致。

参考文献

[1] Liu Yang.An overview of Distance Metric Learning.

[2] Aharon Bar-Hillel,Tomer Hertz,Noam Shental,Daphna Weinshall.Learning Distance Functions using Equivance Relations.

[3] Jacob Goldberger,Sam Roweis,Geoff Hinton,Ruslan Salakhutdinov.Neighbourhood Components Analysis.

[4] Liu Yang.Distance Metric Learning:A comprehensive Survey.

[5] De-Chuan Zhan.Research on Distance Metric Learning Approaches.

[6] https://all-umass.github.io/metric-learn.

[7] http://blog.csdn.net/chlele0105/article/details/13006443

[8] http://www.cnblogs.com/rcfeng/p/3958926.html