智能软件开发方向基础

第10章 非监督式特征降维与低维可视化 --PCA与t-SNE

张朝晖

2022~2023学年第二学期

序号	内容
1	概述
2	机器学习的基本概念
3	模型的选择与性能评价
4	数据的获取、探索与准备
5	近邻模型------分类、回归
6	决策树模型------分类、回归
7	集成学习------分类、回归
8	(朴素)贝叶斯模型------分类
9	聚类
10	特征降维及低维可视化(PCA, t-SNE)
11	总复习

河北师范大学软件学院
Software College of Hebei Normal University

主要内容

1 主成分分析法原理

Principal Component Analysis: PCA

问题1. 如何学习主成分提取模型?

问题2. 如何提取主成分?

如何基于主成分确定样本的降维表示?

问题3. 如何由r(r<p)个主成分，重构x?

PCA应用----人脸识别中的特征提取

2. t-SNE

河北师范大学软件学院
Software College of Hebei Normal University

主成分分析实质:

借助正交线性变换，将一组观测数据由可能相关的特征描述转化由一系列互不相关的特征(主成分)描述。

主成分分析的目的:

- 特征提取
- 降维
- ......

河北师范大学软件学院
Software College of Hebei Normal University

问题描述:

设 (1)原始特征空间特征向量$\boldsymbol{x}=\left[x_1,\cdots,x_p\right]^T$,

其中 $x_1,\cdots,x_p$统计相关

由于相关性，使得各特征存在信息冗余

(2)原始特征空间经**正交**线性变换$\boldsymbol{A}$，得新特征空间

新特征$\xi_i, i=1,\cdots,p$无信息冗余

$$\xi=\left[\xi_1,\cdots,\xi_p\right]^T=\boldsymbol{A}^T\boldsymbol{x}=\left[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_p\right]^T\boldsymbol{x}$$

其中 $\xi_i=\sum_{j=1}^{p}a_{ij}x_j=\boldsymbol{a}_i^T\boldsymbol{x}$

目标: 寻求**最优正交变换A**，得到若干重要新特征(**大方差**)。

$\boldsymbol{A}=\left[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_p\right]$，其中 $\boldsymbol{a}_i^T\boldsymbol{a}_j=\delta_{ij}=\begin{cases}1 & i=j\\0 & i\neq j\end{cases}$

河北师范大学软件学院
Software College of Hebei Normal University

各个变换特征(主成分)的确定原则

(1)ξ_i与ξ_j互不相关$(i,j=1,\ldots,p$并且$i\neq j)$

(2)ξ_1是原始特征线性组合中的方差最大者(最重要特征).

(3)ξ_2是与ξ_1不相关的原始特征线性组合中的方差最大者;
ξ_2对原始数据中不能被ξ_1解释的剩余部分，拥有最大解释能力.

(4)ξ_3是与ξ_1,ξ_2均不相关的原始特征线性组合中的方差最大者;ξ_3 对于原始数据中不能被 ξ_1,ξ_2解释的剩余部分，具有最大解释能力.....

(5)$\xi_1,\xi_2,\cdots,\xi_p$分别称为关于原始特征向量$\boldsymbol{x}$的第一,第二,...,第$\boldsymbol{p}$个主成分。

河北师范大学软件学院
Software College of Hebei Normal University

STEP1.确定a_1，得到第一主成分ξ_1

考虑新特征 $\xi_1=\sum_{j=1}^{p} a_{1j}x_j = a_1^T x = x^T a_1$

各样本关于新特征ξ_1的方差

$\mathrm{var}(\xi_1)=\boxed{E\left[\left(\xi_1-\hat{\xi}_1\right)^2\right]=E\left[\left(a_1^T x-a_1^T\hat{x}\right)\left(x^T a_1-\hat{x}^T a_1\right)\right]}$

$=E\left[a_1^T\left(x-\hat{x}\right)\left(x-\hat{x}\right)^T a_1\right]$ 数学期望的线性性质

$=a_1^T E\left[\left(x-\hat{x}\right)\left(x-\hat{x}\right)^T\right]a_1$

$=a_1^T\Sigma a_1$

最优a_1应满足$\begin{cases}\max\limits_{a_1} & a_1^T\Sigma a_1\\ s.t. & a_1^T a_1=1\end{cases}$

构造*Lagrange*目标函数

$\begin{cases}f(a_1,v)=a_1^T\Sigma a_1-v\left(a_1^T a_1-1\right)\\ v\text{为}Lagrange\text{乘子}\end{cases}$

$$\frac{\partial f}{\partial a_1}=\begin{bmatrix}\frac{\partial f}{\partial a_{11}}\\ \cdot\\ \frac{\partial f}{\partial a_{1p}}\end{bmatrix}=2\left(\Sigma a_1-va_1\right)=0$$

$\Rightarrow\ \Sigma a_1=va_1$

a_1为协方差矩阵Σ的**本征值**v对应**本征列向量**

河北师范大学软件学院 Software College of Hebei Normal University

$\mathrm{var}(\xi_1)=a_1^T\Sigma a_1=va_1^T a_1=v$

要使$\mathrm{var}(\xi_1)$最大：

v应为Σ最大本征值λ_1

a_1为Σ最大本征值λ_1对应的本征列向量

称ξ_1为观测样本x的**第一主成分**

河北师范大学软件学院 Software College of Hebei Normal University

STEP2.确定a_2，得到第二主成分ξ_2

a_2应满足两个要求：

$\begin{cases}\textbf{A.}\text{新特征}\xi_2\text{与第一主成分}\xi_1\text{不相关}\\ \textbf{B.}\text{除去}\xi_1\text{外，新特征}\xi_2\text{方差最大}\end{cases}$

其中 $\xi_2=a_2^T x=x^T a_2=\sum_{j=1}^{p} a_{2j}x_j$

河北师范大学软件学院 Software College of Hebei Normal University

$\boxed{\textbf{A. }\xi_2\text{与}\xi_1\text{不相关}}$

ξ_2与ξ_1之间的协方差 $E\left[\left(\xi_1-\hat{\xi}_1\right)\left(\xi_2-\hat{\xi}_2\right)\right]=0$

$E\left[\left(\xi_1-\hat{\xi}_1\right)\left(\xi_2-\hat{\xi}_2\right)\right]=E\left[a_1^T\left(x-\hat{x}\right)\left(x-\hat{x}\right)^T a_2\right]$

$=a_1^T E\left[\left(x-\hat{x}\right)\left(x-\hat{x}\right)^T\right]a_2=a_1^T\Sigma a_2=a_2^T\Sigma a_1=0$

由于$\Sigma a_1=\lambda_1 a_1$，所以$a_2^T\Sigma a_1=\lambda_1 a_2^T a_1=0$。

即：$\boxed{\xi_2\text{与}\xi_1\text{不相关}\Leftrightarrow a_2,a_1\text{正交}}$

河北师范大学软件学院 Software College of Hebei Normal University

$\boxed{B.\ \text{除去}\xi_1\text{外，}\xi_2\text{方差最大}}$

$\mathrm{var}(\xi_2)=E\left[\left(\xi_2-\hat{\xi}_2\right)^2\right]=E\left[a_2^T\left(x-\hat{x}\right)\left(x-\hat{x}\right)^T a_2\right]$

$=a_2^T\Sigma a_2$

最优a_2应满足 $\begin{cases}\max\limits_{a_2} & a_2^T\Sigma a_2\\ s.t. & a_2^T a_2=1,\ a_1^T a_2=0\end{cases}$

构造*Lagrange*目标函数

$f(a_2,v_2,\mu)=a_2^T\Sigma a_2-v_2\left(a_2^T a_2-1\right)-\mu a_1^T a_2$

河北师范大学软件学院 Software College of Hebei Normal University

$Lagrange$目标函数

$$f\left(a_2, \nu_2, \mu\right) = a_2^T \Sigma a_2 - \nu_2\left(a_2^T a_2 - 1\right) - \mu a_1^T a_2$$

$$\frac{\partial f}{\partial a_2} = 2\Sigma a_2 - 2\nu_2 a_2 - \mu a_1 = 0$$

两边左乘a_1^T，有：

$$2a_1^T \Sigma a_2 - \boxed{2\nu_2 a_1^T a_2} - \mu a_1^T a_1 = 2a_1^T \Sigma a_2 - \boxed{0} - \mu = 0$$

又 $a_1^T \Sigma a_2 = a_2^T \Sigma a_1 = \lambda_1 a_2^T a_1 = 0$

所以 $\mu = 0$

$$\Sigma a_2 - \nu_2 a_2 = 0$$

$$\Sigma a_2 - \nu_2 a_2 = 0$$

显然ν_2为Σ的**本征列向量a_2**对应的**本征值**

所以 $\mathrm{var}\left(\xi_2\right) = a_2^T \Sigma a_2 = \nu_2 a_2^T a_2 = \nu_2$

要使 $\mathrm{var}\left(\xi_2\right) = \nu_2$最大

$$\begin{cases} \nu_2\text{应为}\Sigma\text{第二大本征值}\lambda_2 \\ a_2\text{必为}\Sigma\text{剩余本征值中最大本征值}\lambda_2\text{对应的本征向量} \\ \xi_2\text{为观测}x\text{的}\textbf{第二主成分} \end{cases}$$

STEP3.确定其它$a_i, i = 3, \ldots, p$,及正交变换矩阵A

样本协方差矩阵Σ共有p个本征值，将其排列为

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$$

基于上述本征值对应本征向量$a_i, i = 1, \ldots, p$,可以构造**所有主成分**$\xi_i, i = 1, \ldots, p$

各个主成分的方差满足 $\sum_{i=1}^{p} \mathrm{var}\left(\xi_i\right) = \sum_{i=1}^{p} \lambda_i$

并且有 $\xi = \left[\xi_1, \cdots, \xi_p\right]^T = A^T x = \left[a_1, \cdots, a_p\right]^T x$

河北师范大学软件学院
Software College of Hebei Normal University

结论：

(1)线性变换矩阵$A = \left[a_1 \ \cdots \ a_p\right]$的各个列向量由协方差矩阵$\Sigma$的正交归一本征向量组成，$A^T = A^{-1}$.

--A为正交矩阵，$\xi = A^T x$，$x = A\xi$

(2)主成分方差满足 $\mathrm{var}\left(\xi_i\right) = \lambda_i$ $\quad \sum_{i=1}^{p} \mathrm{var}\left(\xi_i\right) = \sum_{i=1}^{p} \lambda_i$

(3)基于前k个主成分可描述数据信息比例

$\rightarrow$前k个主成分的**累积方差解释比** $d = \dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$

河北师范大学软件学院
Software College of Hebei Normal University

通常求取的是**"零均值化"的主成分**ξ

$$\begin{cases} \xi = A^T\left(x - \mu\right) \\ x = A\xi + \mu \end{cases}$$

河北师范大学软件学院
Software College of Hebei Normal University

基于主成分分析的特征降维

输入：p维特征空间的观测样本集$D = \left\{x_1, \ldots, x_N\right\}$，累积方差解释比$\alpha$

输出：线性特征提取模型，任意观测样本x的特征提取结果

A. 基于p维特征空间的观测样本集$D = \left\{x_1, \ldots, x_N\right\}$，估计**样本中心**$\mu$及**协方差矩阵**$\Sigma$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \Sigma = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)\left(x_i - \mu\right)^T$$

B. 确定Σ的p个**本征值**及**本征向量**，

得 p个本征值 $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$

对应本征向量 $a_i, i = 1, \ldots, p$

C. 确定主成分数目r--即：在$\dfrac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \ge \alpha$条件下$k$的最小值

D. 确定$p \times r$的变换矩阵$A_r = \left[a_1 \ a_2 \ \cdots \ a_r\right]$

E. 提取任意观测样本x的前r个主成分：$\xi_r = A_r^T\left(x - \mu\right)$

河北师范大学软件学院
Software College of Hebei Normal University

方差解释比、
累积方差解释比

主要内容

1 主成分分析法原理

Principal Component Analysis: PCA

问题1. 如何确定主成分?

问题2. 如何基于主成分确定样本的降维表示?

问题3. 如何由r(r<p)个主成分，重构x?

PCA应用----人脸识别中的特征提取

河北师范大学软件学院 Software College of Hebei Normal University

人脸识别的一般概念：给定一个场景中的静态图像或视频，利用给定的人脸数据库信息，鉴别或确认场景中的一位或多位人身份的过程。

一个完整的人脸识别系统通常包括三个部分：

（1）图像获取

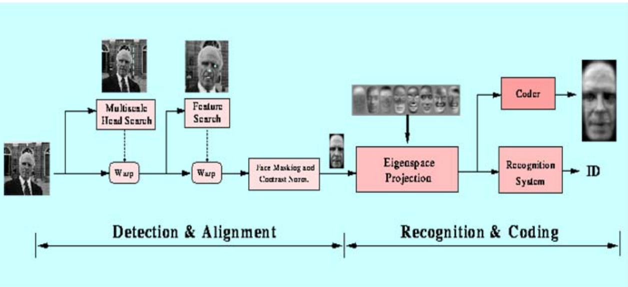
（2）人脸检测与分割

（3）人脸识别（特征选择与提取、模式匹配）

河北师范大学软件学院 Software College of Hebei Normal University

例：MIT人脸识别系统流程图

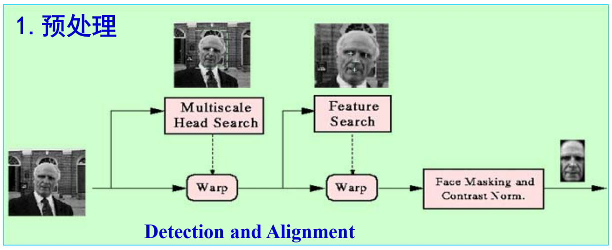
http://vismod.media.mit.edu/vismod/demos/facerec/



河北师范大学软件学院 Software College of Hebei Normal University

---

1. 预处理



(1)Original Input Image; (2)Estimated Head Location & Scale; (3)Head-Centered Image; (4)Estimated Facial Feature Locations; (5) Warped & Masked Facial Region

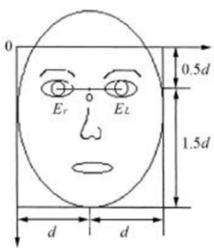
---

预处理阶段：

- 人脸图像的分割与脸部主要器官定位：

  人脸正面图像；

  左右两眼中心位置



- 图像归一化与裁剪

图像几何校准：

图像旋转，使两眼中心连线水平；

图像裁剪：2d×2d

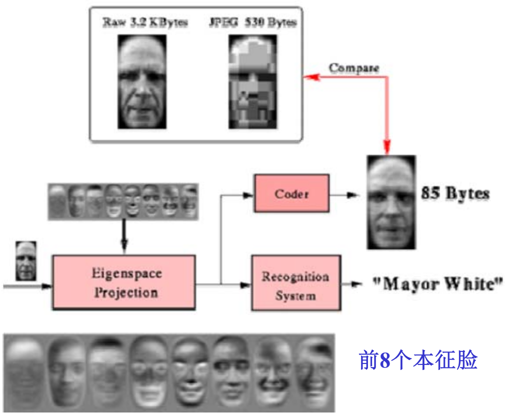
图像放缩，使其大小统一，如：2d×2d ➔ 128×128

灰度拉伸：直方图修正，消除光照影响

预处理阶段：

归一化部分人脸图像

2. 特征提取：本征脸的提取和表示



前8个本征脸

新的特征空间的获取、特征提取：本征脸的提取和表示

- 训练样本集：

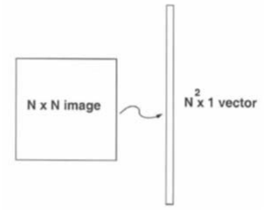
假定进行身份注册的人脸库中含有p个人，共计m幅图像。

每一幅图像都是归一化的标准图像。

所有图像构成训练样本集：

$x_i \in R^{N^2} \quad i = 1, \ldots, m$

m ——标准图像的数量



- PCA

特征提取：投影子空间的样本描述

$$y = U^T(x - \mu)$$

思考：特征提取后，如何做进一步的身份识别或鉴别？

基于压缩信息的人脸近似重构

$$x^* = \sum_{j=1}^{k} y_j u_j + \mu = Uy + \mu$$

思考：

如何基于PCA实现观测样本集的低维可视化？

如何提取观测样本的前m个主成分？

河北师范大学软件学院 Software College of Hebei Normal University

主要内容

1 主成分分析法原理

Principal Component Analysis: PCA

问题1. 如何确定主成分提取模型？

问题2. 如何提取主成分？

如何基于主成分确定样本的降维表示？

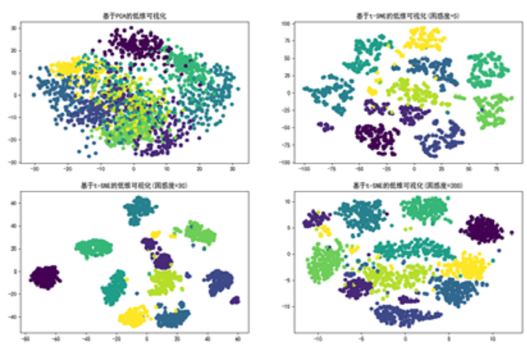
问题3. 如何由r(r<p)个主成分，重构x？

PCA应用----人脸识别中的特征提取

2. 基于t-SNE的特征降维与低维可视化

河北师范大学软件学院 Software College of Hebei Normal University

基于PCA与t-SNE的手写体数字样本集的低维可视化



- 基于PCA的低维可视化，会导致“拥堵”现象
- 基于t-SNE的低维可视化，更注重局部结构保持性

河北师范大学软件学院 Software College of Hebei Normal University

- t-SNE(t-distribution Stochastic Neighbor Embedding，t分布随机近邻嵌入法)
- 本质是基于流形学习(manifold learning)的降维方法，即寻找高维数据中可能存在的低维流形

- 在SNE方法的基础上发展而来
- 利用概率分布来度量样本间的距离，将高维空间中的欧式距离转化为条件概率密度函数来表示样本间的相似度
- 特点是能够保持样本间的局部结构，使得在高维数据中距离相近的点投影到低维中仍然相近
- 常用作样本的低维可视化分析

基于t-SNE的低维可视化要保留数据的局部结构：

- 原始空间彼此相近的样本点，降维之后距离应该也很近
- 原始空间彼此远离的样本点，降维之后距离应该也很远

- 将降维前后的原始空间、以及低维空间样本点之间“距离的远近关系”转化为相应“概率分布”

- 将t-SNE的局部结构保持性转化为寻找两种概率分布尽可能一致的低维可视化过程。

问题描述：

给定原始高维空间N个观测点$x_1,\cdots,x_N$,基于$t-SNE$寻找低维空间相应的像点$y_1,\cdots,y_N$,使变换后的点集具有较好的局部结构保持性。

实现步骤：

1. 在原始空间，针对每个样本点x_i，定义条件概率分布

$$p_{j|i} = \frac{\exp\{-\|x_i - x_j\|^2/2\delta_i^2\}}{\sum_{k\neq i}\exp\{-\|x_i - x_k\|^2/2\delta_i^2\}} \quad j\neq i,\ \ i=1,2,\dots,N$$

并定义$p_{i|i}=0$

进而，构造原始空间两个样本x_i、x_j间对称的概率分布

$$p_{ij} = \frac{p_{i|j}+p_{j|i}}{2N}$$

作为i和j之间在高维空间中的距离度量

N----总样本数

2. 在低维可视化空间，采用t-分布，度量像点之间的相似度

针对像点对y_i、y_j，定义对称概率分布

$$q_{ij} = \frac{\left(1+\|y_j - y_i\|^2\right)^{-1}}{\sum_{k\neq l}(1+\|y_k - y_l\|^2)^{-1}} \quad j\neq i, \qquad i=1,2,\dots,N$$

同样，$q_{ii}=0$

--允许高维与低维样本点之间的相似度使用不同的度量

--不同于原始高维空间，在低维空间采用t-分布度量像点之间的相似度，以缓解SNE法导致的低维可视化的“拥挤”问题

3. 构建度量两种概率分布差异的损失函数

----KL散度(Kullback-Leibler divergence)、也称相对熵

$$L(y_1,\dots,y_N) = KL(P\|\,Q) = \sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

t-SNE的最优解为：

$$\{y_1^*,\dots,y_N^*\} = argmin_{y_1,\dots,y_N}\ L(y_1,\dots,y_N)$$

注意：此处的最小KL损失，等价于最小交叉熵损失

4. 基于梯度下降法的$L(y_1,\dots,y_N)$寻优

- 首先，产生随机初始解$Y^{(0)}=\{y_1,y_2,\cdots,y_N\}$。
- 由$Y^{(0)}$可以计算得到低维可视化空间像点的初始的概率分布Q，进而求得目标函数L关于重构样本的梯度dL/dY，其中第i个梯度分量为：

$$\frac{\partial L}{\partial y_i}=4\sum_j(p_{ij}-q_{ij})(y_i-y_j)\left(1+\|y_j-y_i\|^2\right)^{-1}$$

$$i=1,2,\dots,N$$

- 更新得到新的重构样本集(采用带冲量的梯度下降法)

$$Y^{(t)}=Y^{(t-1)}+\eta(dL/dY)+\alpha(t)\left(Y^{(t-1)}-Y^{(t-2)}\right)$$

η----学习率

$\alpha(t)$----动量遗忘率$\alpha(t)$

- 算法迭代执行事先指定的T步后停止，得到最终的低维可视化结果Y

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: dataset $\mathcal{X}=\{x_1,x_2,\dots,x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)}=\{y_1,y_2,\dots,y_n\}$.
begin
compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
set $p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n}$
sample initial solution $\mathcal{Y}^{(0)}=\{y_1,y_2,\dots,y_n\}$ from $\mathcal{N}(0,10^{-4}I)$
for t=1 **to** T **do**
compute low-dimensional affinities q_{ij} (using Equation 12)
compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 13)
set $\mathcal{Y}^{(t)}=\mathcal{Y}^{(t-1)}+\eta\frac{\delta C}{\delta \mathcal{Y}}+\alpha(t)\left(\mathcal{Y}^{(t-1)}-\mathcal{Y}^{(t-2)}\right)$
end
end

Journal of Machine Learning Research 1 (2008) 1-48 Submitted 4/00; Published 10/00

Visualizing Data using t-SNE

Laurens van der Maaten L.VANDERMAATEN@MICC.UNIMAAS.NL
MICC-IKAT
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

控制参数----困惑度(perplexity)

- 嵌入概率P的取值受到方差σ_i的影响
- 定义

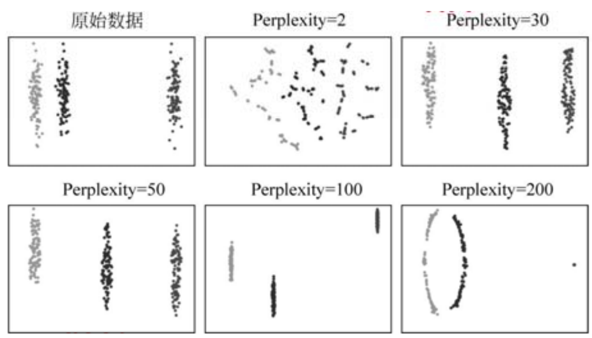
$$Perp(P_i)=2^{H(P_i)}$$

其中，$H(P_i)$是概率分布P_i的信息熵

$$H(P_i)=-\sum_{j=1}^{N}p_{ji}\log_2 p_{ji}$$

- 困惑度是控制样本点是否适合算法的主要参数。
- 推荐范围5－50。
- 困惑度应始终小于样本点数目N。
- 低困惑度，关注本地结构，并关注彼此最接近的样本点。
- 高困惑度，关注全局结构
- 原始空间维数过高时，需先结合PCA进行降维

- 困惑度大致等价于在匹配每个样本点的原始和拟合分布时，考虑的最近邻数。
- 在一些情况下，投影后的低维空间中的可视化结果受到困惑度参数的影响非常大



算法特点

- 收敛和优化情况与初值有关，不能保证收敛到全局最优解
- 在原始空间特征维数较高时，由于t分布的重尾特性，可能会使算法不能很好的保持样本间局部关系结构
- t-SNE不能将训练集上学习得到的投影方式直接用于测试集上进行降维
- 在最终可视化投影中相距较远的聚团之间的距离没有意义

应用

- 降维可视化和非监督学习，即在没有明确分类目标的样本数据中发现内在的分布规律并在低维空间中直观的展示出来