山东大学 计算机科学与技术 学院

可视化技术 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：201900130176 | 姓名： 李伟国 | | 班级： 智能 |
| 实验题目： | | | |
| 实验学时： | | 实验日期： | |
| 实验目的： | | | |
| 硬件环境：  处理器：AMD Ryzen 5 3600 6-Core Processor 3.60 GHz  Ram 16.0 GB | | | |
| 软件环境： | | | |
| 实验步骤与内容：   1. 体验tensorflow project   同样的数据集：    将其方法，然后用鼠标可以看见每一个样本如下。    Google 这个交互式的网站，让数据仿佛活了起来  2) 使用matlab toolbox 比较t-sne ,pca,isomap 等方法的区别  使用的实验数据入下：      有9个维度，其中第一个是性别。       * t-sne：非线性的降维技术（non-linear）   t 是指 t-分布（也叫学生分布），stochastic Neighbor Embedding，叫做随机邻居嵌入。此方法相当的适合用来做高维度的数据可视化。该技术可以通过Barnes-but 来近似的实现，这样的速度比较快。  t-sne ：相似的目标通过其附近的样本点来建模而不相似的目标通过远的目标点来建模，有很高的概率。所以t-sne对原始空间中的数据集进行相似性的建模和降维后的空降中的相似性建模都是概率密度。  如果两个点在原始的空间中比较接近，那么他们在降为后的空间中任然是比较接近的。    t-distributed：设随机变量X和Y相互独立，并且X~N(0,1),Y~则称随机变量 是服从自由度为n的t分布；  接下来，对从网站上下载的数据集进行t-sne降维，在运用t-sne 之前，人们总是先对数据进行PCA降维处理，再进行tsne降维。  Tsne 有一种能够聚类的效果  下面看一下该方法降维后的效果    显而易见，我们可以从图中发现一些拥挤的感觉  随着迭代次数的增加，error一直在减少，但是并不是线性的减少。     * PCA   PCA计算的核心就是计算一个convariance matrix ，并对其进行特征分解（奇异值分解），并按照特征值从大到小的顺序进行选择其对应的特征向量，特征向量的方向就是主成分的方向，然后根据情况选择前k个最大的特征向量作为方向，此时就可以把高维的数据降到了K维，在这里，K取3或者2，这样就可以将我们的高维数据在低维度的空间中可视化出来。    降到2维    降到3维度     * Isomap 一种无监督的算法（unsupervised）   这是一种非线性的降维方式，是一种嵌入的方式。该算法提供了一种简单的方法，用于基于流形上的每个数据点的邻居的粗略估计来估计数据流形的内在几何形状。  这是一种对数据种类和数目很丰富的有效的可行的方法。其能够使用非线性的方式去降维并能保留局部的结构  Isomap 使用嵌入的邻域图的geodesic distance 而不是欧式距离。  给定数据集，经过最近邻等方式构造一个数据图(data graph，nieghborhood graph)。而后，**计算任意两个点之间的最短路径（即测地距离）**。对于全部的任意两个点对，指望在低维空间中保持其测地距离。  算法步骤如下：   1. 决定每一个点的邻居：K nearest neighbors   任然用欧氏距离来确定每个point的最近的k个邻居，而非最近的k个邻居的距离设置为无穷大   1. 构建邻域图（neighborhood graph）   如果数据点和数据点是邻居关系，那么就将他们相连接。如果不是就保持不连接的状态。   1. 计算两个点的最短路径（Dijkstra 或者 floyd warshall）   这一步也叫做在两个点之间通过邻居图  求geodisic distance，   1. 计算低维空间的嵌入（MDS）   因为现在任意两个点之间的额距离已经知道了，MDS尽可能的保持任意两个点之间的距离在embedding space 和 original space 不变    经过isomap 降维后的数据 | | | |
| 结论分析与体会：   1. iosmap （等距特征映射）可以保留一些数据点之间的非线性关系，isomap 也经常用于NLP分析中去。让两个点之间的距离近似的等于依次多个临近点的连线的长度之和 2. PCA 实际上式一个线性变换，将原始的数据 变换到一个新的坐标系中，使得任何数据的投影的方差在一个坐标上最大，在二个坐标上第二大，etc。该方法减少数据的维数的同时保持数据集的对方差贡献最大的特征。 3. 不同于ISOMAP 中距离不变的思想，而是现将欧式距离转变成条件概率，来表示点与点的相似度，在优化loss function（KL 散度），从而保证点一点之间的分布概率不变   并且在低维空间先使用更重长尾的t-distributed 来避免拥挤的问题 | | | |

附录：程序源代码