金牌讲师团模式识别与机器学习复习提纲

（系笔者自行总结，仅作为复习参考）

一、 概论部分

- 什么是无监督学习？监督学习？
- 信息论（熵、KL 散度、信息增益）
- 概率论（条件概率、全概率公式、贝叶斯公式）

二、 机器学习

- MLE 和 MAP 的思想与区别
- 过拟合与欠拟合
- 什么是线性分类器？
- 朴素贝叶斯（离散、连续）
- 逻辑回归
- 支持向量机
- 聚类（层次聚类、K-means、高斯混合模型）

三、 深度学习

- 常用的激活函数？损失函数？
- 梯度消失与梯度爆炸？
- 常用的优化算法（梯度下降和随机梯度下降）
- 多层感知机 MLP（前馈传播和反向传播）
- 卷积神经网络 CNN（卷积的计算、lenet、resnet、alexnet）
- 序列神经网络 RNN、LSTM、Transformer
- 图像生成网络 VAE

2023.11.15

金牌讲师团

附：一些常用算法的数学推导

多项式拟合（回归）

设观测样本数为N，多项式函数形式为$y(x,a)=a_m x^m+a_{m-1}x^{m-1}+\cdots+a_1x^1+a_0=\sum_{j=1}^{m}a_j x^j$，可形式化为一个最优化问题：

$$\min E(w), E(w)=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,a)-t_n\}^2$$

而$y(x,a)$又可写成矩阵相乘形式为

$$\begin{bmatrix} x_1^m & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_N^m & \cdots & 1 \end{bmatrix}\begin{bmatrix} a_m \\ \vdots \\ a_0 \end{bmatrix}=\begin{bmatrix} y_m \\ \vdots \\ y_0 \end{bmatrix}$$

为方便表示可记为$X\cdot A=Y$，其中X为范德蒙德行列式(Vandermonde determinant)，A为模型参数，则残差平方和$loss$为：

$$loss=\|X\cdot A-Y\|^2$$

则目标函数为min $(loss)$，将$loss$展开得到，

$$loss=(X\cdot A-Y)^T\cdot(X\cdot A-Y)=(A^T\cdot X^T-Y^T)\cdot(X\cdot A-Y)$$
$$=A^TX^TXA-A^TX^TY-Y^TXA+Y^TY=A^TX^TXA-2A^TX^TY+Y^TY$$

由于$loss$是A的函数，记作$L(A)$，则对A求导$\frac{\partial(L(A))}{\partial A}$即为$\nabla L(A)$。

$$\nabla L(A)=\frac{\partial\big(L(A)\big)}{\partial A}=\frac{\partial(A^TX^TXA-2A^TX^TY+Y^TY)}{\partial A}=2X^TXA-2X^TY$$

使用$\nabla L(A)$对$L(A)$进行梯度下降优化迭代，即可求得局部最优模型参数A_θ。

最小二乘法

上述$L(A)$对A求导$\nabla L(A)=\frac{\partial(loss)}{\partial A}=0$即求得模型参数$X$，解得$A=(X^TX)^{-1}X^TY$，将数据点信息代入上述公式即可求得多项式拟合模型。

梯度下降

梯度下降法(Gradient Descent)的基本思想是通过迭代地更新模型参数，沿着目标函数的负梯度方向，按照一定步长逐步逼近最优解。

ALGORITHM Gradient Descent (梯度下降)

1: **input** $L(A)$ ←目标优化函数，α ←学习率，$iter$ ←迭代次数，A_0 ←参数初值；
2: $A\leftarrow A_0$；
3: **while** $i<iter$ **do**
4: $gradient\leftarrow\nabla L(A_{i-1})$；
5: $A_i\leftarrow A_{i-1}-\alpha\times gradient$；
6: **end while**
7: **return** A_θ //返回最优参数；

共轭梯度法

上述$loss$最小化问题可等价于求解等式$X^TXA = X^TY$，可转化为最小化问题

$$\min_{A^*} \frac{1}{2}A^TX^TXA - X^TYA$$

构造n个相互关于X^TX共轭的向量$d_1, d_2, d_3, \dots, d_n$，则空间任何向量$A'$都可以表示为$A' = \sum \alpha_i p_i$，则上述优化问题可以等价于

$$\min_{r_1,\dots,r_n} \frac{1}{2}(\sum \alpha_i p_i)^T X^T X\left(\sum \alpha_j p_j\right) - X^TY(\sum \alpha_i d_i) = \frac{1}{2}\sum\sum \alpha_i\, \alpha_j p_i^T X^T X p_j - \sum \alpha_i X^T Y p_i$$

由于相互共轭，因此当$i \neq j$时，有$p_i^T X^T X p_j = 0$，所以上式可变为

$$\min_{r_1,\dots,r_n} \frac{1}{2}\sum \alpha_i^2 p_i^T X^T X p_i - \sum \alpha_i X^T Y p_i = \sum \left(\frac{1}{2}\alpha_i^2 p_i^T X^T X p_i - \alpha_i X^T Y p_i\right)$$

对其各项求导，即可得$\alpha_i^* = \frac{X^T Y d_i}{d_i^T X^T X d_i}$，因而最终最优解$A^* = \sum \alpha_i^* p_i$。

通过寻找共轭梯度进行求解能够较梯度下降优化更快的收敛，设残差$r_k = b - X^T X A_k$，残差向量相互**正交**，迭代过程中满足$r_i^T r_j = 0, i \neq j$。令p_k表示每次迭代过程中的搜索方向，且满足与过去的搜索方向都**共轭**，即$p_{i+1} = r_{i+1} + \beta_k p_i$, $p_i^T X^T X p_j = 0$，联立即可解得α_k, β_k的更新公式。

ALGORITHM Conjugate Gradient Descent(共轭梯度下降)

1：**input** $L(A)$ ←目标优化函数，$iter$ ←迭代次数，A_0 ←参数初值；
2：$Q \leftarrow X^TX$，$b \leftarrow X^TY$，残差$r_0 \leftarrow b - QA_0$，搜索方向$p_0 \leftarrow r_0$，$i \leftarrow 0$；
3：**while** $i < iter$ **do**
4：　　$\alpha_i \leftarrow \frac{r_i^T r_i}{p_i^T Q p_i}$;
5：　　$r_{i+1} \leftarrow r_i - \alpha_i Q p_i$，$\beta_{k+1} \leftarrow \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$;
6:　　$p_{i+1} \leftarrow r_{i+1} + \beta_k p_i$，$A_{i+1} \leftarrow A_i + \alpha_i p_i$;
7：**end while**
8：**return** A_θ //返回最优参数；

逻辑回归

二项逻辑回归可用于一些简单的二分类问题，不妨设$Z \in R^n$是问题的输入，$X \in \{0,1\}$是输出。则逻辑回归假设$X|Z$服从伯努利分布。具体模型为

$$P(X = 1|Z) = \frac{1}{1 + e^{-\theta^T z}};\ P(X = 0|Z) = 1 - \frac{1}{1 + e^{-\theta^T z}}$$

其中，θ^T为模型的参数，$\theta^T z = \theta_0 + \theta_1 z^{(1)} + \cdots + \theta_n z^{(n)}$。应用时，根据两个条件概率值的大小将实例分到概率值较大的一类。

和线性回归不同的是，逻辑回归通过Sigmoid函数引入了非线性因素，将实数域内的值约束到(0,1)的取值范围内，从而可以较为地轻松处理二分类问题。

$$Sigmoid(Z) = \frac{1}{1+e^{-z}}$$

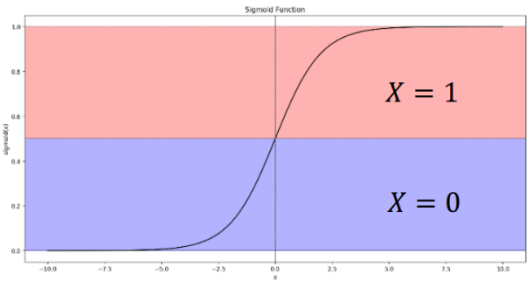


图 1 Sigmoid函数

令观测函数为$h_\theta(z) = 1/(1+e^{-\theta^T z})$，于是，可以整合上述逻辑回归的概率模型得到输入集合的最大似然估计为$L(\theta) = \prod_{i=1}^{m} P(x_i|z_i;\theta) = (h_\theta(z))^x(1-h_\theta(z))^{1-x}$，取对数，乘$-\frac{1}{m}$将最大值问题转化为最小值问题。

$$L(\theta) = \left(-\frac{1}{m}\right)\prod_{i=1}^{m}\log P(x_i|z_i;\theta) = -\frac{1}{m}\sum_{i=1}^{m}[x_i\log h_\theta(z_i) + (1-x_i)\log\big((1-h_\theta(z_i)\big)]$$

$L(\theta)$是凸的、可微的，因此可以运用梯度下降法、牛顿法及其他方法对此问题进行求解此优化问题。

4.1.2 梯度下降

梯度下降法(Gradient Descent)的基本思想是通过迭代地更新模型参数，对上述$L(\theta)$进行求导，得到其梯度各方向为

$$\frac{\partial L(\theta)}{\partial\theta_j} = \left(-\frac{1}{m}\right)\sum_{i=1}^{m}\left[x_i\frac{1}{h_\theta(z_i)}\cdot\frac{\partial h_\theta(z_i)}{\partial\theta_j} - (1-x_i)\frac{1}{(1-h_\theta(z_i)}\cdot\frac{\partial h_\theta(z_i)}{\partial\theta_j}\right]$$

其中，根据观测函数的定义可以展开$\frac{\partial h_\theta(z_i)}{\partial\theta_j}$为

$$\frac{-1}{\left(1+e^{-\theta^T z}\right)^2}(e^{-\theta^T z})\frac{\partial\theta^T z}{\partial\theta_j} = \text{sigmoid}(\theta^T z)\cdot\big(1-\text{sigmoid}(\theta^T z)\big)\frac{\partial\theta^T z}{\partial\theta_j}$$

将其代入原式，最终可以化简得到$\frac{\partial L(\theta)}{\partial\theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(z_i)-x_i)\,z_i^j$。

而后更新参数$\theta_j = \theta_j - \alpha\times\frac{1}{m}\sum_{i=1}^{m}(h_\theta(z_i)-x_i)\,z_i^j, j = 1,2,\dots,n$，重复上述过程，即可逐步向局部最优解逼近。

支持向量机 SVM 的原理

支持向量机(Support Vector Machines, SVM)是一类按有监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面，可以将问题化为一个求解凸二次规划的问题，求解能够正确划分训练数据集

并且几何间隔最大的分离超平面。

考虑在给定特征空间上的训练数据集$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$的二分类问题，其中，$x_i\in\mathbb{R}^n, y_i\in\{+1,-1\}, i=1,2,\dots N$，$x_i$为训练集第$i$个样本向量，$y_i$为其对应的类别。则目标是找到一个函数$g:\mathbb{R}^n\to\mathbb{R}$，而后通过$sgn(g(x))$映射到类别$\{+1,-1\}$上。

对训练集T，假设其是线性可分的，即$\exists w\in\mathbb{R}^n, b\in\mathbb{R}, \varepsilon>0$，使得对于所有$y_i=1$的样本，有$w^T\cdot x+b\geq\varepsilon$；而对所有$y_i=0$的样本，有$w^T\cdot x+b\leq\varepsilon$。

同时，对于给定的数据集T和超平面$w^T\cdot x+b=0$，定义超平面关于样本点(x_i,y_i)的几何间隔为

$$r_i=\frac{|w^Tx_i+b|}{\|w\|}=y_i\left(\frac{w^T}{\|w\|}\cdot x_i+\frac{b}{\|w\|}\right), r=\min_{i=1,2,\dots N} r_i$$

其中r为所有样本点的几何间隔的最小值。则 SVM 模型的求解最大分割超平面问题可以表示为以下约束最优化问题，也叫硬间隔(hard margin)最大化问题。

$$\max_{w,b} r, s.t. y_i\left(\frac{w^T}{\|w\|}\cdot x_i+\frac{b}{\|w\|}\right)\geq r, i=1,2,\dots N$$

将约束条件两边同时除以r，记$\hat{w}=\frac{w}{\|w\|r}, \hat{b}=\frac{b}{\|w\|r}$，因此最大化$r$等价于最小化$\frac{1}{2}\|\hat{w}\|^2$，则问题等价于

$$\min_{w,b}\frac{1}{2}\|\hat{w}\|^2, s.t. y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)\geq 1, i=1,2,\dots N$$

解这个凸二次规划问题，首先构造出拉格朗日函数，

$$L(w,b,\mu)=\frac{1}{2}\|\hat{w}\|^2-\sum_{i=1}^{N}\mu_i\left[y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)-1\right], \mu_i\geq 0$$

则问题的对偶函数及对偶问题为

$$\theta(\mu)=\min_{w,b} L(w,b,\mu);\ \max_{\mu}\theta(\mu), s.t.\ \mu\geq 0$$

首先求解$\theta(\mu)=\min_{w,b} L(w,b,\mu)$，显然最小值在$L$梯度为 0 时得到，即

$$\begin{cases}\frac{\partial L}{\partial w}=w-\sum_{i=1}^{N}\mu_i y_i x_i=0\\ \frac{\partial L}{\partial b}=-\sum_{i=1}^{N}\mu_i y_i=-\mu^T y=0\end{cases}$$

代回$L(w,b,\mu)$中求得最优解$\mu^*=[\mu_1^*,\mu_2^*,..,\mu_n^*]^T$，计算$w^*=\sum_{i=1}^{N}\mu_i^* y_i x_i$，由 KKT 互补条件知$\mu_i^*\left[y_i\left(\widehat{w^*}^T\cdot x_i+\widehat{b^*}\right)-1\right]=0$，当$\mu_i^*>0$时，有$y_i\left(\widehat{w^*}^T\cdot x_i+\widehat{b^*}\right)=1$。即$w^*$和$b^*$仅由对应$\mu_i^*>0$的样本点$(x_i,y_i)$决定，这样的$x_i$即为支持向量。

而如果数据集T线性不可分，则存在不满足约束条件$y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)\geq 1$的样本，会导致优化问题没有可行解。因此需要对目标函数和约束条件引入非负的松

弛变量ξ_i进行修正，即变为软间隔(soft margin)最大化问题。

$$\min_{w,b}\frac{1}{2}\|\hat{w}\|^2+C\sum_{i=1}^{N}\xi_i, s.t.\begin{cases}y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)\geq 1-\xi_i\\ \xi_i\geq 0\end{cases}, i=1,2,\dots N, C>0$$

其中，ξ_i为超平面加入了非线性，C为惩罚参数，约束最优分离超平面在间隔尽量大的同时尽量保持线性。同样地，定义其凸二次规划的拉格朗日函数。

$$L(w,b,\xi,\mu,\lambda)=\frac{1}{2}\|\hat{w}\|^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\mu_i\left[y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)-1+\xi_i\right]-\sum_{i=1}^{N}\lambda_i\xi_i$$

则问题的对偶函数$\theta(\mu)$及对偶问题为

$$\theta(\mu,\lambda)=\min_{w,b}L(w,b,\xi,\mu,\lambda);\ \max_{\mu,\lambda}\theta(\mu,\lambda), s.t.\ \mu,\lambda\geq 0$$

取L梯度为 0 代回可得$\theta(\mu,\lambda)=-\frac{1}{2}\mu^TQ\mu+\mathbf{1}^T\mu$，其中$Q$为对称矩阵，$Q_{ij}=y_iy_jx_i^Tx_j$，因而对偶问题化简为

$$\max_{\mu}-\frac{1}{2}\mu^TQ\mu+\mathbf{1}^T\mu, s.t.\begin{cases}0\leq\mu\leq C\cdot\mathbf{1}\\ \mu^Ty=0\end{cases}$$

解得最优解$\mu^*=[\mu_1^*,\mu_2^*,..,\mu_n^*]^T$，则原问题的最优解$w^*=\sum_{i=1}^{N}\mu_i^*y_ix_i$。而由 KKT 互补条件$b^*$的值由$\mu^*$和支持向量决定。同样地，支持向量对应 $0<\mu_i\leq C$的样本。

$$\begin{cases}\mu_i=0\Leftrightarrow y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)\geq 1\\ 0<\mu_i<C\Leftrightarrow y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)=1\\ \mu_i=C\Leftrightarrow y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)=1-\xi_i\leq 1\end{cases}$$

因此当 $0<\mu_i<C$时，有$y_i\left(\hat{w}^T\cdot x_i+\hat{b}\right)=1$成立，又因为$y_i^2=1$，于是可得$b^*=y_j-\sum_{support}\mu^*y_ix_i^Tx_j$，同理可以解得$\mu_i=C$时对应的$b^*$。

在实际应用中，还可以通过引入核函数，可以使 SVM 适用于非线性分类问题。通过一个非线性映射$\varphi:\mathbb{R}^n\to\mathbb{R}^m(m\gg n)$将低维的输入解耦至高维空间，从而实现数据的线性可分。

K-means 聚类算法

K-means 聚类算法的核心思想为假定聚类内部点之间的距离应该小于数据点与聚类外部的点之间的距离。即使得**每个数据点和与它最近的中心之间距离的平方和最小**。

假设数据集为$X=\{x_1,\dots,x_N\},x_i\in\mathbb{R}^D$，我们的目标是将数据集划分为$K$个类别$Y$。令$\mu_k\in\mathbb{R}^D,k=1,\dots,K$表示各类别的中心。聚类问题等价于求概率分布：

$$P(Y|X)=\frac{P(X|Y)\cdot P(Y)}{P(X)}$$

K-means 聚类相当于假设$P(X|Y)$服从多元高斯分布（**特征之间相互独立，协方差矩阵**$\Sigma=\lambda\boldsymbol{I}$），且$P(Y)$为等概率均匀分布。而$P(X)$为已知数据分布，从似然

的角度看，极大化$P(Y|X)$即等价于极大化$P(X|Y)\sim -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$，即最小化各数据点到其类别的均值。

多元正态分布：$\mathcal{N}(x|\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$

引入二值指示变量$r_{nk}\in\{0,1\}$表示数据点的分类情况，则可定义目标函数为

$$\min_{r,\mu}J=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\|x_n-\mu_k\|^2$$

因此最优化过程可以划分为两步：

1) 固定μ，优化r_{nk}。由于J关于r_{nk}是线性的，因此可以对每个n分别进行最小化，即对r_{nk}根据与聚类中心的距离进行最优化：

$$r_{nk}=\begin{cases}1, & k=argmin_j\|x_n-\mu_j\|^2\\ 0, & \text{其他情况}\end{cases}$$

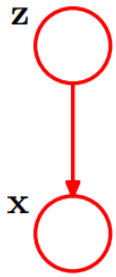
2) 固定r_{nk}，优化μ。由于J是μ的二次函数，对其求导等于零得

$$\frac{\partial J}{\partial\mu_k}=2\sum_{n=1}^{N}r_{nk}(x_n-\mu_k)=0\Rightarrow\mu_k=\frac{\sum_n r_{nk}x_n}{\sum_n r_{nk}}$$

即μ_k等于类别 k 的所有数据点的均值。

混合高斯模型

任意连续概率密度都能用多个高斯分布的线性组合叠加的高斯混合概率分布$p(x)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)$来描述。引入"1 of K"编码的二值随机变量z，满足$z_k\in\{0,1\}$且$\sum_{k=1}^{K}z_k=1$。



由右图模型定义联合概率分布$p(x,z)=p(z)\cdot p(x|z)$，z的边缘**先验**概率分布设为$p(z_k=1)=\pi_k(0\le\pi_k\le 1\text{ 且 }\sum_{k=1}^{K}\pi_k=1)$，也可写作$p(z)=\prod_{k=1}^{K}\pi_k^{z_k}$。

那么，x的条件概率分布为：

$$p(x|z_k=1)=\mathcal{N}(x|\mu_k,\Sigma_k)\Leftrightarrow p(x|z)=\prod_{k=1}^{K}\mathcal{N}(x|\mu_k,\Sigma_k)^{z_k}$$

于是可以给出$p(x)=\sum_z p(z)\cdot p(x|z)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)$，同时，$z$的条件后验概率$\gamma(z_k)$由贝叶斯定理得（已知为$x$，类别为$z_k$的概率）：

$$\gamma(z_k)\equiv p(z_k=1|x)=\frac{p(z_k=1)p(x|z_k=1)}{\sum_{j=1}^{K}p(z_j=1)p(x\mid z_j=1)}=\frac{\pi_k\mathcal{N}(x\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x\mid\mu_j,\Sigma_k)}$$

于是此聚类过程可以看做将概率分布$p(x)$解耦成K个高斯分布，对应K个类别。对于数据集$X=\{x_1,\dots,x_N\},x_i\in\mathbb{R}^D,X\in\mathbb{R}^{N\times D}$，对应隐变量表示为$Z\in\mathbb{R}^{N\times K}$。

则对数似然函数为

$$\ln\ p(X\mid\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n\mid\mu_k,\Sigma_k)\right\}$$

将此似然函数关于μ_k求导（假设Σ_k非奇异），令$N_k=\sum_{n=1}^{N}\gamma(z_{nk}),\gamma(z_{nk})\equiv p(z_k=1|x_n)$为能被分配到聚类$k$的有效数量，可以得到：

$$\sum_{n=1}^{K}\underbrace{\frac{\pi_k\mathcal{N}(x_n\mid\mu_k,\Sigma_k)}{\sum_j\pi_j\mathcal{N}(x_n\mid\mu_j,\Sigma_j)}}_{\gamma(z_{nk})}\Sigma_k^{-1}(x_n-\mu_k)=0\Rightarrow\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

由此式μ_k可视为当前所有点数据为第k类的概率加权平均。

同样地，将此函数关于Σ_k求导等于 0 可以得到：

$$\Sigma_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x-\mu_k)^T$$

最后使用拉格朗日乘子法关于π_k优化 $\ln p(X\mid\pi,\mu,\Sigma)+\lambda(\sum_{k=1}^{K}\pi_k-1)$（$\pi_k$需要满足和为 1 的条件）得到：

$$\sum_{n=1}^{N}\frac{\mathcal{N}(x_n\mid\mu_k,\Sigma_k)}{\sum_j\pi_j\mathcal{N}(x_n\mid\mu_j,\Sigma_j)}+\lambda=0\Rightarrow\pi_k=\frac{N_k}{N}$$

使用 EM 算法优化$\ln\ p(X\mid\pi,\mu,\Sigma)$即可总结为以下步骤：

ALGORITHM EM for Gaussian Mixture Models

1：**input** X ←数据集，K ←类别数目，$iter$ ←迭代次数；
2：初始化均值μ_k、协方差Σ_k和混合系数π_k
3：计算对数似然$\ln\ p(X\mid\pi,\mu,\Sigma)\leftarrow\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n\mid\mu_k,\Sigma_k)\}$
4：**while** $i<iter$ **do**
5：　　$\gamma(z_{nk})\leftarrow\frac{\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x_n|\mu_j,\Sigma_j)}$；　　（E 步）
6：　　$\mu_k^{new}\leftarrow\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$；
7：　　$\Sigma_k^{new}\leftarrow\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k^{new})(x-\mu_k^{new})^T$；　　（M 步）
8：　　$\pi_k^{new}\leftarrow\frac{N_k}{N}$；
9：**end while**
10：**return** μ_k,Σ_k,π_k //返回最优参数；

主成分分析 PCA

主成分分析（Principal Component Analysis,PCA) 基于协方差矩阵进行线性变换从而将高维数据转换为低维空间，并最大程度地保留原始数据的方差。

具体而言，PCA 的主要思想是从原始的空间中顺序地找一组相互正交的坐标轴，将n维特征映射到全新的k维正交特征向量上。而对于k维正交特征向量的选择，一般选择协方差矩阵对应特征值最大的k个特征向量，相当于保留k个原始数据中方差最大的方向，具体伪代码如下。

ALGORITHM Principal Component Analysis (主成分分析)

1：**input** X ←高维数据矩阵$(m \times n)$，k ←降维目标；

2：$X \leftarrow (X - \bar{X})$ //数据中心化；

3：协方差矩阵$C \leftarrow 1/m(X \cdot X^T)$；

4：求出C的特征值和特征矩阵，$W \leftarrow C$前k大的特征值对应特征向量构成；

5：**return** $W \cdot X$ //返回降维后的矩阵；

假设我们有一个样本集$\{x^1, x^2, \dots, x^m\}$，每个样本的特征数为n，那么我们可以用一个$n \times m$矩阵X来表示这个样本集。

$$X = (x^1, x^2, \dots, x^m) = \begin{pmatrix} x_1^1 & \cdots & x_1^m \\ \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^m \end{pmatrix}$$

那么我们希望能够找到一个$k \times n$的投影矩阵$W = [w^1, w^2, \dots, w^k]^T$，$w^i$为$1 \times n$的行向量，使得$W \cdot X = Z(k \times m) \leftarrow \{z^1, z^2, \dots, z^m\}$，实现对$X$的降维。

而对于X中的任意一个样本x，经过W投影过后得到$z = Wx$，z_i表示第i维新的特征，z_i^j表示第j个样本x^j经降维投影后的第i维特征，则有$z_i^j = w^i \cdot x^j$，于是第i维新特征的样本均值$\overline{z_i}$为

$$\overline{z_i} = \frac{1}{m}\sum_{j=1}^{m} z_i^j = \frac{1}{m}\sum_{j=1}^{m} w^i \cdot x^j = w^i \cdot \frac{1}{m}\sum_{j=1}^{m} x^j = w^i \cdot \bar{x}$$

要使在Z矩阵中的k个维度最大程度保留X的数据特征，则等价于此k个方向上方差最大，由此可将问题形式化为

$$max\ Var(z_i) = \frac{1}{m}\sum_{j=1}^{m} \left(z_i^j - \overline{z_i}\right)^2,\ \left\|w^i\right\|_2 = 1 \text{ 且} (w^i)^T \cdot w^j = 0$$

则$Var(z_i)$可等价变形为

$$\begin{aligned} Var(z_i) &= \frac{1}{m}\sum_{j=1}^{m} \left(z_i^j - \overline{z_i}\right)^2 = \frac{1}{m}\sum_{j=1}^{m} \left(w^i \cdot x^j - w^i \cdot \bar{x}\right)^2 \\ &= \frac{1}{m}\sum_{j=1}^{m} \left(w^i \cdot \left(x^j - \bar{x}\right)\right)^2 = \frac{1}{m}\sum_{j=1}^{m} \left(w^i\right)^T \left(x^j - \bar{x}\right)\left(x^j - \bar{x}\right)^T w^i \\ &= \left(w^i\right)^T \frac{1}{m}\sum_{j=1}^{m} \left(x^j - \bar{x}\right)\left(x^j - \bar{x}\right)^T w^i = \left(w^i\right)^T Cov(x) w^i \end{aligned}$$

令$S = Cov(x)$，则问题形式可简化为，

$$max\ (w^i)^T S w^i,\ \|w^i\|_2 = (w^i)^T w^i = 1 \text{ 且} (w^i)^T \cdot w^j = 0$$

则使用拉格朗日乘子法构造函数组$g(w)$如下

$$\begin{cases} g(w^1) = (w^1)^T S w^1 - \alpha((w^1)^T w^1 - 1) \\ g(w^2) = (w^2)^T S w^2 - \alpha((w^2)^T w^2 - 1) - \beta((w^2)^T w^1 - 0) \\ \cdots\cdots \\ g(w^k) = (w^2)^T S w^2 - \alpha((w^2)^T w^2 - 1) - \sum_{j=1}^{k-1} \beta_j \left((w^k)^T w^j - 0\right) \end{cases}$$

对w^i内各元素$w_1^i, w_2^i, \ldots, w_n^i$求偏导得到

$$\partial g(w^1)/\partial w_1^1 = 0, \partial g(w^1)/\partial w_2^1 = 0, \ldots, \partial g(w^k)/\partial w_n^k = 0$$

则对于w^1，解得$Sw^1 - \alpha w^1 = 0$，则两边同乘$(w^1)^T$可以得到等式

$$(w^1)^T S w^1 = \alpha (w^1)^T w^1 = \alpha$$

由此推出w^1为协方差矩阵S对应的最大特征值λ_1的特征向量。

而对于w^2，解得$Sw^2 - \alpha w^2 - \beta w^1 = 0$，则两边同乘$(w^1)^T$可以得到等式

$$(w^1)^T S w^2 - \alpha (w^1)^T w^2 - \beta (w^1)^T w^1 = 0$$

而由于前面两项正交，可得到

$$((w^1)^T S w^2)^T = (w^2)^T S^T w^1 = (w^2)^T S w^1 = \lambda_1 (w^2)^T w^1 = 0;$$

$$\alpha (w^1)^T w^2 = 0$$

所以$\beta = 0$，所以$Sw^2 - \alpha w^2 = 0$，由此推出w^2为协方差矩阵S对应的第二大特征值λ_2的特征向量。

同理$w^3, \ldots, w^k$分别对应协方差矩阵S的前k大特征值$\lambda_3, \ldots, \lambda_k$的特征向量，组合即可得到投影矩阵$W$。

将投影矩阵与输入矩阵做矩阵乘法$W \cdot X = Z$即可得到降维后的目标主成分矩阵Z，包含X中方差最大的k个特征。