机器学习复习笔记

by 慕弋云子

目录

第一章 概述

1.什么是学习？

狭义：获得知识或技能的过程

广义：获得经验而产生的行为或行为潜能

2.什么是机器学习？

研究计算机怎样模拟或实现人类学习行为

e.g. AlphaGo：强化学习

Arthur Samuel 定义：不显式编程地赋予计算机能力的研究领域

Tom Mitchell 定义：先验知识 E，任务 T，表现度量 P，则 ML=在 T 上的表现随着 E 在 P 下得到提高

Quiz?

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

- Classifying emails as spam or not spam.
- Watching you label emails as spam or not spam.
- The number (or fraction) of emails correctly classified as spam/not spam.

此三条分别对应 T, E, P

3.与机器学习相关课程/领域？

模式识别：e.g. 数字识别（车牌）等

数据挖掘：e.g. 消费习惯分析等

计算机视觉：e.g. 特征识别、目标检测等

NLP：e.g. 翻译、对话

其他：交通出行、社交媒体、经济金融等

4.基本概念

样本、标签、训练集（+验证集）、测试集

5.机器学习分类？

田字格分类：离散/连续+监督/非监督：分类、聚类、回归、降维

6.NFL（没有免费的午餐）定理内容？

没有绝对好的算法，必须结合具体问题

7.机器学习的目标？

在欠拟合和过拟合之间寻找平衡点：降低泛化误差。

泛化误差由偏差、方差和噪声组成

第二章 线性回归

1.曲线拟合时，避免过拟合的一些方法？

加入惩罚项，对较大值的系数进行惩罚

常用正则项：$\frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$，q=0.5、1（Lasso）、2（Quadratic）、4

2.贝叶斯（Bayes）公式及其含义？

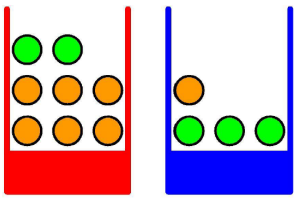
$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)}$$

含义：给出了结果事件已经发生的条件下，原因事件的条件概率。对结果的任何观测，都将增加对原因事件真正分布的知识。

亦可解读为：后验概率等于先验概率（P(B)）与似然概率（P(A|B)）的乘积除以证据因子（P(A)）。

3.贝叶斯概率练习题

● 苹果和桔子



$$P(B=r)=4/10$$
$$P(B=b)=6/10$$

1.取到苹果的概率?

2.如果取到桔子，来自红箱子的概率?

第 1 问：乘法原理

第 2 问：贝叶斯定理

红箱子取到橘子：$\frac{6}{8}\times\frac{4}{10}=\frac{3}{10}$，蓝箱子取到橘子：$\frac{1}{4}\times\frac{6}{10}=\frac{3}{20}$

Bayes：

$$P(\text{取到橘子}\mid\text{来自红箱子})=\frac{P(\text{从红箱子取橘子})}{P(\text{取到橘子})}=\frac{\frac{3}{10}}{\frac{3}{10}+\frac{3}{20}}=\frac{2}{3}$$

4.回归模型

$$y(\mathbf{x},\mathbf{w})=\sum_{j=0}^{M-1}w_j\phi_j(\mathbf{x})=\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

基函数

基函数：一次项→线性回归

此外还有多项式、高斯、Sigmoid 等基函数

5.回归的基本思想

最小化预测值 f 与样本标记值 y 的差异，以确定模型中的权重参数w^T

此差异定义为目标函数/代价函数/损失函数，需要用优化方法将损失函数最小化

6.回归模型有哪些优化方法？

梯度下降：以负梯度为搜索方向，又可分为 BGD、SGD

批处理梯度下降（BGD）：每次都利用所有数据，迭代速度很慢

随机梯度下降（SGD）：每次只用一个样本，收敛速度较快，不容易陷入局部极值，大样本数据较有效，又称在线学习

变化形式：离线学习、Mini-Batch 等

7.什么是标准方程/正规方程？

定义 $X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \cdots \\ \mathbf{x}_N^T \end{pmatrix}$，这样 $X\mathbf{w} - \mathbf{y} = \begin{pmatrix} \mathbf{X}_1^T w_1 - y_1 \\ \mathbf{X}_2^T w_2 - y_2 \\ \cdots \\ \mathbf{X}_N^T w_N - y_N \end{pmatrix}$，对应着误差的列向量形式，那么损失函数即可重写为误差列向量的内积，即：

$$J(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 = (X\mathbf{w} - \mathbf{y})^T (X\mathbf{w} - \mathbf{y})$$

关于 w 求导得到：$2X^T(X\mathbf{w} - \mathbf{y})$

令其为 0（相当于最小化损失函数），反解得到 $\hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$，即为正规方程/标准方程。

标准方程相当于是直接使用矩阵方法求解参数。

→样本量较小时选用标准方程组求解、较大时使用优化方法求解。

8.看曲线拟合的两种角度

观测数据由确定的函数加高斯噪声组成。$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$

最大似然估计：最大化似然函数→最小化 SSE(sum squares error)

贝叶斯估计：最大化后验分布→最小化 SSE+对系数的正则化

9.非参数估计的思想

本质上就是抽样的思想。要估计x_i点的密度，可以独立抽取 N 个样本，落入临近区域体积 V 的样本数为 k，则 $\hat{p}(x) = \frac{k}{NV}$（相当于频率比上区域体积）

第三章 混合模型和聚类

1.什么是聚类，有什么应用？

无监督学习、分类任务。把对象分类成不同的组别或子集合，同类对象具有相似的属性。

如医学影像中组织分类、遥感中地貌分类、降维等

2.简述 k-means 算法流程

初始化：设定要分类成的簇数目 k

（1）先初始化 k 个样本均值 $\mu_1, \ldots, \mu_k$

（2）计算样本点到均值点的距离（一般使用欧氏距离），将样本归属到距离最小的那个样本均值所在簇中

（3）重新计算此时样本均值，更新 $\mu_1, \ldots, \mu_k$

（4）重复 2-3，直至达到停机条件

优点：收敛速度快，效果较好，可解释性强

缺点：对非凸、噪声，异常点比较敏感，不易收敛

3.简述高斯混合模型（GMM）

思想是，假设数据分布是由一系列高斯分布组成的，可以通过 EM 算法求解 GMM 中的参数。

假设是由 k 个高斯分布组成。

（1）定义 k 个 0-1 示性向量，设 $z_i \quad (i = 1, \ldots, k)$ 表示第 i 个维度为 1，其余均为 0，对应地 $\pi_i \quad (i = 1, \ldots, k)$ 表示样本来自第 i 个高斯分布的概率，显然有 π_i 介于 0、1 之间且 sum 为 1。

（2）可以写出概率密度函数 $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$，其中

π_k表示来自第 k 个高斯分布，$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$表示在第 k 个高斯分布中采样出 x 的概率。这里的实质是全概率公式。

（3）写出对数似然函数，考虑约束条件$\sum_{k=1}^{K} \pi_k = 1$，使用拉格朗日乘子法，分别对 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ 求导，一阶导为零得到新的 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ 表达式。

（4）执行 EM 算法：

E 步（Expectation）：使用当前的 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ 和样本数据 $\boldsymbol{x}_n$ 计算 $\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$，表示$\boldsymbol{x}_n$属于第 k 个高斯的概率。

M 步（Maximization）：将 $\gamma(z_{nk})$ 代入（3）中极大似然估计求得的 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ 表达式，更新 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ 。

不断循环迭代 E、M 步，直至收敛。

4.GMM 的一些应用

背景建模：使用背景（前 N 帧）图像进行训练，得到 GMM 的参数；当第 N+i 帧图像中某像素不服从 GMM 时，该像素即为前景。

第四章 线性分类

1.什么是先验（A Priori）概率、后验（A Posteriori）概率

先验概率就是基于统计或是自身经验等得到的概率值，一般形式为 P(A)，如抛一枚硬币正面朝上的概率是 0.5、北航男生的比例是 0.7 等。

从先验概率可以进一步定义类条件概率密度$p(x|w_i)$即在w_i的类别中取到 x 的概率，这个取值与类的分布密度$p(x)$有关。

后验概率是基于观测事件得到的，已知了某些信息后的概率值，一般形式为 P(A|B)，如已知该同学来自计算机学院（信息 B），该同学是男生（事件 A）的概率即为后验概率。

从后验概率可以进一步定义错误概率：

$$P(e|\mathbf{x})\begin{cases}P(w_2|\mathbf{x}) \text{ if } \mathbf{x} \text{ is assigned to } w_1\\ P(w_1|\mathbf{x}) \text{ if } \mathbf{x} \text{ is assigned to } w_2\end{cases}$$

2.贝叶斯决策规则主要有哪些？

最小错误率和最小期望风险；

错误率定义为 $P(e)=\int P(e|x)p(x)dx$

期望风险定义为 $R(\alpha)=\int R(\alpha(x)|x)p(x)dx$

3.贝叶斯决策例题

假设在某个局部区域细胞中正常(w_1)和异常(w_2)两类的先验概率分别为：正常状态 $P(w_1)=0.9$，异常状态 $P(w_2)=0.1$。现有一待识别细胞，其观察值为x，从类条件概率密度曲线上查得 $p(x|w_1)=0.2$，$p(x|w_2)=0.4$。试对该细胞x进行分类。

基于最小错误率：计算后验概率 $\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$，正常细胞：0.818，异常细胞：0.182，谁大选谁→正常细胞

基于最小风险：根据题目给出的决策表：

决策表

决策 \ 损失 \ 状态	w_1	w_2
α_1	0	6
α_2	1	0

解读为决策 1（α_1），即将细胞归为正常细胞，若其本质上属于异常细胞，风险系数为 6；决策 2（α_2），即将细胞归为异常细胞，若其本质上属于正常细胞，风险系数为 1。

计算每种决策的风险，即横着看，计算后验概率乘以风险系数，然后求和即可。

此题中 R1 为 0.182*6=1.092，R2 为 0.818*1=0.818，后者更小，故归为异常细胞。

4.线性判别的思想

将样本值带入某一线性函数，若值大于 0 则分为正类，小于零则分为负类，于是问题的关键在于确定权重 w，便可将分类器设计问题转化为求准则函数极值的问

题。

常用的准则有：Fisher 准则、感知机准则、最小二乘准则等。

5.Fisher 准则的思想

最大化类间散度、最小化类内散度，找到一条最好的、最易于分类的投影线。

类内散度定义为：

$S_i = \sum_{x \in \chi_i} (x - m_i)(x - m_i)^T$，其中$m_i$为第 i 类样本均值向量，$\sum S_i$即为类内散度$S_w$

类间散度定义为：

$$S_b = (m_1 - m_2)(m_1 - m_2)^T$$

一种准则函数即为 $J_F(w) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$，进一步推导得 $J_F(w) = \frac{w^T S_b w}{w^T S_w w}$，即类间$S_b$与类内$S_w$的广义瑞利商之比。最大化该函数的 w 即为所求。

6.感知机准则的思想

随意给定判别函数初始值，通过样本分类训练过程逐步修正确定。

（1）将增广样本向量规范化（对于负类而言，标记不变，样本取反），这样解区就是所有向量正侧的交集

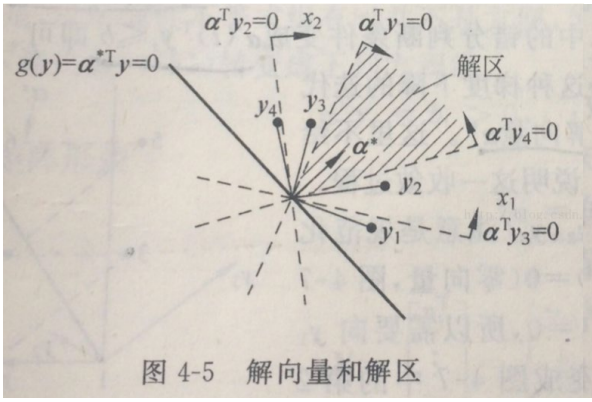


图 4-5　解向量和解区

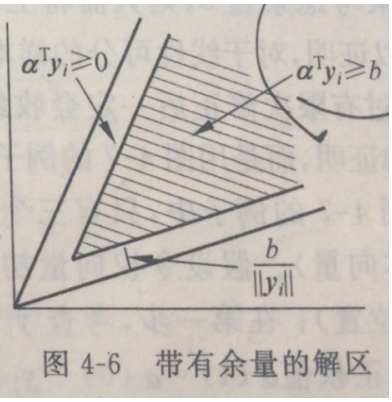


图 4-6　带有余量的解区

（2）采用梯度下降法迭代学习参数

$$a(k+1) = a(k) + \rho_k \nabla \sum_{y \in \eta^k} y$$

（可以采用 BGD、SGD 等方法，其中η^k为错

分样本集合）

（3）不断循环（2）直至η^k变为空集（无错分样本、全部向量内积为正）

7.感知机准则例题

- Class 1

$$\mathbf{x}_1=\begin{pmatrix}-2\\2\end{pmatrix},\mathbf{x}_2=\begin{pmatrix}-2\\-2\end{pmatrix}\qquad \mathbf{y}_1=\begin{pmatrix}1\\-2\\2\end{pmatrix},\mathbf{y}_2=\begin{pmatrix}1\\-2\\-2\end{pmatrix}$$

- Class 2

$$\mathbf{x}_3=\begin{pmatrix}2\\1\end{pmatrix},\mathbf{x}_4=\begin{pmatrix}2\\-1\end{pmatrix}\qquad \mathbf{y}_3=\begin{pmatrix}-1\\-2\\-1\end{pmatrix},\mathbf{y}_4=\begin{pmatrix}-1\\-2\\1\end{pmatrix}$$

- Initial weight vector

$$\mathbf{a}(1)=\begin{pmatrix}0\\2\\1\end{pmatrix}\qquad \mathbf{a}(1)^t=\begin{pmatrix}0&2&1\end{pmatrix}$$

8.最小二乘准则思想

最小化误差的平方和寻找数据的最佳函数匹配。对异常值比较敏感。

第五章 支持向量机

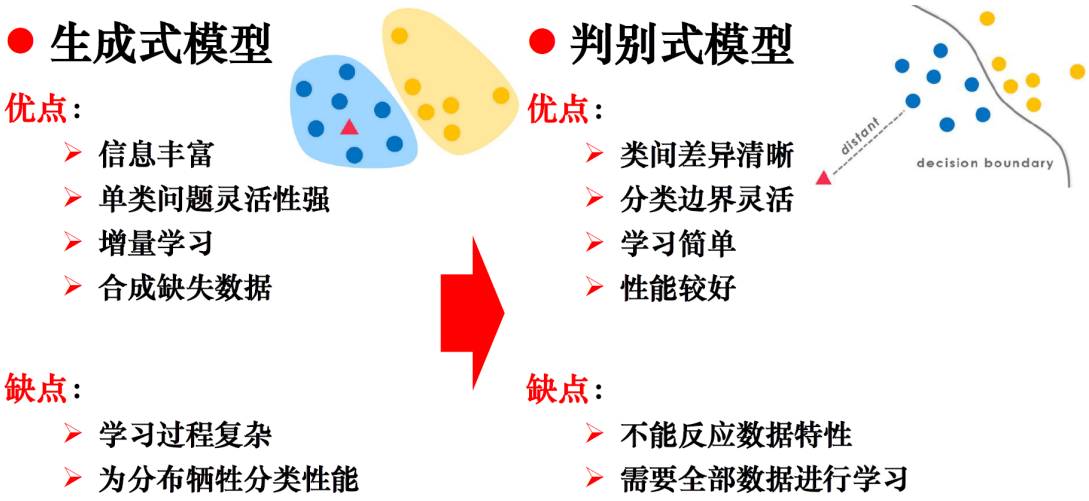
1.生成式（Generative）模型与判别式（Discriminative）模型

生成：分别对各类的类条件密度和先验概率进行建模，之后利用贝叶斯定理计算后验概率，或者直接对联合分布建模得到后验概率。

判别：直接对后验概率建模

如 GMM、朴素贝叶斯、隐马尔可夫等就是生成式模型，线性&逻辑斯蒂回归（+softmax 的判别）、SVM、k 近邻等都属于判别式模型。

优缺点：



另外，由生成模型可得判别模型，反之不可。

2.什么是支持向量机（SVM），其特点是什么？

SVM 从线性可分的最优分类面发展而来，SVM 即希望找到这样的分类超平面，使其不但能将两类正确分开，且分类间隔（Margin）最大。距离分类面最近的几个样本点被称为支撑向量。

对支撑向量，满足 $y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)=1$，SVM 的目标是最大化几何间隔，一般将其转化为二次规划问题：

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1,\quad i=1,2,\ldots,m$$

其特点是必须处理线性可分的问题，且大样本上很难实施，也无法处理多分类问题。由此，也衍生了如软间隔支持向量机、kernel SVM 等诸多处理线性不可分，以及避开高维映射的方法。

一般通过对偶问题和 KKT 条件求解 SVM 的二次规划问题。

3.软间隔支持向量机的思想

软间隔的核心思想就是允许支持向量机在一些样本上出错，以解决一些噪声、离群点等导致的线性不可分问题。实质上，这是一种正则化的思想。

4.核技巧的思想是什么，有哪些常用的核函数？

利用核函数，可以避免复杂的高维映射过程，而在低维空间中直接计算出高维映

射后的向量内积结果。

常用的核函数有线性核、多项式核、高斯核、Sigmoid 核等。

第六章 采样方法

1.蒙特卡洛方法的思想

其思想就是暴力采样，通过大量随机采样，去了解一个系统，进而得到想要的值，比如计算某个复杂函数的定积分等。区别于拉斯维加斯方法，蒙特卡洛方法的采样越多，越接近最优解。

2.基本采样法的思想及例题

其思想就是从一个基本概率分布（如均匀分布），通过函数变换产生新的分布。即$p(y)$是我们想要生成的分布，$p(z)$是均匀分布，我们希望找到一个$y=f(z)$，使得将采样的所有 z 变换后的 y 服从$p(y)$。

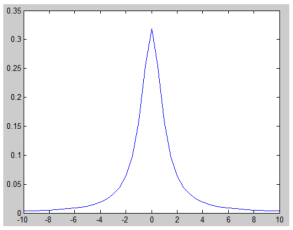
核心关系式为$p(z)dz=p(y)dy$，若在$(0,1)$上采样，可以直接得到$p(z)=1$，从而有$z=h(y)$，为其累积分布函数（CDF）。反解得到$y=h^{-1}(z)$，也即映射$f=h^{-1}$。【实际上就是求分布函数、反解的过程】

例子：标准柯西分布

已知 $z\sim U(0,1)$

求 $y=f(z)$

使 $p(y)=\frac{1}{\pi}\frac{1}{1+y^2}$



解：（1）求分布函数 $h(y)=\int_{-\infty}^{y}p(\hat{y})d\hat{(y)}=\frac{1}{\pi}\arctan y+\frac{1}{2}$，$z=h(y)$

（2）反解：$y=h^{-1}(z)=\pi\tan(z-\frac{1}{2})$

指数分布 $p(y)=\lambda\exp(-\lambda y)\quad y\in[0,\infty)$

求 $y=f(z)$

3.接受-拒绝采样的思想及步骤

思想：借助一个容易采样的分布$q(z)$（也叫建议分布）去逼近一个很复杂的分布

$\tilde{p}(z)$（很可能是积分/反函数不好求，基本采样做不了的）。

其基本步骤为：

（1）首先找到一个满足不等式$kq(z) \geq \tilde{p}(z)$的常数 k （一般取最小值，即让建议分布的倍数盖过目标分布）

（2）从建议分布$q(z)$中抽样，产生z_0

（3）从均匀分布$U(0, kq(z_0))$中抽样，得到样本u_0

（4）若$u_0 \geq \tilde{p}(z_0)$，则拒绝采样，否则接受采样

循环 2-4 直至达到停机条件。

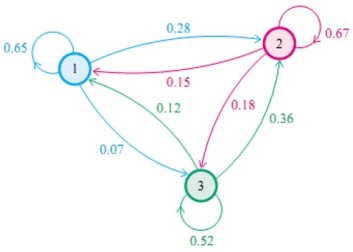
【直观理解：步骤 2 生成横坐标，步骤 3 生成纵坐标，步骤 4 比较大家，最终生成一个目标分布的类蒙特卡洛采样】

【若建议分布也为均匀分布，这时就退化为了蒙特卡洛方法】

4.什么是马尔科夫链？

如果一个过程的“将来”仅依赖于“现在”而不依赖“过去”，则此过程具有马尔可夫性,或称此过程为马尔可夫过程。时间和状态都离散的马尔可夫过程称为马尔可夫链。

5.马尔科夫链例题（阶级固化）



		子代		
	State	1	2	3
父代	1	0.65	0.28	0.07
	2	0.15	0.67	0.18
	3	0.12	0.36	0.52

$$P = \begin{bmatrix} 0.65 & 0.28 & 0.07 \\ 0.15 & 0.67 & 0.18 \\ 0.12 & 0.36 & 0.52 \end{bmatrix}$$

递推式：$\pi_n = \pi_i P^{n-i}$ 其中π_i表示状态向量，即当前的分布情况，P 为状态转移矩阵，经过不断迭代后发现收敛，此时达到细致平稳条件，称此时分布为平稳分布。

6.简述 M-H 方法的思想

对于需要采样的一个分布$p(z)$，构造一个转移矩阵为 T 的马尔可夫链，使它的平稳分布恰好为$p(z)$。如果马尔可夫链在第n步已经收敛了，于是我们就得到了$p(z)$的样本$x_n, x_{n+1}, \ldots$

基本步骤就是，在收敛后对构造的马尔科夫链采样，再从均匀分布中采样，若均匀分布采样值 u 小于接受率则接受。初始 Metropolis 方法的接受率较小采样效率低，之后 Hastings 改进了接受率，形成了 M-H 方法

（接受率改进为：$\min\left\{1, \frac{\tilde{p}(z)q(z'|z)}{\tilde{p}(z')q(z|z')}\right\}$）

7.Gibbs 采样与 M-H 方法的关系

Gibbs 采样是一种特殊的 M-H 采样方法（实际上两者都是 MCMC 方法），其特点是：每次只改变一个维度上的值，保持其他维度不变。

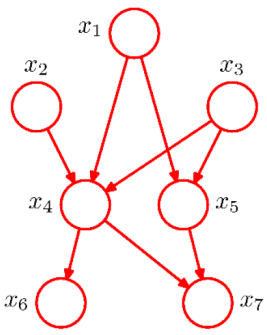
第七章 概率图模型

1.什么是概率图模型（PGM）？有哪些常见的概率图模型？

PGM 是一类用图来表达变量相关关系的概率模型，一般地，结点表示一个或一组随机变量，边表示变量之间的概率相关关系。

又可分为有向图模型，如贝叶斯网（Bayesian network）；或无向图模型，如马尔可夫网络等。

2.贝叶斯网络联合概率密度计算

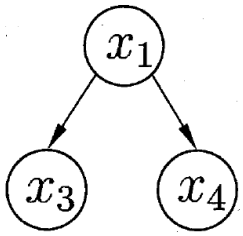


写出 $p(x_1, \ldots, x_7)$ ？

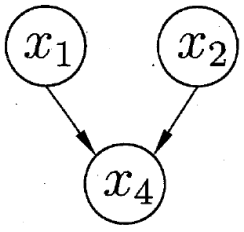
解：

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

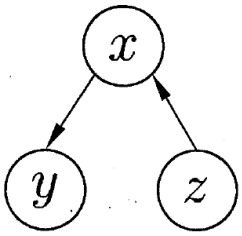
3.判断条件独立性



同父结构



V型结构



顺序结构

同父：给定 x1 时，x3 和 x4 条件独立

顺序：给定 x 时，y 和 z 条件独立

V 型结构：给定 x4，x1 和 x2 不独立，x4 未知时却独立！（反常）

需要会证明以上内容。

4.什么是隐马尔可夫模型（HMM）

区别于状态序列可见的马尔科夫过程，HMM 的状态序列不可见，只能得到对状态的观测序列。

HMM 是一个双重随机过程：一是马尔可夫随机，即**状态之间的转移**是随机的（比如选择箱子是随机的），使用**转移概率**描述；二是一般随机过程，即状**态生成观测的过程**也是随机的（比如从箱子里面摸球），使用**观测概率**描述。

5.HMM 的两个基本假设和三个基本问题及对应的算法

两个基本假设：齐次马尔可夫假设、观测独立假设（针对以上双重随机过程而言）

三个基本问题： λ表示模型参数，O 表示观测序列，I 表示状态序列

（1）概率计算问题：给定λ和O，求$P(O|\lambda)$？$\Longrightarrow$ 前向、后向算法

（2）学习问题：已知O，求λ使$P(O|\lambda)$最大？$\Longrightarrow$ EM算法（也叫Baum-Welch算法）

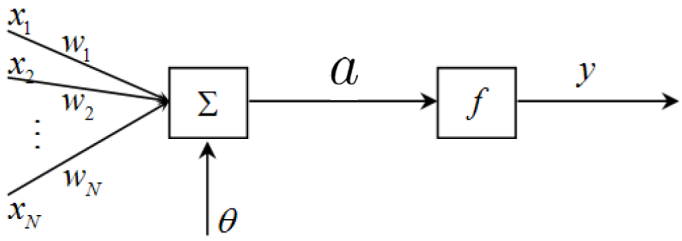
（3）预测（解码）问题：已知λ和O，求I使$P(I|O)$最大？$\Longrightarrow$ Viterbi算法

6.什么是条件随机场（CRF）

条件随机场是马尔可夫网络的一种，由无向图表示，比较常用的是线性链条件随机场，可以用于标注等问题，是一种判别模型。

第八+九章 神经网络

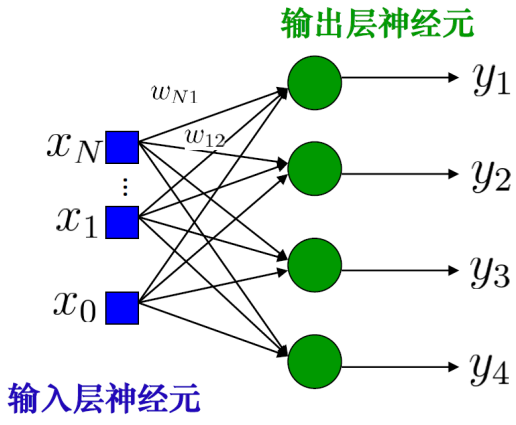
1.什么是人工神经元的 M-P 模型？画出模型并写出表达式。



$y = f(a) = f(\sum_{i=1}^{N} w_i x_i - \theta) = f(\sum_{i=0}^{N} w_i x_i)$ 其中 f 称为激活函数，MP 模型采用阶跃函数，此外激活函数还有 Sigmoid、双曲正切等

2.什么是感知机模型？相对 M-P 模型有何区别？

感知机模型相对 M-P 模型在输出层神经元上进行了扩展，M-P 模型是一个 PE（Processing Element，处理单元），而感知机模型则是由单层的多 PE 构成，形式上表现为：



，即 $y_j = f(a_j) = f\left(\sum_{i=1}^{N} w_{ij} x_i\right)$

3.感知机模型有何局限性，如何解决？

对于线性不可分问题，感知机问题难以处理，解决办法就是增加感知机层数构成多层感知机（MLP）。理论证明，三层感知机可以实现任意的逻辑运算，在激活函数为 Sigmoid 函数的情况下，可以逼近任何非线性多元函数。

4.推导 BP 反向传播算法的过程？

本质上就是链式求导法则。画出 MLP 模型会更有助于推导。

要明确核心目标：要求的是误差对于权重更新的影响。

$$E_n(\mathrm{w}) = \frac{1}{2}\sum_{k=1}^{3}(y_{nk} - t_{nk})^2$$

误差，也即准则函数定义为：

此处表示多层感知机有 n 层，这里的 Sum 上限为第 n 层神经元个数，此处为 3。BP 的思路是将此误差逐层求导，得到逐层的梯度，然后每层在负梯度上以一定学习率调整优化。

另外要把握核心关系式：$y_k = h(a_k) = h(\sum_j w_{kj}x_j)$，也就是说，对某一层（神经元）$j$求导，就要往该层（神经元）的输出$z_j = h(a_j)$和上层给予的输入$a_j = \sum_i w_i z_i$（j 的上一层有 i 个神经元）上靠，用链式求导法则。

定义δ_k表示对隐层 k 的“误差”（隐层需要做出的调整步长），那么有：

对于输出层而言： $\delta_k \equiv \frac{\partial E_n}{\partial a_k} = y_k - t_k$ （这里a_k实际上就是y_k）

对于隐藏层而言： $\delta_j \equiv \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} = \sum_k \delta_k \frac{\partial a_k}{\partial a_j} = \sum_k \delta_k \frac{\partial a_k}{\partial z_j}\frac{\partial z_j}{\partial a_j} = \sum_k \delta_k w_{kj}\frac{\partial h(a_j)}{\partial a_j}$

$= h'(a_j)\sum_k \delta_k w_{kj}$ （注意这里就是往 z 和 a 上靠，无论多少层都可以这样链式求导）

5.人工神经网络的优化方法有哪些，各有什么优劣？

主要有 SGD、BGD、Mini-BGD、Adagrad、RMSprop、Adadelta、动量 SGD、NAG、Adam 等。

Adam 是最常用、最稳定的方法，能够跳出局部极值和鞍点，但是不易收敛到最优值，常用在 NLP 领域（毕竟 NLP 经常就是局部坑），由于是自适应，也不用怎么调参；

相比之下 SGD 就容易在局部极值震荡，且高维空间存在大量鞍点，效果不好，但是收敛时易收敛至最优值。为了抑制 SGD 的震荡，可以对其加入动量（Momentum），或使用 Nesterov 梯度等方法改进。

采用变学习率则衍生了 Ada（adaptive）族的优化方法，进入了“自适应学习率”优化算法时代。这类算法往往开始学习的速度较快，之后学习速度较慢。

6.什么是梯度消失问题？有哪些解决方法？

在神经网络中，当隐藏层的数目增加后，因为 Sigmoid 函数的导数呈驼峰状，其

最大值是$\frac{1}{4}$，小数与小数相乘最终导致梯度接近于 0，这就是梯度消失问题。这会造成距离输出较近的几层正常更新，而距离输入较近的几层几乎没有变化。

可以使用 ReLU 激活函数避免梯度消失和爆炸，并且简化计算过程；采用 Dropout、输出归一化等正则手段可以避免过拟合，也可以缓解梯度消失/爆炸问题。

第十章 决策树

1.什么是决策树？

决策树的基本思想是，采用自顶向下的递归方法，以信息熵（系统的不确定度）为度量构造一棵熵值下降最快的树，到叶子结点的熵值为 0，此时叶结点中的实例都属于同一类。

可以讲决策树看成一个 if-then 的规则集合，从特征空间角度来看，决策树就是将特征空间划分为多个不相交的单元或区域，以达到分类的目的。

2.构建决策树的基本过程，有哪些决策树算法？

（1）选择一个属性作为决策树的根节点

（2）a.若当前实例都属于同一类别，则标记为叶结点，不再划分

b.否则，选取一个从该节点到根路径中没有出现过的属性，以该属性的不同特征作为分支，将样本划分为多个集合。

不断循环（2），直至所有的样本被划分完毕。

如何选择属性进行划分是决策树构建的关键，也因此衍生了不同决策树算法，但无论如何，选择的属性应该使结点的“纯度”提高。根据目标函数的不同，主要有以下决策树算法：

ID3： 信息增益

C4.5： 信息增益率

CART：基尼指数

3.什么是 ID3 决策树？

ID3 使用的目标函数是信息增益，在每次选择属性时，优先选取信息增益最大的属性，构造一棵熵值下降最快的决策树。

设样本集 D，属性 a，则信息增益定义为经验熵与经验条件熵的差值：

$$G(D,a) = H(D) - H(D \mid a)$$
$$= H(D) - \sum_{n=1}^{N} \frac{|D^n|}{|D|} H(D^n)$$

其中 N 表示当前属性下特征的数目，D^n表示当前属性、特征为 n 的样本子集，$H(D)$定义为：

$$H(D) = -\sum_{c=1}^{C} p_c \log_2 p_c, \quad p_c = \frac{D_c}{D}$$

其中 C 表示样本类别总数，p_c为类别为 c 的样本所占比例。

4.ID3 例题（见 ppt）

5.什么是 C4.5 决策树，与 ID3 的区别？

C4.5 是 ID3 的一种改进。信息增益准则对可取值数目较多的属性有所偏好，故 C4.5 不直接使用信息增益，而是使用“增益率”来选择最优划分属性。增益率定义为信息增益与属性固有值的比值，固有值的定义类似于信息熵，可能取值的数目越多，固有值越大。

不过增益率准则对可取值数目较少的属性有所偏好，所以 C4.5 使用了一种启发式选择方法：先从划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的。

6.什么是 CART 决策树？

CART 决策树使用基尼指数来划分属性，基尼指数定义为：

$$Gini(D) = \sum_{c=1}^{C} \sum_{c' \neq c} p_c p_c' = 1 - \sum_{c=1}^{C} p_c^2$$

属性 a 的基尼指数定义为：

$$Gini(D,a) = \sum_{n=1}^{N} \frac{|D^n|}{|D|} Gini(D^n)$$

直观上讲，基尼指数反映了从数据集中随机抽取两个样本，其类别标记不一致的概率，基尼指数越小纯度自然越高。

7.决策树剪枝的目的和基本策略？各有什么优劣？

基本策略分为预剪枝和后剪枝。剪枝是对付过拟合，提高决策树泛化能力的主要手段。

预剪枝是在生成过程中，每次结点划分前进行验证集上的估计：若划分不能提高决策树泛化性能，则停止划分并设为叶结点。

后剪枝是先生成决策树，自底向上对非叶结点进行考察，若将该结点子树替换为叶结点（删除对应子树），能带来泛化性能的提升，则将该子树替换为叶结点。

预剪枝使决策树没有“展开”，减少了训练和测试开销，但因其本质是贪心，所以存在欠拟合的风险；后剪枝通常保留了更多的分支，泛化性能往往优于预剪枝，但其训练时间开销要大得多。

8.决策树中如何处理连续值？

对 n 个连续值，从小到大排列，每两个相邻值取中位点得到 n-1 个划分候选点，对每个划分候选点都可以计算信息增益，选择其中信息增益最大的作为属性的信息增益即可。

与离散值不同，连续值在其后代结点还可以做为划分属性。

9.决策树中如何处理缺失值？

（1）在计算信息熵时忽略掉缺失样本，计算出信息增益后乘以系数 rho，其中 rho 为完整样本占整个样本的数目比例。

（2）在确定划分属性后，对于缺失值，以不同权重同时进入各个属性，权重分别对应各个特征的样本数目比例，其余完整样本的权重保持为 1。

第十一章 集成学习

1.什么是集成学习？

通过将多个学习器进行集成，通常可获得比单一学习器显著优越的泛化性能，这对弱分类器尤为明显。当然了，多个分类器不一定比单一学习器性能更好，集群可能提升性能，也可能起副作用。

【集成学习并不是具体的训练方法，往往可以和之前的这些算法行程组合拳，比如 RF 就是结合了决策树】

2.集成学习方法如何分类，各有哪些代表，及其思想？

首先是串行化方法，个体学习器间存在强依赖关系：Boosting，Adaboost

Boosting：串行生成学习器。初始训练一个基学习器，随后使错分样本受到更多关注（比如提升其相应的权重），然后训练下一个基学习器。重复进行至学习器数目达到设定，最终将这些基学习器加权结合。

Adaboost：Boosting 族最具有代表性的一种算法，其步骤具体为：

（1）基于初始分布D_t从数据集D中训练处分类器h_t

（2）估计h_t的误差

（3）确定h_t的权重α_t

（4）基于指数损失函数更新样本分布，并加入规范化因子确保D_{t+1}是一个分布

循环迭代 1-4，直至达到训练轮数。

其次是并行化方法，不存在强依赖关系：Bagging、随机森林（RF）

Bagging（Bootstrap AGGregatING）：基于自助采样法，可以获得 T 个含 m 个训练样本的采样集，这样我们可以直接训练 T 个基学习器（这里是并行的），然后再将这些基学习器结合即可。通常，对分类任务使用简单投票法，对回归任务使用简单平均法。

【从偏差-方差分解的角度讲，Boosting 比较关注降低偏差，Bagging 比较关注降低方差】

RF（Random Forest）：是 Bagging 的一个变体，其学习算法是基于决策树的，只是在其上引入了随机属性选择。具体地，对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含 k 个属性的子集，再从这个子集中选择最优属性划分。若 k=d（当前结点属性数），则退化为传统决策树；若 k=1，则完全随机，一般取$k = \log_2 d$。再将这些决策树结合即形成了随机森林。

RF 简单、容易实现、计算开销小、效果好，被誉为“代表集成学习技术水平的方法”。

第十二章 数据降维

1.为什么要进行数据降维，有哪些常用降维方法？

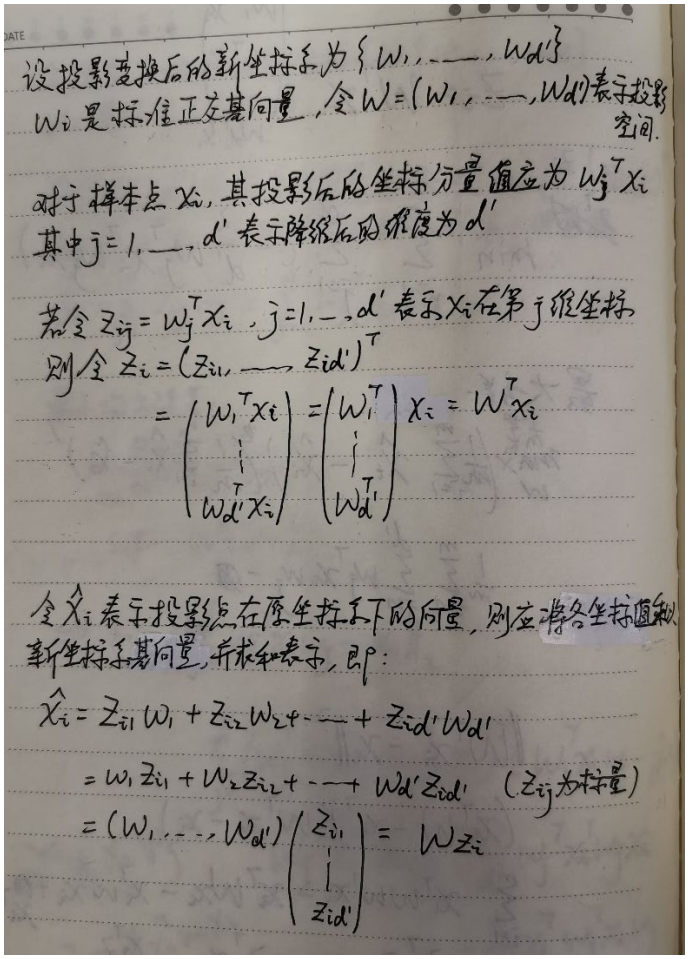
数据经过降维以后，可以保留原有数据的主要信息、提取本质结构，减少冗余信息和噪声信息造成的误差，提高应用中的精度。此外还可以避免维数灾难，训练和预测的时间效率将大为提高。

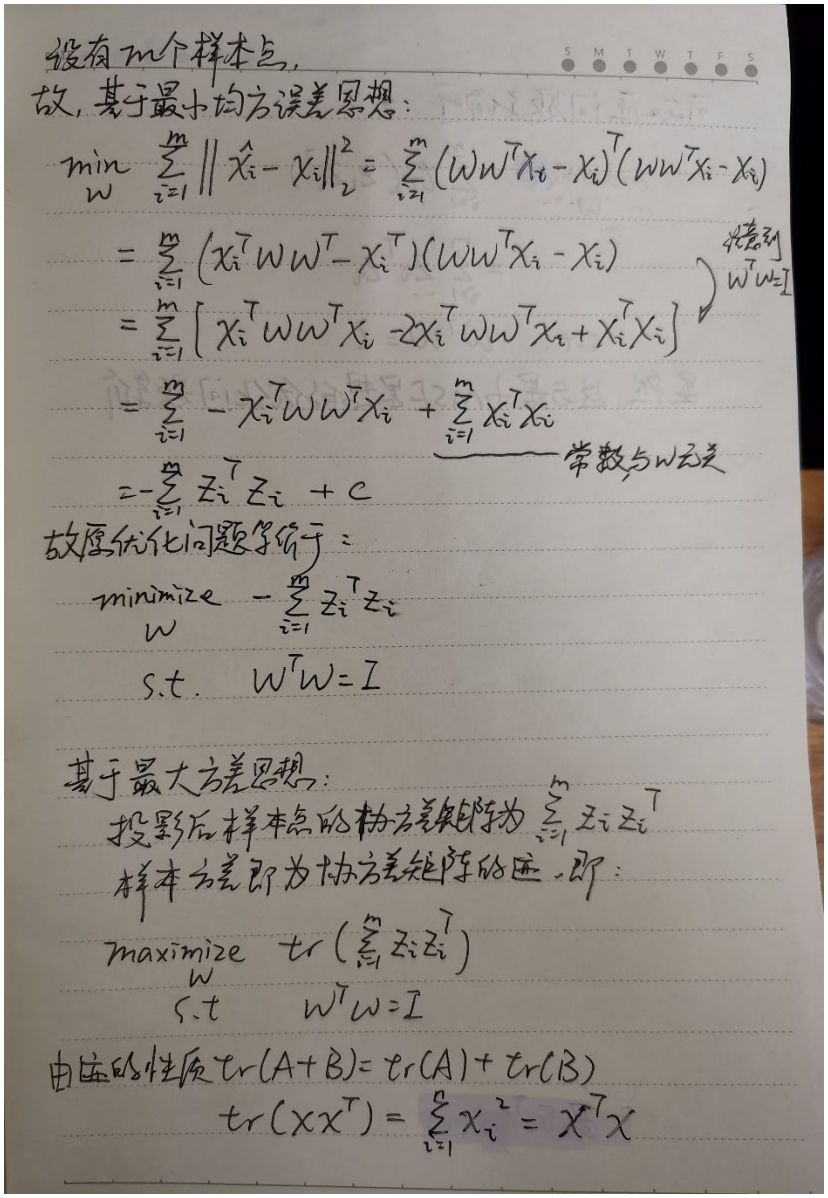
常用的方法有主成分分析（PCA）、等距映射、局部线性嵌入等。

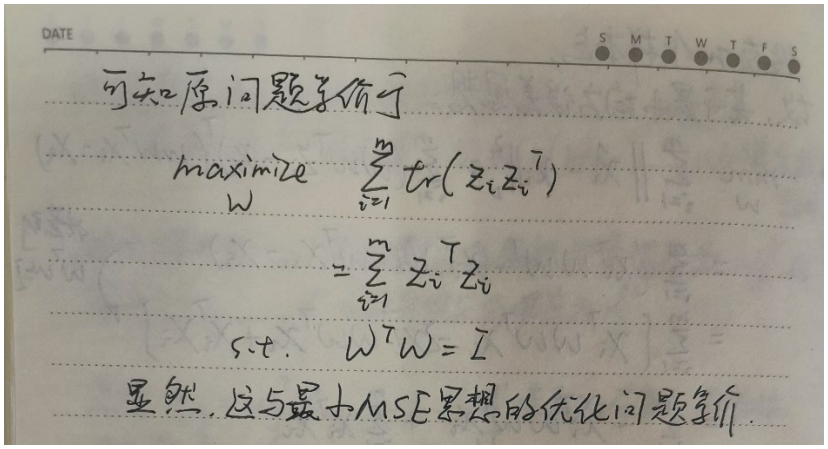
2.PCA 的思想，从两个角度分别解释和推导 PCA 的优化目标？

如果我们希望用一个超平面来表达一组数据，我们一般希望这个超平面具有最近重构性，即样本点到超平面足够近；还希望具有最大可分性，即样本点在超平面上的投影都尽可能分开。这就是 PCA 的主要思想。

也就是说，PCA 希望降维后的样本方差尽可能大，均方误差尽可能小，这两个角度殊途同归，从任何一个角度都可以推出 PCA 的优化目标。




设有m个样本点，
故，基于最小均方误差思想：
$\min_W \sum_{i=1}^m \|\hat{x}_i - x_i\|_2^2 = \sum_{i=1}^m (WW^Tx_i - x_i)^T(WW^Tx_i - x_i)$
$= \sum_{i=1}^m (x_i^TWW^T - x_i^T)(WW^Tx_i - x_i)$
$= \sum_{i=1}^m [x_i^TWW^Tx_i - 2x_i^TWW^Tx_i + x_i^Tx_i]$
提到 $W^TW=I$
$= \sum_{i=1}^m -x_i^TWW^Tx_i + \sum_{i=1}^m x_i^Tx_i$
常数与W无关
$= -\sum_{i=1}^m z_i^Tz_i + c$
故原优化问题等价于：
$\underset{W}{\text{minimize}} \; -\sum_{i=1}^m z_i^Tz_i$
$\text{s.t.} \quad W^TW = I$
基于最大方差思想：
投影后样本点的协方差矩阵为 $\sum_{i=1}^m z_iz_i^T$
样本方差即为协方差矩阵的迹，即：
$\underset{W}{\text{maximize}} \; tr(\sum_{i=1}^m z_iz_i^T)$
$\text{s.t} \quad W^TW = I$
由迹的性质 $tr(A+B) = tr(A) + tr(B)$
$tr(xx^T) = \sum_{i=1}^n x_i^2 = x^Tx$


DATE
可知原问题等价于
$\underset{W}{\text{maximize}} \; \sum_{i=1}^m tr(z_iz_i^T)$
$= \sum_{i=1}^m z_i^Tz_i$
$\text{s.t.} \quad W^TW = I$
显然，这与最小MSE思想的优化问题等价。

3.PCA 有哪些应用？

图像压缩，还有模式识别任务，如人脸检测和匹配等。

4.PCA 与 LDA？

从用途角度说：

PCA 追求降维后最大化保持数据内在信息，对于数据的区分作用不大，反而可能使数据点混杂在一起

LDA 追求数据在降维后能够很容易被区分开。

从降维角度说：

PCA 可以将数据降至任意维度，而 LDA 只能降一个维度。

5.什么是等度量映射（ISOMAP），有什么优缺点？

其出发点是，低维空间在映射到高维后，直接计算直线距离具有误导性，而应该使用测地线距离，于是可以构建一个近邻连接图，求解测地线距离就成为了一个求图上最短路径问题，这就有很多成熟的求解方法了。

其优点是，非线性非迭代求解方便，且保证了全局最优；缺点是易受到噪声干扰，且在大曲率区域存在短路现象，样本量不宜过大等

6.什么是局部线性嵌入（LLE）

ISOMAP 试图保持邻域间的样本距离，而 LLE 则试图保持邻域间样本的线性关系，即某一样本点可以被其邻域样本点的线性组合重构出来。也因此，其目标就是最小化重构误差，可以应用在点云等信息系统中。