Assignment (I)

- Linear Discriminant Analysis (LDA) and Neighborhood component Analysis (NCA) are two widely used methods for dimensionality reduction. Please compare them with PCA and answer what are their rationales for data reduction.

数据挖掘作业一

810594956@qq.com
171860607
白晋斌

目录

一、LDA

1.LDA 思想

线性判别分析(LDA, Linear Discriminant Analysis)是一种监督学习的分类和降维的方法,但更多是被用来降维.LDA 的原理是让投影后同一类中数据的投影点之间尽可能地靠近,而类不同类别中数据的类别中心之间的距离尽可能远,用一句话概括就是"投影后类内方差最小,类间方差最大".

2.二类 LDA 原理

假设我们的数据集 D={(x_1,y_1),(x_2,y_2),...,((x_m,y_m))},其中任意样本x_i为 n 维向量,$y_i \in \{0,1\}$.我们定义N_j (j=0,1)为第 j 类样本的个数, X_j (j=0,1)为第 j 类样本的集合，而$\mu_j = \frac{1}{N_j}\sum_{x \in X_j} x$ (j = 0,1)为第 j 类样本的均值向量,定义$\sum_j = \sum_{x \in X_j}(x - \mu_j)(x - \mu_j)^T$ (j = 0,1)为第 j 类样本的协方差矩阵.

由于是两类数据,因此我们只需要将数据投影到一条直线上即可.假设我们的投影直线是向量 w,则对任意一个样本x_i,它在直线 w 的投影为$w^T x_i$,对于我们的两个类别中心点μ_0, μ_1,点在直线 w 的投影为$w^T\mu_0, w^T\mu_1$.由于 LDA 需要让不同类别的数据的类别中心之间的距离尽可能的大, 也就是我们要最大化 $||w^T\mu_0 - w^T\mu_1||_2^2$,同时我们希望同一种类别数据的投影点尽可能的接近,也就是要同类样本投影点的协方差$w^T \sum_0 w$, $w^T \sum_1 w$尽可能的小,即最小化$w^T \sum_0 w + w^T \sum_1 w$.

综上所述,我们的优化目标是

$$\underbrace{arg\ max}_{w}\ J(w) = \frac{||w^T\mu_0 - w^T\mu_1||_2^2}{w^T \sum_0 w + w^T \sum_1 w} = \frac{w^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w}{w^T(\sum_0 + \sum_1)w}$$

我们一般定义类内散度矩阵S_w为

$$S_w = \sum_0 + \sum_1 = \sum_{x \in X_0}(x - \mu_0)(x - \mu_0)^T + \sum_{x \in X_1}(x - \mu_1)(x - \mu_1)^T$$

同时定义类间散度矩阵S_b为

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$$

这样我们的优化目标重写为

$$\underbrace{arg\ max}_{w}\ J(w) = \frac{w^T S_b w}{w^T S_w w}$$

利用广义瑞利商的性质,我们知道 J(w')的最大值为矩阵$S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}}$的最大特征值,而对应的 w'为$S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}}$最大特征值对应的特征向量.而$S_w^{-1} S_b$的特征值和$S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}}$的特征值相同, $S_w^{-1} S_b$的特征向量和$S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}}$的特征向量 w'满足$w = S_w^{-\frac{1}{2}} w'$的关系.

注意到对于二类的时候, $S_b w$的方向恒平行于$\mu_0 - \mu_1$,不妨令$S_b w = \lambda(\mu_0 - \mu_1)$, 将其带入$(S_w^{-1} S_b)w = \lambda w$,可以得到$w = S_w^{-1}(\mu_0 - \mu_1)$,也就是说我们只要求出原始二类样本的均值和方差就可以确定最佳的投影方向 w 了.

3.多类 LDA 原理

假设我们的数据集 D={(x_1,y_1),(x_2,y_2),...,((x_m,y_m))},其中任意样本x_i为 n 维向量,y_i∈{C_1, C_2,..., C_k}.我们定义N_j (j=1,2...k)为第 j 类样本的个数, X_j (j=1,2...k)为第 j 类样本的集合，而μ_j (j=1,2...k)为第 j 类样本的均值向量,定义$\sum_j$ (j = 1,2...k)为第 j 类样本的协方差矩阵.

由于我们是多类向低维投影,则此时投影到的低维空间就不是一条直线,而是一个超平面了.假设我们投影到的低维空间的维度为 d,对应的基向量为(w_1, w_2, ..., w_d), 基向量组成的矩阵为 W, 它是一个 n×d 的矩阵.

此时我们的优化的目标应该可以变成

$$\frac{W^T S_b W}{W^T S_w W}$$

其中,$S_b = \sum_{j=1}^{k} N_j \ (\mu_j - \mu)(\mu_j - \mu)^T$, μ为所有样本的均值向量.$S_w = \sum_{j=1}^{k} S_{wj} = \sum_{j=1}^{k} \sum_{x \in X_j} (x - \mu_0)(x - \mu_0)^T$

此时, $W^T S_b W$和$W^T S_w W$都是矩阵,不是标量,无法作为一个标量函数来优化,故我们要用其他的一些替代优化目标来实现.

常见的一个 LDA 多类优化目标函数定义为

$$\underbrace{arg\ max}_{w}\ J(w) = \frac{\prod_{diag} w^T S_b w}{\prod_{diag} w^T S_w w}$$

其中, $\prod_{diag} A$为 A 的主对角线元素的乘积,W 为 n x d 的矩阵.

$J(w)$的优化过程可以转化为

$$J(w) = \frac{\prod_{i=1}^{d} w_i^T S_b w_i}{\prod_{i=1}^{d} w_i^T S_w w_i} = \prod_{i=1}^{d} \frac{w_i^T S_b w_i}{w_i^T S_w w_i}$$

右式即为广义瑞利商,最大值是矩阵$S_w^{-1}S_b$的最大特征值,最大的 d 个值的乘积就是矩阵$S_w^{-1}S_b$的最大的 d 个特征值的乘积,此时对应的矩阵 W 为这最大的 d 歌特征值对应特征向量组成的矩阵.

由于 W 是一个利用样本的类别得到的投影矩阵,因此它的降维到的维度 d 的最大值为 k-1.因为S_b中的每个$\mu_j - \mu$的秩为 1,因此协方差矩阵相加后的最大的秩为 k(矩阵的秩小于等于各个相加矩阵的秩的和),但是由于如果我们知道前 k-1 个μ_j后,最后一个μ_k可以由前 k-1 个μ_j线性表示,因此S_b的秩最大是 k-1,即特征向量最多有 k-1 个.

4.LDA 算法流程

输入: 数据集 D={(x_1,y_1),(x_2,y_2),...,((x_m,y_m))},其中任意样本x_i为 n 维向量,y_i∈{C_1, C_2,..., C_k},降维到的维度 d.

输出:降维后的样本集 D'

算法:1) 计算类内散度矩阵S_w

2) 计算类间散度矩阵S_b

3) 计算矩阵$S_w^{-1}S_b$

4) 计算$S_w^{-1}S_b$的最大的 d 个特征值和对应的 d 个特征向量(w_1, w_2, ..., w_d), 得到投影矩阵 W

5) 对样本集中的每一个样本特征x_i,转化为新的样本$z_i = W^T x_i$

6) 得到输出样本集 D'={(z_1,y_1),(z_2,y_2),...,((z_m,y_m))}

以上是使用 LDA 进行降维的算法流程。实际上 LDA 除了可以用于降维以外，还可以用于分类。一个常见的 LDA 分类基本思想是假设各个类别的样本数据符合高斯分布，这样利用 LDA 进行投影后，可以利用极大似然估计计算各个类别投影数据的均值和方差，进而得到该类别高斯分布的概率密度函数。当一个新的样本到来后，我们可以将它投影，然后将投影后的样本特征分别带入各个类别的高斯分布概率密度函数，计算它属于这个类别的概率，最大的概率对应的类别即为预测类别。

5.LDA 算法特色

优点:

1)在降维过程中可以使用类别的先验知识经验，而像 PCA 这样的无监督学习则无法使用类别先验知识。

2)LDA 在样本分类信息依赖均值而不是方差的时候，比 PCA 之类的算法较优。

缺点:

1)LDA 不适合对非高斯分布样本进行降维，PCA 也有这个问题。

2)LDA 降维最多降到类别数 k-1 的维数，如果我们降维的维度大于 k-1，则不能使用 LDA。当然目前有一些 LDA 的进化版算法可以绕过这个问题。

3)LDA 在样本分类信息依赖方差而不是均值的时候，降维效果不好。

4)LDA 可能过度拟合数据。

6.LDA 与 PCA 的对比

6.1 相同点

1)两者均可以对数据进行降维

2)两者在降维时均使用了矩阵特征分解的思想

3)两者都假设数据符合高斯分布

6.2 不同点

1)LDA 是有监督的降维方法,而 PCA 是无监督的降维方法

2)LDA 降维最多降到类别数 k-1 的维数，而 PCA 没有这个限制

3)LDA 除了可以用于降维,还可以用于分类

4)LDA 选择分类性能最好的投影方向,而 PCA 选择样本点投影具有最大方差的方向

二、NCA

1.NCA 思想

近邻成分分析（NCA, Neighborhood Component Analysis）是由 Jacob Goldberger 和 Geoff Hinton 等人在 2005 年发表于 NIPS 上的一项工作，属于度量学习（Metric Learning）和降维（Dimension Reduction）领域。NCA 中的 Neighborhood 是指近邻，可以理解为相似的样本（这里默认距离小代表相似度高）,NCA 的原理是以马氏距离为距离度量的 KNN 为基础，通过不断优化 KNN 分类的准确率来学习转换矩阵，最终得到对原数据进行降维的转换矩阵。

2.NCA 原理

给定数据集 D={(x_1,y_1),(x_2,y_2),…,((x_m,y_m))},其中任意样本x_i为 d 维向量,y_i∈{C_1, C_2,…, C_k}.考虑在 KNN 中,使用 leave one out 交叉验证法,假设现在要预测第 i 个样本的标签,那么我们可以这样做:

1)计算样本 i 和其余所有样本之间的欧式距离 $d_{ij} = ||x_i - x_j||_2$

2)选择距离最小的 k 个

3)利用这 k 个样本的标签进行投票得到预测结果

上述过程是一般的 KNN 过程,这里引入 Stochastic 1-NN 改进方法:

1)计算样本 i 的近邻分布:$p_{ij} = \frac{e^{-||x_i - x_j||_2^2}}{\sum_{k \neq i} e^{-||x_i - x_k||_2^2}}, p_{ii} = 0$

2)根据概率分布p_{ij}, j $\neq i, j \in [1, n]$采样得到一个样本 k,然后将第 i 个数据的样本预测为y_k

从上面可以看出,第 i 个样本的真实标记为y_i,假如$y_k = y_i$,那么预测正确,记$C_i = \{j | y_j = y_i\}$表示与第 i 个样本类别一样的下标集合,那么利用上述 Stochastic 1-NN 正确预测第 i 个样本标签的概率为: $p_i = \sum_{j \in C_i} p_{ij}$

那么,对于所有的样本,优化目标为$f = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \sum_{\in C_i} p_{ij}$

鉴于使用欧式距离计算会导致计算量特别大,并且维度空间特别高,这里再引入度量学习的思想,引入可学习的马氏距离.

3.马氏距离和度量学习

3.1 马氏距离

3.1.1 定义

数据样本矩阵表示为X $= [x_1; x_2; \ldots; x_n]^T$,这是以样本角度表示的,还可以以特征角度表示为X $= [f_1; f_2; \ldots; f_d]^T$,假设样本间的协方差矩阵为S $\in R^{d \times d}$,那么有

$$S_{ij} = Cov(i, j) = \frac{1}{n}(f_i - mean(f_i))^T (f_i - mean(f_i))$$

马氏距离的定义为

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

3.1.2 马氏距离的优点

1) 相当于是对数据中心化和标准化,数据中心化和标准化后的马氏距离与原来的马氏距离一致

2)由于是相当于中心化和标准化,所以不受特征单位的影响

3)可以推导出可学习的马氏距离,是度量学习的基础

4)考虑样本总体特性,一般来说,两个样本放入不同的总体,计算得到的马氏距离不相等

3.2 度量学习

度量学习是基于可学习的马氏距离,也称伪(pseudo)马氏距离:

$$d_{\mathrm{M}}(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)}, M \in S_+^d$$

其中 M 是半正定矩阵(Positive Semi-Definite, PSD),是可以学习的参数,称为度量. 度量学习的目标就是通过一些约束(比如,must-link pair 和 must-not-link pair)进行优化矩阵 M ,从而学习到一个距离度量.

由于 M 是半正定矩阵,那么存在$A \in R^{k \times d}, k < d, k \geq rank(M)$,满足$M = A^T A$,那么

$$d_{\mathrm{M}}(x_i, x_j) = \sqrt{(x_i - x_j)^T A^T A (x_i - x_j)} = ||Ax_i - Ax_j||_2$$

可以看出马氏距离做的事情是先把数据降维到低维空间,然后再求欧氏距离.

4.NCA 完整表达

1)令$A \in R^{k \times d}$为参数

2)计算样本 i 的近邻分布:$p_{ij} = \frac{e^{-||Ax_i - Ax_j||_2^2}}{\sum_{k \neq i} e^{-||Ax_i - Ax_k||_2^2}}, p_{ii} = 0$

3)优化目标: $f(A) = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} \sum_{\in C_i} p_{ij}$

4)NCA 求解:$f(A)$对 A 求偏导,得到

$$\frac{\partial f}{\partial A} = 2A \sum_i (p_i \sum_{k \neq i} p_{ik}(x_i - x_k)(x_i - x_k)^T - \sum_{j \in C_i} p_{ij}(x_i - x_j)(x_i - x_j)^T)$$

求出梯度之后,利用梯度下降法优化即可得到 NCA 的训练结果.训练得到的映射矩阵 A 可以用来对数据进行降维,以便于降低数据维度.

5.NCA 算法特色

1)NCA 是有监督的,需要提供样本标签

2)借鉴度量学习引入距离度量 A 来计算样本间距离

3)目标是使得 Stochastic 1-NN 的准确率最高

4)参数学习过程使用了 Leave one out 交叉验证的方法

5)NCA 以及其他度量学习的复杂度至少是N^2,可能没办法在大数据上应用

6.NCA 与 PCA 的对比

6.1 相同点

1)两者均可以对数据进行降维

2)两者对降维的维数均没有限制

6.2 不同点

1)NCA 是有监督的降维方法,而 PCA 是无监督的降维方法

2)NCA 对数据分布没有假设,而 PCA 要求数据服从高斯分布

3)NCA 除了降维还是一种度量学习的方法

4)NCA 基于 KNN 选择分类性能最好的投影方向,而 PCA 选择样本点投影具有最大方差的方向