3.4 机器学习相关数学基础

3.4.1 线性代数

线性代数是**应用数学**的一个重要分支，被广泛应用于自然科学和工程领域。线性代数主要处理线性问题。**线性关系**是指数学对象之间的关系是以**一次形式**来表达的。**行列式**和**矩阵**为处理线性问题提供了有力的工具。特别是的矩阵运算，是很多机器学习算法，尤其是深度学习算法设计思想的基础。在本节中仅涉及机器学习、深度学习相关的线性代数基础知识。包括：向量空间和范数、矩阵运算、投影变换、特征值和特征向量、矩阵特征分解、因式分解、对称矩阵、正交化/标准正交化、主成分分析（PCA）、奇异值分解（SVD）等。

1.标量、向量、矩阵和张量

在机器学习中，数据样本通常要由线性代数中以下 4 种结构类型量组成，并进行数值计算。

标量（scalar）：一个数据即一个数值，是计算的最小单元。

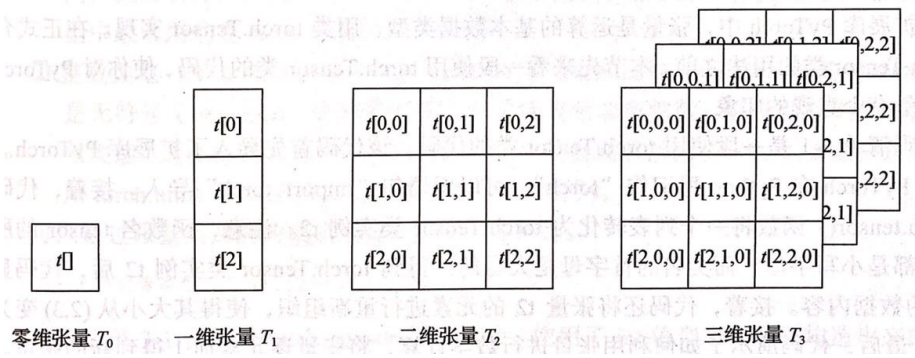
向量（vector）：物理上指具有大小和方向的量，形象化地表示为带箭头的线段。一个 n 维向量类型数据（如特征向量）是指由 n 个标量组成的有序数列。通常用小写字母表示，如 n 维向量$\mathbf{a} \in R^n$, 如果没特殊说明, 一般表示为**列向量 a**=$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$, 即大小为 n×1 的**矩阵**（matrix）。采用下标索引获取向量的每一个元素，a_i表示向量 **a** 的第 i 个**分量**，或第 i 维**元素**。为书写简便，有时会以加上转置符号的 1×n 矩阵表示列向量 **a**=$[a_1, a_2, \cdots, a_n]^T$。

矩阵：一个大小为 m×n 的矩阵类型数据（例如一帧位图）是指一个由 m 行 n 列标量排列成的矩形阵列。通常采用大写字母表示，如矩阵A $\in R^{m\times n}$。

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a_{11}} & \boldsymbol{a_{12}} & \cdots & \boldsymbol{a_{1n}} \\ \boldsymbol{a_{21}} & \boldsymbol{a_{22}} & \cdots & \boldsymbol{a_{2n}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{a_{m1}} & \boldsymbol{a_{m2}} & \cdots & \boldsymbol{a_{mn}} \end{bmatrix}$$

也可简记为$A_{m\times n} = (a_{ij})_{m\times n}$。与向量类似，可以通过给定行和列的下标索引来获取矩阵中元素，矩阵A从左上角数起第 i 行第 j 列上的元素（项），记为a_{ij}或$a_{i,j}$。一个 n 维向量可以看作是 n×1 的矩阵。特殊的，行数与列数都等于n的矩阵称为n阶矩阵或n阶方阵。

张量（tensor）:在深度学习中，很多时候是高于二维空间的数据，需要能表示任意维度空间中数据的结构类型，这就是张量。一个大小为 m×n×k 的张量数据是如下图所示的 m×n×k 个标量排列成的立方体。



这 4 种数据类型对应 NumPy 的多维数组类型 ndarray, 也对应 PyTorch 的 Tensor 类型，创建方式可见第二讲相关内容。

2. 线性相关与向量空间

矩阵可以看成由行向量或者列向量组成的**向量组**。即向量组与矩阵具有等价关系。

线性组合：给定向量组A: $a_1, a_2, \cdots, a_n (a_i \in R^m)$，对于任意一组实数$k_1, k_2, \cdots k_n$，把表达式：

$$k_1 a_1 + k_2 a_2 + \cdots + k_n a_n$$

称为向量组A的一个**线性组合**，$k_1, k_2, \cdots k_n$称为向量组的系数。对于任意一个 m 维向量b $\in R^m$，如果存在一组实数，使得：

$$b = k_1 a_1 + k_2 a_2 + \cdots + k_n a_n$$

成立，则称向量 b 可以被向量组A: $a_1, a_2, \cdots, a_n$**线性表示**。

向量空间（Vector Space）：也称**线性空间**（Linear Space），是指由向量组成的集合，并满足向量加法和标量乘法运算是**空间封闭**的，记作 **V**。

一个常用的线性空间是**欧氏空间**，通常表示为R^n，n 为空间**维度**。即由任意 n 个实数做系数构成的所有线性组合向量集合$\{k_1 a_1 + k_2 a_2 + \cdots + k_n a_n, k_i \in R\}$是一个 n 维欧式空间。例如，由向量组$\{[1,0,0], [0,1,0], [0,0,1]\}$和任意 3 个实数$k_1, k_2, k_3$构成的所有线性组合向量$[k_1, k_2, k_3]$构成了三维欧式空间$R^3$。

线性相关：给定向量组A: $a_1, a_2, \cdots, a_n (a_i \in R^m)$，如果存在不全为 0 的实数$k_1, k_2, \cdots k_n$，使得线性组合$k_1 a_1 + k_2 a_2 + \cdots + k_n a_n = \mathbf{0}$成立，则称向量组A是**线性相关**的，反之向量组A是线性无关的。

向量组的线性组合$k_1 a_1 + k_2 a_2 + \cdots + k_n a_n = \mathbf{0}$也可以看做$k_1, k_2, \cdots k_n$为未知量，向量组为系数的 **N 元齐次线性方程组**，当这个齐次线性方程组的系数矩阵是一个方阵时，这个系数矩阵存在**行列式为 0**，即方程组有非零解，从而系数向量组$a_1, a_2, \cdots, a_n$**线性相关**。

基向量组（坐标系）：n 维向量空间 V 的基向量是指 V 的有限子集$\{e_1, e_2, \cdots, e_n\}$，其中**两两向量线性无关**。即$e_i \in R^n, e_i^T e_j \neq 1, i \neq j$，V 中所有向量可以按唯一方式表示为基向量$e_i$的线性组合。即任意v $\in$ V，有$v = k_1 e_1 + k_2 e_2 + \cdots + k_n e_n$。称$(k_1, k_2, \cdots, k_n)$为向量$v$关于该基向量组的**坐标**。如果满足$\|e_i\|_2 = 1$, $e_i^T e_j = 0, i \neq j$, 则称**标准正交基**向量组（坐标系）。

向量组的**秩**（Rank）：给定向量组A: $a_1, a_2, \cdots, a_n (a_i \in R^m)$，如果能从中选出 r 个向量构成的子集$A_0: a_1, a_2, \cdots, a_r$, $r < n$，满足：

$a_1, a_2, \cdots, a_r$线性无关；向量组A的任意 r+1 个向量构成的子向量组都是线性相关的。则称子集$A_0: a_1, a_2, \cdots, a_r$, $r < n$是向量组A的一个**最大线性无关组**，最大线性无关组包含的向量个数 r 称为向量组A的**秩**。把和矩阵对应的行向量组的秩，称为矩阵的行秩，对应的列向量组的秩称为矩阵的列秩，矩阵的行秩与列秩相等，统称为**矩阵的秩**。

☆从线性空间映射角度来看，一个矩阵A $\in R^{m \times n}$定义了一个从线性空间R^n到R^m的**线性映射**（也叫**线性变换**）函数$f: R^n \to R^m$。即对于 n 维空间向量$\boldsymbol{x} \in R^n$和 m 维向量$\boldsymbol{y} \in R^m$，可以分别表示为的矩阵：

则 n 个 m 维列**向量组**对应的矩阵 A 为 y 到 x 的**线性映射函数**表示。即

$$\boldsymbol{y} = f(\boldsymbol{x}) = \mathrm{A}\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 a_{11} + x_2 a_{12} + \cdots + x_n a_{1n} \\ x_1 a_{21} + x_2 a_{22} + \cdots + x_n a_{2n} \\ \vdots \\ x_1 a_{m1} + x_2 a_{m2} + \cdots + x_n a_{mn} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix},$$

☆**仿射变换**：是指通过一个线性变换和平移，将一个空间向量变换为另一空间向量的函数。可以表示为:

$$\boldsymbol{y} = f(\boldsymbol{x}) = \mathrm{A}\mathbf{x} + \boldsymbol{b} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

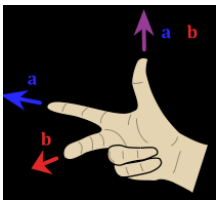
$$= \begin{bmatrix} x_1 a_{11} + x_2 a_{12} + \cdots + x_n a_{1n} + b_1 \\ x_1 a_{21} + x_2 a_{22} + \cdots + x_n a_{2n} + b_2 \\ \vdots \\ x_1 a_{m1} + x_2 a_{m2} + \cdots + x_n a_{mn} + b_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

其中$\boldsymbol{b} \in R^m$为平移项，当$\boldsymbol{b} = 0$时，仿射变换退化为线性变换。仿射变换不改变原始空间的向量相对位置关系。

向量运算

☆**内积**（Inner Product）：是 n 维线性空间中向量与向量的乘积，其结果是一个标量。$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}| \cos\theta = \sum_{i=1}^{n} a_i b_i$，也叫**点积**（Dot Product）或数量积、**标量积**（Scalar Product）。

特别地，向量自身的内积为**模**的平方，即$\boldsymbol{a} \cdot \boldsymbol{a} = |\boldsymbol{a}|^2$。两个向量方向的夹角可利用来内积计算，公式为$cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\sqrt{\boldsymbol{a} \cdot \boldsymbol{a}}\sqrt{\boldsymbol{b} \cdot \boldsymbol{b}}}$，如果空间中两个向量方向垂直$a \perp b$，其内积$\boldsymbol{a} \cdot \boldsymbol{b}$ =0，称两个向量**正交**（Orthogonal）。



叉积（Cross product），又称**向量积**（Vector product）、**叉乘**，对三维空间中的两个向量的乘积，结果是个向量。定义为$\boldsymbol{a} \times \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}| sin\theta \boldsymbol{i}$，$\boldsymbol{i}$是与$\boldsymbol{a}, \boldsymbol{b}$都垂直的单位向量。几何意义是所得乘积向量与两个被乘向量所在平面垂直，方向由右手定则规定，大小是两个被乘向量张成的平行四边形的面积。所以向量积不满足交换律。

外积（Outer Product）：两个向量$\boldsymbol{a} \in R^m$和$\boldsymbol{b} \in R^n$的外积是一个$m \times n$的矩阵，定义为

$$\boldsymbol{a} \otimes \boldsymbol{b} = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_m b_1 & a_m b_2 & \cdots & a_m b_n \end{bmatrix} = \boldsymbol{a}\boldsymbol{b}^T$$

，也称张量积，看作矩阵的克罗内克积（Kronecker）的一种特例。

3. **向量与矩阵的范数**

向量范数（Norm）：是**向量大小的一种度量方式**。是指一个表示“长度”概念的函数，把一个向量 a 映射为一个非负实数值，即满足$f: R^m \to R$，从而为向量空间内的所有向量赋予了一个非零的正长度或大小。在机器学习中，利用范数度量模型的复杂度，是优化模型参数时进行模型正则化，防止过拟合的主要手段。对于一个 n 维的向量a，常见的范数函数为 p 范数，$\| a \|_p = (\sum_{i=1}^{n} |a_i|^p)^{\frac{1}{p}}$，p 为一个标量参数，常用的 p 的取值有 1,2,∞等。

1 范数：向量各元素绝对值之和

$$\| a \|_1 = \sum_{i=1}^{n} | a_i |$$

2 范数：又称 E 范数或 F 范数，向量的**模**，是向量各元素的平方和求平方根

$$\| \mathrm{a} \|_2 = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a}_{\mathrm{i}}^2} = \sqrt{\mathrm{a}^{\mathrm{T}}\mathrm{a}}$$

∞范数：是向量各元素的最大绝对值

$$\| \mathrm{a} \|_\infty = \max\left(\{|\mathrm{a}_i|\}_{i=1}^{n}\right)$$

4. 矩阵的范数：满足$f: R^{m\times n} \to R$的非负函数，有很多种形式，

1 范数：也称为列和范数，定义为：$\| \mathbf{A} \|_1 = max_j\left(\{\sum_{i=1}^{m} |a_{ij}|\}_{j=1}^{n}\right)$

2 范数：也称为谱范数，定义为：$\| \mathbf{A} \|_2 = max_j\left(\{\sqrt{\lambda_j(A^T A)}\}_{j=1}^{n}\right)$，其中$\lambda_j(A^T A)$表示矩阵$A^T A$的第 j 个特征值。

∞范数：也称为行和范数，定义为：$\| \mathbf{A} \|_\infty = max_i\left(\{\sum_{j=1}^{n} |a_{ij}|\}_{i=1}^{m}\right)$

F 范数：

$$\| \mathbf{A} \|_F = \left(\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{m}} \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{n}} |\mathbf{a}_{\mathbf{ij}}|^2\right)^{1/2}$$

5. 矩阵的基本运算

如果A和B都为m× n 的矩阵，则A和B的加减结果也是m× n 的矩阵，其每个元素是A和B相应元素相加或相减。

$$(\mathrm{A}+\mathrm{B})_{\mathrm{ij}} = \mathrm{a}_{\mathrm{ij}} + \mathrm{b}_{\mathrm{ij}}$$

$$(\mathrm{A}-\mathrm{B})_{\mathrm{ij}} = \mathrm{a}_{\mathrm{ij}} - \mathrm{b}_{\mathrm{ij}}$$

A 和 B 的**点乘**$A \odot B \in R^{m\times n}$为 A 和 B 相应元素相乘：

$$(\mathrm{A} \odot \mathrm{B})_{\mathrm{ij}} = \mathrm{a}_{\mathrm{ij}}\mathrm{b}_{\mathrm{ij}}$$

一个**标量 c 与矩阵*A*乘积**为 A 的每个元素是 A 的相应元素与 c 的乘积

$$(\mathrm{cA})_{\mathrm{ij}} = \mathrm{ca}_{\mathrm{ij}}$$

两个**矩阵**的**乘积**：仅当第一个矩阵A的列数和另一个矩阵 B 的行数相等时才有定义。如 A 是 m×p 矩阵和 B 是 p×n 矩阵，则乘积AB是一个 m×n 的矩阵

$$(\mathrm{AB})_{\mathrm{ij}} = \sum_{\mathrm{k}=1}^{\mathrm{p}} a_{\mathrm{ik}}\, \mathrm{b}_{\mathrm{kj}}$$

☆**两个矩阵的乘积**可以看作矩阵表示的线性变换合成的**复合线性变换**。

$$\mathrm{y} = (\mathrm{AB})\mathrm{x} = \mathrm{A}(\mathrm{Bx}) = \mathrm{f}_A(\mathrm{f}_B(\mathrm{x}))$$

矩阵的乘法满足结合律和分配律，但不满足交换律。

结合律：$(AB)C = A(BC)$

分配律：$(A+B)C = AC + BC$，$C(A+B) = CA + CB$

矩阵**转置**：m×n 矩阵A的**转置**是一个 n×m 的矩阵，记为A^T，A^{T}第 i 行第 j 列的元素是原矩阵 A 第 j 行第 i 列的元素：

$$(\mathrm{A}^{\mathrm{T}})_{\mathrm{ij}} = a_{\mathrm{ji}}$$

矩阵转置的运算律：

$(A^T)^T = A$

$(\lambda A)^T = \lambda A^T$

$(A + B)^T = B^T + A^T$

矩阵的向量化是将矩阵表示为一个列向量。这里，vec是向量化算子。设$\mathrm{A} = \left[\mathrm{a_{ij}}\right]_{\mathrm{m\times n}}$则

$$\mathrm{vec(A)} = \left[\mathrm{a_{11}},\ \mathrm{a_{21}},\ \cdots,\ \mathrm{a_{m1}},\ \mathrm{a_{12}},\ \mathrm{a_{22}},\ \cdots,\ \mathrm{a_{m2}},\ \cdots,\ \mathrm{a_{1n}},\ \mathrm{a_{2n}},\ \cdots,\ \mathrm{a_{mn}}\right]^{\mathrm{T}}$$

6. 矩阵的行列式：记作$det(A)$，是一个将 n 阶方阵$A \in R^{n\times n}$映射到实数的函数。

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

行列式的意义：行列式等于矩阵**特征值的乘积**。可以看作有向面积或体积概念在 N 维欧氏空间中的推广，行列式的绝对值可以用来衡量矩阵参与矩阵乘法（线性变换）后向量空间体积扩大或缩小了多少。如果行列式是 0，那么向量空间至少沿着某一维完全收缩了，使其失去了所有的体积；如果行列式是 1，那么这个线性变换保持向量空间体积不变。

7. 常见的向量

全 1 向量：指所有值为 1 的向量。用$\underline{1}_{\mathrm{n}}$表示，n 表示向量的维数。$\underline{1}_{\mathrm{k}}=[1,\cdots,1]_{\mathrm{K}\times 1}$是 K 维的全 1 向量。

one-hot 向量：表示一个 n 维向量，其中只有一维元素为 1，其余元素都为 0。例如，在数字电路中，one-hot 是一种状态编码，指对任意给定的状态，状态寄存器只有 1 位为 1，其余位都为 0。

零向量：始点与终点重合，也就是重合点的向量。具有方向不确定性。因此零向量与任一向量重合。

单位向量：如果向量$\boldsymbol{a}$的 2 范数为 1，即满足：$\|\boldsymbol{a}\|_2 = 1$成立，则称为单位向量。

8. 常见矩阵

对称矩阵：是转置等于自己的任意 **n 阶方阵**A，即满足$\mathrm{A} = \mathrm{A}^T$。

对角矩阵：是一个主对角线之外的元素皆为 0 的矩阵。对角线上的元素可以为 0 或其它值。一个 **n 阶方阵**A是对角矩阵，则满足：

$$\mathrm{A_{ij}} = 0 \ \text{ if } \mathrm{i} \neq \mathrm{j} \quad \forall \mathrm{i,j} \in \{1,\cdots,\mathrm{n}\}$$

n 阶对角矩阵 A 也可以记为$diag(\boldsymbol{a})$，$\boldsymbol{a}$为一个 n 维向量，并满足：$a_{\mathrm{ii}} = a_{\mathrm{i}}$

n 阶对角矩阵$\mathrm{A} = diag(\boldsymbol{a})$和 n 维向量**b**的乘积为一个 n 维向量。

$$\mathrm{Ab} = diag(\boldsymbol{a})\mathrm{b} = \mathrm{a} \odot \mathrm{b}$$

其中⊙表示向量**点乘**，即$(\mathrm{a} \odot \mathrm{b})_{\mathrm{i}} = \mathrm{a_i b_i}$

单位矩阵是一种特殊的**对角矩阵**，其主对角线元素为 1，其余元素为 0。n 阶单位矩阵$\mathrm{I_n}$，是一个 n×n 的方阵。可以记为$\mathrm{I_n} = \mathrm{diag}(1,1,\cdots,1)$。

一个矩阵与单位矩阵的乘积等于其本身。$\mathrm{AI} = \mathrm{IA} = \mathrm{A}$

正交矩阵：对于一个 n 阶方阵 A，若矩阵的行向量之间相互正交，且行向量均为单位向量，即满足：$AA^T = A^T A = I$成立，则称方阵 A 是一个正交矩阵。若 A 是正交矩阵，必有$A^T = A^{-1}$成立。正交矩阵也是对称矩阵

正交矩阵的性质（略）

☆**9. 矩阵特征值和特征向量**：设A是 n 阶方阵，如果存在实数和 n 维非零向量 x，满足：

$$\mathrm{Ax} = \lambda \mathrm{x}$$

成立，则把λ称为方阵A的**特征值**，非零向量x称为方阵A对应特征值λ的**特征向量**。从几何角度来看，特征向量x的方向经过方阵A的线性变换后，结果向量λx与x保持在同一直线上，这时或者方向不变（$\lambda > 0$），或者方向相反（$\lambda = 0$），或者变为 0 向量。

☆**10. 矩阵的特征分解**：假设方阵 A 有 n 个线性无关的特征向量$\{x^{(i)}\}_{i=1}^{n}$，它们对应的特征值为$\{\lambda^{(i)}\}_{i=1}^{n}$，把 n 个特征向量作为列向量构成一个新方阵Q。用特征值作对角线元素构成一个对角矩阵

$$\Sigma = \mathrm{diag}([\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(n)}]^T)$$

则方阵 A 可以表示为**特征向量矩阵**Q和**特征值矩阵**Σ的分解形式：

$$\mathrm{A} = \mathrm{Q}\Sigma\mathrm{Q}^{-1} = \mathrm{Q}\Sigma\mathrm{Q}^{T}$$

其中Q为正交矩阵。一个 n 阶方阵A能够进行特征分解的充要条件是方阵 A 含有 n 个线性无关的特征向量。**矩阵分解**目的是把一个矩阵用一些较简单的子矩阵来表示，通过观察每个子矩阵可以得到矩阵本身的有用性质。通过矩阵的特征分解可以得到方阵 A 简单形式的等价矩阵Σ。

☆**11. 奇异值分解**（Singular Value Decomposition，简称 SVD）：一个$\mathrm{m} \times \mathrm{n}$的矩阵 A 的奇异值分解定义为：$\mathrm{A} = \mathrm{U}\Sigma V^T$，其中 U 和 V 分别是$\mathrm{m} \times \mathrm{m}$和 $\mathrm{n} \times \mathrm{n}$的正交矩阵，Σ为$\mathrm{m} \times \mathrm{n}$的对角矩阵，对角线上元素称为奇异值（Singular Value），一般按从大到小排列。
由于AA^T和$A^T A$都为对称矩阵，必有 n 个线性无关的特征向量。

$$\mathrm{AA}^T = \mathrm{U}\Sigma V^T V \Sigma \mathrm{U}^T = \mathrm{U}\Sigma^2 \mathrm{U}^T$$

$$\mathrm{A}^T\mathrm{A} = \mathrm{V}\Sigma U^T U \Sigma \mathrm{V}^T = \mathrm{V}\Sigma^2 \mathrm{V}^T$$

因此，U 和 V 分别为AA^T和$A^T A$的**特征向量**，因此，A 的非零奇异值为AA^T或$\mathrm{A}^T A$的非零特征值的平方根。

☆**12. 迹运算**

n 阶方阵 A 的所有主对角线的元素之**和**称为它的**迹**（Trace），记作：$\mathrm{tr(A)} = \sum_{i=1}^{n} a_{ii}$。矩阵转置的迹等于其自身的迹，即$\mathrm{tr(A)} = \mathrm{tr}(\mathrm{A}^T)$。

矩阵 A 的迹与矩阵 A 的 F 范数有紧密联系，满足$\|A\|_F = \sqrt{tr(AA^T)}$

矩阵 A 的迹与矩阵 A 的特征值有密联系，设$\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$为矩阵 A 的所有特征值，则满足：$\mathrm{tr(A)} = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$

13. 距离计算

余弦距离：$cos\,\theta = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$，弦夹角=余弦相似度，表示两个向量方向的差异，机器学习中借用这一概念来衡量**向量之间的相似性**。夹角越小，趋近于 0 度，余弦值越接近 1，它们的方向更加吻合则越相似。当两个向量的方向完全相反，夹角余弦取最小值-1，当余弦值为 0 时，两**向量正交**，夹角 90 度。

在二维空间中向量$A(x_1, y_1)$与向量$B(x_2, y_2)$的夹角余弦公式为：

$$cos\,\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

明可夫斯基距离：也被称为明式距离，它不仅仅是一种距离，而是将多个距离公式（曼

哈顿距离、欧式距离、切比雪夫距离）总结为一个公式。

首先假设两个 n 维向量$a(x_1,x_2,\cdots x_n)$与$b(y_1,y_2,\cdots y_n)$，对于这两个 n 维向量则有明式距离公式：

$$d_{12}=P\sqrt{\sum_{k=1}^{n}|x_i-y_i|^P}$$

当 P=2 时明式距离为欧式距离。

欧氏距离：两点间（直线）距离

三维空间两点$a(x_1,y_1,z_1)$与$b(x_2,y_2,z_2)$之间的欧式距离：

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$

两个 n 维向量 a 与 b 之间的欧式距离

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+\cdots+(x_n-y_n)^2}$$

$$=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

标准化欧氏距离：针对简单欧式距离的缺点做的改进，由于数据各维分量的分布不一致，先将各个分量都“标准化“到均值、方差相等。假设样本集 X 的均值为 m，标准差为 s，则 X 的“标准化变量”—标准化后的值=（标准化前的值−分量的均值）/分量的标准差。

$$X^*=\frac{X-m}{s}$$

公式描述：两个 n 维向量$a(x_1,x_2,\cdots x_n)$与$b(y_1,y_2,\cdots y_n)$之间的标准化欧式距离的公式：

$$d_{12}=\sqrt{\sum_{k=1}^{n}\left(\frac{x_i-y_i}{s_k}\right)^2}$$

当 P=1 时明式距离为曼哈顿距离。

曼哈顿距离：两个 n 维向量$a(x_1,x_2,\cdots x_n)$与$b(y_1,y_2,\cdots y_n)$之间的曼哈顿距离$d_{12}=\sum_{i=1}^{n}|x_i-y_i|$。

当 P=∞时，明式距离为切比雪夫距离。

切比雪夫距离：向量空间中的一种度量，两点之间的距离定义为其各坐标数值差的最大值。

二维平面两点$a(x_1,y_1)$与$b(x_2,y_2)$之间的切比雪夫距离$d_{12}=max(|x_1-x_2|,|y_1-y_2|)$。

两个 n 维向量$a(x_1,x_2,\cdots x_n)$与$b(y_1,y_2\cdots y_n)$之间的切比雪夫距离$d_{12}=max(|x_i-y_i|)$。

还可表示为$d_{12}=\lim_{k\to\infty}(\sum_{i=1}^{n}|x_i-y_i|)^{\frac{1}{k}}$

杰卡德相似系数与杰卡德距离的应用：可将杰卡德相似系数用在衡量样本的相似度上。

杰卡德相似系数：两个集合 A 和 B 的交集元素在 A、B 的并集中所占的比例，称为两个集合的杰卡德相似系数，用符号 J（A，B）表示。

$$J(A,B)=\frac{A\cap B}{A\cup B}$$

杰卡德相似距离：与杰卡德相似系数相反的概念，即杰卡德距离用两个集合中不同的元素占所有元素的比例来衡量两个集合的区分度。

$$J_\delta(A,B) = 1 - J(A,B) = \frac{A \cup B - A \cap B}{A \cup B}$$

汉明距离：两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数。换句话说，它就是将一个字符串变换成另外一个字符串所需要替换的个数。

3.4.2 微积分

以函数为研究对象，研究函数的极限、微分 (Differentiation)、积分 (Integration) 以及应用的数学分支。其中求导运算是一套关于函数变化率的理论。

1. 导数

对于定义域和值域都是实数域的函数$y = f(x)$，若$f(x)$在点 x_0 的某个邻域 $\triangle x$ 内，极限

$$f'(x_0) = \lim_{\triangle x \to 0} \frac{f(x_0 + \triangle x) - f(x_0)}{\triangle x}$$

存在，则称函数$y = f(x)$在点 x_0 处可导，导数为$f'(x_0)$。

若函数$f(x)$在其定义域包含的某区间内每一个点都可导，那么也可以说函数$f(x)$在这个区间内可导。定义函数$f'(x)$为函数$f(x)$的导函数，也称为导数。

函数$f(x)$的导数$f'(x)$也可记作$\nabla_x f(x)$，$\frac{\partial f(x)}{\partial x}$ 或$\frac{\partial}{\partial x} f(x)$。

2. 向量导数

对于一个 p 维向量$x \in R^p$,函数$y = f(x) = f(x_1, \cdots, x_p) \in R$，则 y 关于 x 的导数为

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_p} \end{bmatrix} \in R^p$$

对于一个 p 维向量$x \in R^p$，函数$y = f(x) = f(x_1, \cdots, x_p) \in R^q$，则 y 关于 x 的导数为

$$\frac{\partial f(x)}{\partial x} = \begin{matrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_q(x)}{\partial x_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_1(x)}{\partial x_p} & \cdots & \frac{\partial f_q(x)}{\partial x_p} \end{matrix} \in R^{p \times q}$$

3. 导数法则

加（减）法则：$y = f(x)$，$z = g(x)$则

$$\frac{\partial (y + z)}{\partial x} = \frac{\partial y}{\partial x} + \frac{\partial z}{\partial x}$$

乘法法则：$(uv)' = u'v + uv'$

（1）若$x \in R^p$，$y = f(x) \in R^q$，$z = g(x) \in R^q$，则

$$\frac{\partial y^T z}{\partial x} = \frac{\partial y}{\partial x} z + \frac{\partial z}{\partial x} y$$

（2）若$x \in R^p$，$y = f(x) \in R^s$，$z = g(x) \in R^t$，$A \in R^{s \times t}$和 x 无关，则

$$\frac{\partial y^T A z}{\partial x} = \frac{\partial y}{\partial x} A z + \frac{\partial z}{\partial x} A^T y$$

（3）若$x \in R^p$，$y = f(x) \in R$，$z = g(x) \in R^p$，则

$$\frac{\partial yz}{\partial x} = y \frac{\partial z}{\partial x} + \frac{\partial y}{\partial x} z^T$$

☆**链式法则**

（1） $x \in R^p$，$y = g(x) \in R^s$，$z = f(y) \in R^t$，则

$$\frac{\partial z}{\partial x} = \frac{\partial y}{\partial x} \cdot \frac{\partial z}{\partial x}$$

（2） 若$X \in R^{p \times q}$为矩阵，$Y = g(x) \in R^{s \times t}$，$z = f(Y) \in R$，则

$$\frac{\partial z}{\partial X_{ij}} = tr\left(\left(\frac{\partial z}{\partial Y}\right)^T \frac{\partial Y}{\partial X_{ij}}\right)$$

（3） 若$X \in R^{p \times q}$为矩阵，$Y = g(X) \in R^s$,$z = f(y) \in R$，则

$$\frac{\partial z}{\partial X_{ij}} = tr\left(\left(\frac{\partial z}{\partial Y}\right)^T \frac{\partial Y}{\partial X_{ij}}\right)$$

4. 常用函数及其导数

（1）标量函数及其导数

指示函数:

指数函数$I(x = c)$为

$$I(x = c) == \begin{cases} 1 & \text{if} \quad x = c \\ 0 & \text{else} \quad 0. \end{cases}$$

指数函数$I(x = c)$除了在 c 外，其导数为 0。

多项式函数：如果$f(x) = x^r$,其中 r 是非零实数，那么导数

$$\frac{\partial x^r}{\partial x} = r x^{r-1}$$

当 r=0 时，常函数的导数是 0。

指数函数：底数为 e 的指数函数$f(x) = \exp(x) = e^x$的是它本身。

对数函数：$\frac{\partial \log(x)}{\partial x} = \frac{1}{x}$

☆ **（2）向量函数及其导数**

$$\frac{\partial x}{\partial x} = \mathbf{I}$$

$$\frac{\partial \mathrm{Ax}}{\partial \mathrm{x}} = \mathbf{A}^{\mathbf{T}}$$

$$\frac{\partial \mathrm{x}^{\mathrm{T}}\mathrm{A}}{\partial \mathrm{x}} = \mathbf{A}$$

（3）按位计算的向量函数及其导数

假设一个函数 f(x)的输入是标量 x。对于一组 K 个标量 x_1，…，x_k，可以通过 f(x)得到另外一组 K 个标量 z_1，…，z_k，

$$z_{k=}f(x_k), \forall k = 1,\cdots, K$$

定义$x = [x_1, \cdots, x_k]^T$，$z = [z_1, \cdots, z_k]^T$，

$$z = f(x)$$

其中，f(x)是按位运算的，即$\big(f(x)\big)_i = f(x_i)$。

如果 f(x)的导数记为$f'(x)$。当这个函数的输入为 K 维向量$x = [x_1, \cdots, x_k]^T$时，其导数为一个对角矩阵。

$$\frac{\partial(x)}{\partial x} = \left[\frac{\partial f(x_j)}{\partial x_i}\right]_{K\times K} = \begin{bmatrix} f'(x_1) & 0 & \cdots & 0 & \\ 0 & f'(x_2) & \cdots & 0 & \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & & 0 & \cdots & f'(x_K) \end{bmatrix}$$

☆（4）**logistic 函数**

logistic 函数经常用来将一个**实数空间的数**映射到(0,1)区间，记为σ(x)

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

其导数为

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

当输入为 K 维向量$x = [x_1, \cdots, x_k]^T$时，其导数为

$$\sigma'(x) = \mathrm{diag}(\sigma(x) \odot (1 - \sigma(x)))$$

☆（5）**softmax 函数**

softmax 函数是将多个标量映射为一个概率分布。

对于 K 个标量$x_1, \cdots, x_k$，softmax 函数定义为

$$z_k = \mathrm{softmax}(x_k) = \frac{\exp(x_k)}{\sum_{i=1}^{k}\exp(x_i)}$$

可以将 K 个变量$x_1, \cdots, x_k$转换为一个分布：$z_1, \cdots, z_k$，满足

$$z_k \in [0,1], \ \forall k, \ \sum_{i=1}^{K} z_k = 1$$

当 softmax 函数的输入为 K 维向量 x 时，

$$\hat{z} = \mathrm{softmax}(x)$$

$$= \frac{1}{\sum_{k=1}^{K}\exp(x_k)}\begin{bmatrix}\exp(x_1)\\ \vdots \\ \exp(x_k)\end{bmatrix}$$

$$= \frac{\exp(x)}{\sum_{k=1}^{K} \exp(x_k)}$$
$$= \frac{\exp(x)}{\underline{1}_K^T \exp(x)}$$

其中，$\underline{1}_K = [1, \dots, 1]_{K\times 1}$是 K 维的全 1 向量。其导数为

$$\frac{\partial \ \mathrm{softmax}(x)}{\partial x}$$
$$= \frac{\partial\left(\frac{\exp(x)}{\underline{1}_K^T \exp(x)}\right)}{\partial x}$$
$$= \frac{1}{\underline{1}_K \exp(x)} \frac{\partial \exp(x)}{\partial x} + \frac{\partial\left(\frac{1}{\underline{1}_K^T \exp(x)}\right)}{\partial x} (\exp(x))^T$$
$$= \frac{\mathrm{diag}(\exp(x))}{\underline{1}_K \exp(x)} - \left(\frac{1}{(\underline{1}_K^T \exp(x))^2}\right) \frac{\partial(\underline{1}_K^T \exp(x))}{\partial x} (\exp(x))^T$$
$$= \frac{\mathrm{diag}(\exp(x))}{\underline{1}_K \exp(x)} - \left(\frac{1}{(\underline{1}_K^T \exp(x))^2}\right) \mathrm{diag}(\exp(x)) \underline{1}_K^T (\exp(x))^T$$
$$= \frac{\mathrm{diag}(\exp(x))}{\underline{1}_K \exp(x)} - \left(\frac{1}{(\underline{1}_K^T \exp(x))^2}\right) \exp(x) (\exp(x))^T$$
$$= \mathrm{diag}\left(\frac{(\exp(x))}{\underline{1}_K \exp(x)}\right) - \frac{\exp(x)}{\underline{1}_K^T \exp(x)} \cdot \frac{(\exp(x))^T}{\underline{1}_K^T \exp(x)}$$
$$= \mathrm{diag}\big(\mathrm{softmax}(x)\big) - \mathrm{softmax}(x)\mathrm{softmax}(x)^T$$

其中，$\mathrm{diag}(\exp(x))\underline{1}_K = \exp(x)$

3.4.3 概率论

概率论是集中研究概率和随机对象的数学分支，是研究随机性或不确定性等现象的学科，概率论主要研究对象为**随机事件、随机变量和随机过程**。概率论是深度学习的重要基础。

1. 事件

基本事件（单位事件）：再一次随机试验中可能发生且不能再细分的结果。

事件空间：在随机试验中可能发生的所有单位事件的集合，用 S 来表示。事件空间是由可数有限单位事件或者可数无限及不可数单位事件组成。

随机事件：事件空间 S 的子集，它由事件空间 S 中的单位元素构成，用大写字母表示。概率

2. 古典概率：如果一个随机试验所包含的单位事件是有限的，且每个单位事件发生的可能性均相等。

古典概率的特点：（1）样本空间的元素只有有限个。（2）实验中每个基本事件发生的可能性相同。

P(A)=构成事件 A 的元素数目/构成事件空间 S 的所有元素数目

3. 统计概率：大数概率

通过对事件进行 100 次、1000 次或者甚至 10000 次的前后相互独立的 n 次随机实验，针对每次试验均记录下绝对频率值和相对频率值，随着试验次数 n 的增加，会出现如下事实，**即相对频率值会趋于稳定**，它在一个特定的值上下浮动，也就是说存在一个极限值 P(A)，相对频率值趋向于这个极限值。这个极限值被称为统计概率表示为$P(A) = \lim_{n\to\infty} f_n(A)$

4. 概率公理

事件空间的概率值为 1。

互斥事件加法法则（可应用于可数个互斥事件的联集）

若 A $\cap B = \emptyset$　　$P(A \cup B) = P(A) + P(B)$

如果若干事件$A_1, A_2 \cdots, A_n \in S$每两两之间是空集关系，那么这些所有事件集合的概率等于单个事件的概率的和。

$$P(A_1 \cup \cdots \cup A_n) = \sum_{j=1}^{n} P(A_j)$$

对于事件空间 S 中的任意两个事件 A 和 B，有如下定理：

P(A∪ B)=P(A)+P(B)-P(A∩ B)

☆**乘法法则**：事件 A、B 同时发生的概率是 P(A∩ B)=P(A)·P(B|A)=P(B)·P(A|B)

无关事件乘法法则：两个不相关联的事件 A、B 同时发生的概率 P(A∩B)=P(A)·P(B)

☆**全概率**：n 个事件$H_1, H_2, \ldots, H_n$相互独立，且共同组成整个事件空间 S，这是 A 的概率可以表示为$P(A)$=$\sum_{j=1}^{n} P(A|H_j) \cdot P(H_j)$

3.4.4 统计学

1. 数据度量标准

平均值的优缺点

优点：可以用它来反映一组数据的一般情况，也可以用它进行不同数组的比较，以看出组与组之间的差别。

缺点：只能应用于数值型数据，不能用于分类数据和顺序数据。

中位数特点

一个数集中最多有一半的数值小于中位数，最多有一半的数值大于中位数。如果大于和小于中位数的数值个数均少于一半，那么数集中必有若干值等同于中位数。

众数

众数主要用于分类数据，也可以用于顺序数据和数值型数据。

众数的特点：

在离散概率分布中，众数是指概率质量函数有最大值的数据，也就是最容易取样的数据。在连续概率分布中，众数是指概率密度函数有最大值的数据，也就是概率密度函数的峰值。

在高斯分布中，众数位于峰值，和平均数、中位数相同。若分布是高度偏斜分布，则众数可能会和平均数、中位数有很大差异。

用众数代表一组数据，适合在数据量较多时使用，且众数不受极端数据的影响。

☆**期望**

在概率论和统计学中，“期望”为期望值的简称，是指在一个离散型随机变量实验中每次可能结果的概率乘以其结果的总和。$E(x) = \Sigma xp(X = x)$

☆**方差**

在概率论和统计学中，一个随机变量的方差描述的是他的离散程度，也就是该变量离其期望值的距离。方差是度量数据分散性的一种方法，是数据与均值距离的平方数的平值。设 X 为服从分布 F 的随机变量，如果 E[X]是随机变数 X 的期望值（平均数μ=E[X]）。则随机变量 X 或者分布 F 的方差为$Var(X) = \sigma^2 = \frac{\Sigma E(X-\mu)^2}{N}$

☆**标准差**

标准差是描述典型值与均值距离的一种方法。标准差越小，数值离均值越近。

标准分

标准分也叫 z 分数，是一种具有相等单位的量数。它是将原始分数与团体的平均数之差除以标准差所得的商数，是以标准差为单位度量原始分数离开平均数的分数之上多少个标准差，或是在平均数之下多少个标准差，它是一个抽象值，不受原始测量单位的影响，并可接受进一步的统计处理。z=(x-μ)/σ z 值的量代表着原始分数和母体平均值之间的距离，以标准差为单位计算，在原始分数低于平均值时，z 为负数，反之则为正数。

标准差的特点：标准分数是一种不受原始测量单位影响的数值，其除了能够表明原数据在其分布中的位置，还能对未来不能直接比较的各种不同单位的数据进行比较。

2. 概率分布

几何分布：n 重伯努利实验，前 k-1 次失败，第 k 次成功的概率。

二项分布：n 重伯努利实验，成功 k 次失败 n-k 次的概率

☆**正态分布（连续型）**：概率密度函数中包含两个可以设置的参数（均值和标准差），一种包含了多种好用特征的分布，用于各种研究。

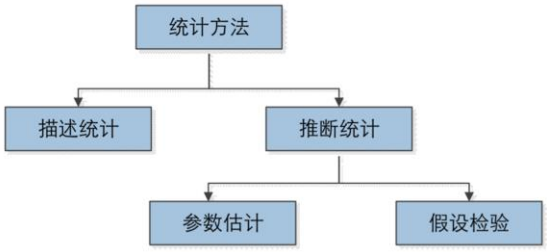
☆**中心极限定理**：特定条件下，大量统计独立的随机变量的**平均值的分布趋于正态分布**，因此可以通过调整均值和标准差来**用正态分布近似其他分布**。

泊松分布：描述单位事件内随机事件发生次数的分布。用单位时间或单位面积内随机事件发生次数作为参数。二项分布中的 n 足够大时可以用泊松分布近似。

3. 参数估计

参数估计：设有一个统计总体，总体的分布函数为$F(x;\theta)$，其中θ为未知参数，现从该总体抽样，得样本$X_1, X_2, \dots, X_n$，要依据该样本对参数θ作出估计$\hat{\theta}$，或估计θ的某个已知函数$g(\theta)$，这类问题称为**参数估计**。参数估计分为**点估计**和**区间估计**。

参数估计在统计方法中的地位：



点估计量：

已知某地区新生婴儿的体重$X \sim N(\mu, \sigma^2)$（μ，σ^2未知），随机抽查 100 个婴儿，得 100 个体重数据：10,7,5,6,6,5.2，…。而全部信息就由这 100 个数组成，据此我们应如何估计μ和σ呢？

我们需要构造出适当的样本的函数 $T(X_1, X_2, \dots, X_n)$，每当有了样本，就代入到该函数中算出一个值，用来作为μ的估计值。$T(X_1, X_2, \dots, X_n)$称为**参数**的**点估计量**。

问题是：使用什么样的估计量去估计μ？如何评价估计量的好坏？

- 可以用样本均值
- 也可以用样本中位数
- 或者是其他的统计量

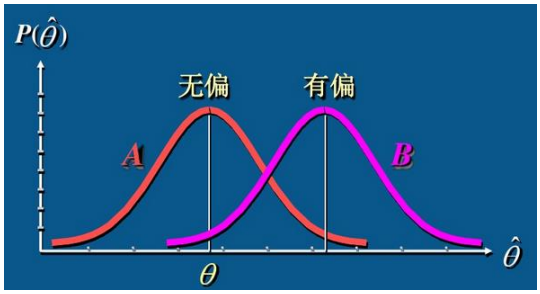
衡量统计量的标准

- **无偏性**：估计的偏差被定义为：

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

如果$bias(\hat{\theta})$=0，即$E(\hat{\theta}) = \theta$，那么估计量$\hat{\theta}$被称为是无偏的。

如果$\lim bias(\hat{\theta})$=0 ，即$\lim E(\hat{\theta}) = \theta$，$\hat{\theta}$被称为是渐进无偏的。

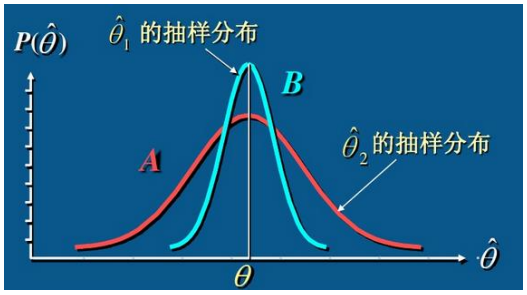


例：X_1，X_2，X_3是总体X的样本。下列统计量是否为总体均值μ的无偏估计量？谁更优？

$\widehat{\mu_1} = \frac{2}{5}X_1 + \frac{1}{10}X_2 + \frac{1}{2}X_3$，$\widehat{\mu_2} = \frac{1}{3}X_1 + \frac{3}{4}X_2 - \frac{1}{12}X_3$，

$\widehat{\mu_3} = \frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3$，$\widehat{\mu_4} = \frac{1}{5}X_1 + \frac{1}{10}X_2 + \frac{7}{10}X_3$.

- **有效性：**设$\widehat{\theta_1}$和$\widehat{\theta_2}$是θ的两个无偏估计量，若方差$D(\widehat{\theta_1}) < D(\widehat{\theta_2})$，称$\widehat{\theta_1}$比$\widehat{\theta_2}$有效。



- **一致性（相合性）**：在概率P的意义下，$\hat{\theta} \to \theta(n \to \infty)$，即对于$\forall \epsilon > 0$，有$\lim\limits_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon)$=0.

例：设总体X的期望$EX = \mu$，方差$DX = \sigma^2$，X_1，X_2，…，X_n是来自X的样本，$\widehat{\mu_1}$=$\frac{1}{n}\sum_{i=1}^{n} X_i$，

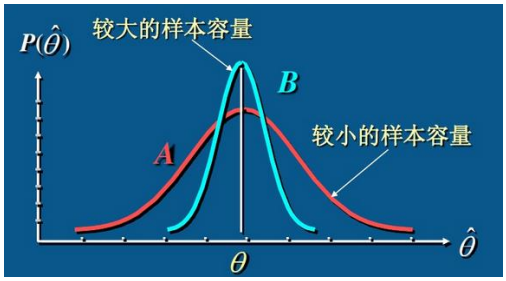
$\widehat{\mu_2}$=$\frac{1}{k}\sum_{i=1}^{k} X_i$，$k < n$，试证$\widehat{\mu_1}$，$\widehat{\mu_2}$都是$\mu$的无偏估计量，且$\widehat{\mu_1}$比$\widehat{\mu_2}$更有效。

证明：$E(\widehat{\mu_1}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \times n\mu = \mu$.

$E(\widehat{\mu_2}) = E\left(\frac{1}{k}\sum_{i=1}^{k} X_i\right) = \frac{1}{k} \times k\mu = \mu$.

$D(\widehat{\mu_1}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}$，$D(\widehat{\mu_2}) = \frac{\sigma^2}{k}$.

由于$n > k$，故$D(\widehat{\mu_1})$<$D(\widehat{\mu_2})$，所以估计量$\widehat{\mu_1}$更有效。



☆**最大似然估计**

最大似然估计原理：设X_1，X_2，…，X_n是取自总体X的一个样本，样本的联合密度或联合分布律为$f(x_1,x_2,\cdots,x_n;\theta)$。定义似然函数为：

$$L(\theta)=f(x_1,x_2,\cdots,x_n;\theta),$$

其中，$x_1,x_2,\cdots,x_n$是样本的观察值；$L(\theta)$看做参数θ的函数，它可作为θ将以多大可能产生样本值$x_1,x_2,\cdots,x_n$的一种度量。

最大似然估计法就是用使$L(\theta)$达到最大值的$\hat{\theta}$去估计θ。

$$L(\hat{\theta})=\max_{\theta}L(\theta),$$

称$\hat{\theta}$为θ的最大似然估计值。而相应的统计量$\hat{\theta}(X_1，X_2，\cdots，X_n)$称为$\theta$的最大似然估计量。

例：设X_1，X_2，…，X_n是取自总体$X\sim B(1,p)$的一个样本，求参数p的最大似然估计量。

解：似然函数：

$$L(p)=f(x_1,x_2,\cdots,x_n;p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}.$$

对数似然函数：

$$\ln L(p)=\sum_{i=1}^{n}x_i\ln(p)+(n-\sum_{i=1}^{n}x_i)\ln(1-p).$$

对p求导并令其为 0：

$$\frac{d\ln L(p)}{dp}=\frac{1}{p}\sum_{i=1}^{n}x_i-\frac{1}{1-p}(n-\sum_{i=1}^{n}x_i)=0.$$

得 $\hat{p}=\bar{x}$.从而p的最大似然估计量为：

$$\hat{p}(X_1，X_2，\cdots，X_n)=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}.$$

☆**贝叶斯定理（法则）**：贝叶斯定理提供了一种由$P(D)$、$P(h)$、$P(D|h)$计算后验概率$P(h|D)$的方法。公式如下：

$$P(h|D)=\frac{P(D|h)\cdot P(h)}{P(D)}$$

其中，D是观测数据样本，通常用n个属性的测量值组成的特征向量描述。h为某种假设，如数据样本D属于某个特定类C。$P(h\,|\,D)$ 是后验概率，或在条件D下，h的后验概率。$P(h)$是先验概率，或h的先验概率，$P(h)$独立于D。$P(D)$是D的先验概率。

极大后验概率假设

$P(h|D)$是假设h的后验概率，使$P(h|D)$最大化的假设h，称为极大后验（maximum a posteriori，MAP）假设。形式化地，MAP 假设h_{MAP}满足：

$$h_{MAP} = \underset{h\in H}{arg\max} P(h|D) = \underset{h\in H}{arg\max} \frac{P(D|h)P(h)}{P(D)} = \underset{h\in H}{arg\max} P(D|h)P(h)$$

在实际使用中，我们往往不能得到各个假设的先验概率，我们只能认为假设空间中的所有假设都是等可能的，此时，P(h)可认为是常数，有：

$$h_{MAP} = \underset{h\in H}{arg\max} P(D|h)P(h) = \underset{h\in H}{arg\max} P(D|h) = h_{ML}$$

其中，$P(D|h)$常被称为**给定h时数据D的似然度**，因此，使得$P(D|h)$最大的假设称为极大似然假设h_{ML}。

当假设空间中的假设都是等可能的假设时，h_{MAP}=h_{ML}。

4. 假设检验

假设检验是统计推断中用于检验统计假设的一种方法。而“统计假设”是可通过观察一组随机变量的模型进行检验的科学假说。一旦能估计未知参数，就会希望根据结果对未知的真正参数值做出适当的推论。假设检验的种类包括：t 检验、Z 检验、卡方检验、F 检验等。

☆5. 相关和回归

在概率论和统计学中，**相关**称为**相关系数或关联系数**，表示两个随机变量之间的**线性关系的强度和方向**。在统计学中，**相关**的意义用来**衡量两个变量相对于其相互独立的距离**。相关性即**变量之间的数学关系**，通过**散点图**上的点的独特构成模式，可以识别图上的各种相关性。如果**散点图**上的集合**呈直线分布**，则相关性为线性。存在正线性相关、负线性相关和无相关。

通过**相关系数判断**所求**最佳拟合线**与**数据的拟合度**，规则如下：

如果相关系数的绝对值越接近 1，则所求最佳拟合线的拟合度越高，可用于数据预测。

如果相关系数的绝对值越接近 0，则所求最佳拟合线的拟合度越低，不推荐用于进行预测。

线性回归和逻辑回归

在线性回归中，结果是连续的。它可以是无限数量的可能值中的任何一个。在逻辑回归中，结果只有有限数量的可能值。

相关和回归的联系

两者区别：

回归和相关都是研究两个随机变量相互关系的分析方法。相关分析研究两个变量之间相关的**方向和密切程度**。但是相关分析不能指出两个变量相互关系的具体形式，也无法从一个变量的变化来推测另个变量的变化关系。回归方程则是通过一定的数学方程来反映变量之间相互关系的具体形式，以便从一个已知量来推测另一个未知量。为估算预测提供个重要的方法。具体区别有：

(1)相关分析中变量之间处于平等的地位；回归分析中，因变量处在被解释的地位，自变量用于预测因变量的变化。

(2)相关分析中不必确定自变量和因变量，所涉及的变量可以都是随机变量；而回归分析则必须事先确定具有相关关系的变量中，哪个是因变量。一般来说，回归分析中因变量是随

机变量，而把自变量作为研究时给定的非随机变量。

(3)相关分析研究变量之间相关的方向和程度，但相关分析不能根据一个变量的变化来推测另一个变量的变化情况；回归分析是研究变量之间相互关系的具体表现形式，根据变量之间的联系确定一个相关的数学表达式，从而可以从已知量来推测未知量。

(4)对两个变量来说，相关分析只能计算出一个相关系数；而回归分析有时可以根据研究目的的不同建立两个不同的回归方程。

两者联系：

相关分析与**回归分析**是广义相关分析的两个阶段，两者有着密切的联系：

(1)相关分析是回归分析的基础和前提，回归分析则是相关分析的深入和继续。相关分析需要依靠回归分析来表现变量之间数量相关的具体形式，而回归分析则需要依靠相关分析来表现变量之间数量变化的相关程度。**只有当变量之间高度相关时**，进行回归分析寻求其相关的具体形式才有意义。如果在没有对变量之间是否相关以及相关方向和程度做出正确判断之前，就进行回归分析，则很容易造成“虚假回归”。

(2)由于相关分析只研究变量之间**相关的方向和程度**，不能推断变量之间相互关系的具体形式，也无法**从一个变量的变化来推测另一个变量的变化情况**。因此在具体应用过程中，只有**把相关分析和回归分析结合起来**，才能达到研究和分析的目的。

3.4.5 信息论

信息论是由香农发展的，用来找出**信号处理与通信操作的基本限制**，如数据压缩、可靠的存储和数据传输等。

☆**1. 信息熵**：是信息的一个关键度量，通常用**一条消息**中需要**存储或传输**一个符号的平均比特数来表示。熵衡量了预测随机变量的值时涉及的不确定度的量。其定义为离散随机事件的出现概率。一个系统越有序，信息熵越低，反之一个系统越混乱，信息熵就越高。

如果一个随机变量 X 的可能取值为 X={$x_1, x_2, \dots, x_n$},其概率分布为$P(X = x_i)$, i=1,2,⋯,n, 则随机变量 X 的信息熵定义为$H(X)$。

$$H(x) = -\sum_{i=1}^{n} p(x_i)\, log\, P(x_i) = \sum_{i=1}^{n} P(x_i) \frac{1}{logP(x_i)}$$

☆**2. 联合熵**：两个随机变量 X 和 Y 的联合分布可以形成联合熵，定义为联合自信息的数学期望，他是二维随机变量 X、Y 的不确定行的度量，用$H(X,Y)$表示：

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} P(x_i, y_j) logP(x_i, y_j)$$

☆**3. 条件熵**：用来衡量在已知随机变量 X 的条件下，随机变量 Y 的不确定性，在随机变量 X 发生的前提下，随机变量 Y 发生新带来的熵，定义为 Y 的条件熵，用$H(Y|X)$表示：$H(Y|X) = -\sum_{x,y} P(x,y)\, log\, P(y|x)$

由贝氏定理，我们有$p(x,y) = p(y|x)p(x)$,代入联合熵的定义，可以分离出条件熵，于是得到联合熵与条件熵的关系式：$H(Y|X) = H(X,Y) - H(X)$

☆**4. 相对熵（信息增益）**：描述两个概率分布 P 和 Q 差异的一种方法，记为 D(P||Q)。在信息论中，D(P||Q)表示当用概率分布 Q 来拟合真实分布 P 时，产生的信息损耗，其中 P 表示真实分布，Q 表示 P 的拟合分布。

$$\mathrm{D(P||Q)} = \sum_{i=1}^{N} P(x_i)\, log \frac{P(x_i)}{Q(x_i)}$$

☆**5. 互信息**：两个随机变量 X、Y 的互信息，X、Y 的联合分布和各自独立分布乘积的相对熵称为互信息，用 I(X,Y)表示。互信息是信息论里一种有用的信息度量方式。它可以看作一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不确定性。互信息是两个事件集合之间的相关性。

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\, log \frac{P(x,y)}{P(x)P(y)}$$

互信息、信息熵和条件熵之间存在以下关系：$H(Y|X) = H(Y) - I(X,Y)$

☆**6. 最大熵**：最大熵原理是概率模型学习的一个准则，学习概率模型时，在所有可能的概率分布中，满足约束条件的模型集合中选取熵最大的模型，熵最大的模型是最好的模型。

3.4.6 数值计算与最优化方法（规划）

数值计算与最优化理论是上世纪中期形成和发展起来的一门应用数学分支学科，其理论和方法愈来愈多，如**线性规划、非线性规划、动态规划**、排队论、对策论、决策论、**博弈论**等。广泛应用于机器学习、工程设计、生产管理、交通运输、国防等重要领域。而其中机器学习又是与最优化理论联系最紧密的学科之一，机器学习模型的训练，最终归结为最优化问题，也就是寻找学习模型最优的参数，使得模型的**误差损失函数**（目标函数）达到最小值，而寻找模型最优化参数的方法，称为**最优化方法**。本节内容将深入分析在**机器学习领域常用的最优化方法及其原理**，并分析不同的最优化方法之间各自的特点。

最优化理论是关于系统的**最优设计、最优控制、最优管理**问题的理论与方法。最优化，就是在一定的约束条件下，使系统具有所期待的最优功能的组织过程。是从众多可能的选择中做出最优选择，使系统的**目标函数**在约束条件下达到最大或最小。最优化是**系统化方法**的基本目标。优化方法有几个基本因素：建立系统**模型**、制定系统**评价**标准、**度量**系统执行方案目标的**代价**等。

数学优化问题也叫最优化问题，是指在一定的约束条件下，求解**目标函数** $f: \boldsymbol{R^{D_1\times D_2\times\cdots\times D_k}} \to \mathbf{R}$的最大值或最小值的问题。即给定一个能度量系统模型代价的**目标函数**（Objective Function），也叫**代价函数、损失函数**（Loss Function），寻找自变量θ（也叫参数）的一个取值$\theta^* \in D \subset \boldsymbol{R^{D_1\times D_2\times\cdots\times D_k}}$，使得对于所有$\theta \in D$，满足$f(\theta^*) \le f(\theta)$（称为目标函数最小化）或$f(\theta^*) \ge f(\theta)$（称为目标函数最大化），其中 D 为自变量$\theta$的一个约束集，也叫**可行域**，D 中的自变量值被称为优化问题的**可行解**。

依据目标函数的自变量为连续变量或离散变量，优化问题可分为**连续优化问题**或**离散优化问题**。根据是否有自变量的约束条件函数，将优化问题分为**无约束优化问题**和**约束优化问题**。

如果目标函数和所有的约束条件函数都为线性函数，则该问题为**线性规划问题**。如果目标函数或任意一个约束条件函数为非线性函数，则该问题为**非线性规划问题**。

在非线性规划问题中，有一类比较特殊的问题是**凸优化问题**，目标函数自变量θ的可行域 D 为凸集，即对于集合中任意两点的连线全部位于集合 D 内部，目标函数也必须为凸函数。**凸优化问题**是一种**特殊的约束优化问题**。并且**等式约束条件函数**为**线性函数**，**不等式约束条件函数为凸函数**。

优化求解算法：优化问题一般都可以通过**迭代**的算法来解决，通过猜测一个初始的估计θ_0，然后不断迭代产生新的估计$\theta_0, \theta_1, \theta_2, \cdots$，希望$\theta_t$最终收敛到期望的最优解$\theta^*$，一个好的

优化算法，应该是在一定的时间或空间复杂度下，能够快速、准确的找到最优解，同时好的优化算法，受初始猜测点θ_0的影响较小，通过迭代能稳定的找到最优解的邻域，然后迅速收敛到最优解。

优化算法中常用的迭代方法有**线性搜索**和**置信域**方法等。线性搜索的策略是寻找**方向**和步长，具体有**梯度下降法**，**牛顿法**和**共轭梯度法**等算法。

很多非线性优化问题会存在若干个局部最小值，其对应的解称为**局部最小解**。求解局部最小解一般是比较容易的，但很难保证其为**全局最小解**，对于**线性规划或凸优化问题**，**局部最小解就是全局最小解**。

确认一个点是否为局部最小解，通过比较它的领域内有没有更小的函数值是不现实的，如果目标函数$f(\theta)$是二次连续可微的，可以通过检查目标函数在点θ^*的梯度$\nabla f(x^*)$和汉森（Hessian）矩阵$\nabla^2 f(x^*)$来判断。

定理 1 如果x^*为局部最小解，并且函数 f 在x^*的邻域内一阶可导，则存在$\nabla f(x^*)$导数等于 0。函数 f 的一阶导数为零的点，也称为驻点或临界点，驻点不一定为局部最小解。

定理 2 如果x^*为局部最小解，并且函数 f 在x^*的邻域内二阶可微，且$\nabla f(x^*)=0$，则二阶导数$\nabla^2 f(x^*)$为半正定矩阵。

梯度下降法：也叫最速下降法，经常用来求解**无约束优化的极小值问题**。对于函数 f(x)，如果函数 f(x)在点x_t附近是连续可微的，那么 f(x)下降最快的方向是 f(x)在点x_t的梯度方向的反方向。

梯度下降法为一阶收敛算法，当靠近局部最小解梯度变小，收敛速度会变慢，并且可能以之字形的方式下降，如果目标函数为二阶连续可微函数，可以采用牛顿法，牛顿法为二阶收敛算法，收敛速度更快，但是每次迭代需要计算汉森矩阵的逆矩阵，复杂度较高。如果我们要求解一个最大值问题，就要向梯度正方向迭代进行搜索，逐渐接近函数的局部最大解，这个过程则称为梯度上升法。

拉格朗日乘数法是一种有效求解约束优化问题的优化方法。优化问题可以表示为，

$$\min_x f(x)$$

$$\text{s.t. } h_m(x)=0, m=1,\dots,M$$

$$g_n(x)\le 0, n=1,\dots,N$$

包括等式约束函数$h_m(x)$和不等式约束函数$g_n(x)$，x 的可行域为函数 f(x)和$h_m(x)$、$g_n(x)$定义域的交集。

等式约束优化问题，

不等式约束优化问题，

kkt 条件是拉格朗日乘数法在不等式约束优化问题上的泛化。当原问题是凸优化问题时，满足 kkt 条件的解也是原问题和对偶问题的最优解，在 kkt 条件中，称为互补松弛条件，如果最优解出现在不等式约束的边界上，如果出现在不等式约束的内部，互补松弛条件说明当最优解出现在不等式约束的内部，则约束失效。

3.5 实验 2：基于 Pytorch 的条件随机场模型（CRF）实现

3.5.1 马尔可夫链

随机过程是一个时间函数，其随着时间变化而变化，随机过程在每个时刻上的函数值是不确定的、随机的，即每个时刻上函数值按照一定的概率进行分布。

马尔可夫（随机）过程：在当前状态下，过程的未来状态与它的原始状态及过去状态无关，这种形式的随机过程就是马尔可夫过程。

马尔可夫链：在随机过程中，每个随机试验的当前状态，仅依赖于此前的过去状态，这

种随机过程就是马尔可夫链，该链可以当做一种数据的观察序列。

如果数据序列是字符串，字符串中每个字符的出现是随机独立的，就是**独立链**。由于实际场景中句子每个语言符号的出现概率不是相互独立的，如果每个字符的出现与前面字符相关，既不独立并且具有依赖性，就称之为**马尔可夫链**。

N 元马尔可夫链：考虑前一个语言符号对后一个语言符号出现概率的影响，这样得出的语言成分的链叫做**一阶马尔科夫链**，也叫**二元文法**；考虑前两个语言符号对后一个语言符号出现概率的影响，这样得出的语言成分的链叫做**二阶马尔科夫链**，也叫**三元文法**。以此类推。

马尔可夫链在数学上描述了自然语言句子的**生成过程**，是一个自然语言形式化的早期模型，后来 N 元文法的研究，都是建立在马尔可夫模型的基础上，马尔可夫链和隐马尔可夫模型都是有限自动机的扩充。注意马尔可夫链不能表示固有歧义的问题，当概率指派没有歧义时，马尔科夫链才有用。

有限自动机：是状态集合之间的转移集。如果说语法用来精确描述语言和它的结构，那么自动机是用来机械地刻画对输入字符串的处理过程。**加权有限状态机**中每个弧与一个概率相关，这个概率说明通过这个弧的可能性，且某个点出发的弧具有归一化的性质，即某点出发的弧概率之和为 1。

3.5.2 动态规划 Viterbi 算法

3.5.3 条件随机场模型（CRF）

条件随机场（Conditional Random Field，CRF）模型由 Lafferty 等人 2001 年提出，是一种判别式无向图模型。CRF 是用来**标记**和**切分**序列化数据的**统计模型**。在中文分词、命名实体识别、序列标注等问题上都取得了较好的结果。其主要思想来源于**隐马尔科夫模型**（HMM），加上了一些观察值（特征）。就是对给定的观察序列 X 和输出的标注序列 Y，通过定义给定一组输入变量 X 条件下输出变量 Y 的条件概率 P(Y|X)，而不是定义联合概率 P(X,Y)进行建模，CRF 可以根据给定的观察序列条件，捕捉全局信息，计算整个输出标记序列的联合概率。

HMM 模型认为当前时刻的观测值依赖于当前时刻的隐状态，当前时刻的隐状态又依赖上一时刻的隐状态。HMM 假设状态转移过程中当前状态只和前一状态有关，可以根据当前状态来定义下一个状态的分布，属于**生成式模型**。CRF 与之不同，属于**判别式模型**。

对于像中文分词、词性标注、命名实体识别等序列标注类结构化预测问题，一个最大的缺点就是由于其**输出独立性假设**，导致其不能考虑上下文的特征，限制了特征的选择。考虑相邻标签之间的相关性并**联合解码**给定输入句子的标签序列是非常有帮助的。例如，在词性标注（Part-Of-Speech, POS）问题中，形容词更可能后跟名词而不是动词，所以，充分考虑相邻标签之间的相关性对于序列标注问题是至关重要的。CRF 模型并不在每一个节点进行归一化，而是所有特征进行全局归一化，因此可以求得全局的最优值。

CRF 模型的优点是通过训练样本直接推断 P(Y|X)，其根据已经观察的序列特征向量 X 对输出变量 Y 进行预测，是对整个输出序列的优化，而不是某个时刻状态的优化，最后在训练预测 P(Y|X)基础上输出预测标签 Y。即 CRF 模型以观察序列 X 为全局条件，并且不对 X 做任何假设，下图为 CRF 的概率图模型。

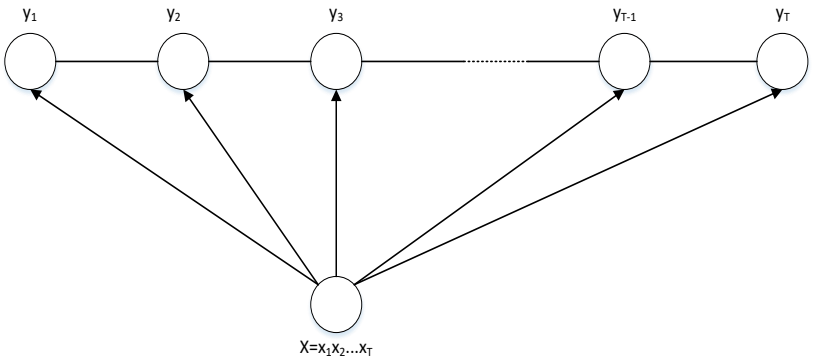


图　CRF 概率图模型

CRF 模型还兼具产生式模型的优点，考虑到输出标签 Y 上下文间的转移概率,以序列化形式进行全局参数优化和解码的特点，解决了其他判别式模型（如最大熵马尔科夫模型）难以避免的标签偏置问题。设定 n 为句子长度，$X=\{x_1,x_2,...,x_n\}$ 为观测序列，$Y=\{y_1,y_2,...,y_n\}$ 为相应的标签序列，那么对于观测序列 X，标签序列 Y 对应的条件概率如公式:

$$p(y\mid x)=\frac{1}{Z(x)}\exp(\sum_{i,j}\lambda_j t_j(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)) \quad (1)$$

$$Z(x)=\sum_{y}\exp(\sum_{i,j}\lambda_j t_j(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)) \quad (2)$$

其中，$Z(x)$ 是归一化因子，t_j 和 s_l 表示两种**特征函数**，对应的权重参数分别为为 λ_j 和 μ_l，j 和 l 为特征函数的个数。t_j 和 s_l 的取值为 0 和 1 两个值，以 t_j 为例，公式为（3）:

$$t_j(y_{i-1},y_i,x,i)=\begin{cases}1, y_{i-1},y_i,x\text{的取值符合条件;}\\ 0,\text{其他}\end{cases} \quad (3)$$

在自然语言处理中，对标记序列进行建模最常见的是链式结构条件随机场模型（linear-chain，CRF）。设输入序列 $X=x_0x_1\cdots x_T$，则标签状态序列 $Y=y_0y_1\cdots y_T$ 的条件概率 $P(Y\mid X)$ 为:

$$P(Y\mid X)=\frac{1}{Z_X}\exp\{\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,X,t)\} \quad (4)$$

其中 $f_k(y_{t-1},y_t,X,t)$ 表示作用于标签状态和输入序列的任意特征函数，$\lambda_k\geq 0$ 是特征函数 $f_k(y_{t-1},y_t,X,t)$ 的权重，表示该函数的贡献大小。特征函数 $f_k(y_{t-1},y_t,X,t)$ 的定义和参数权重 λ_k 的学习是 CRF 模型训练的核心。Z_X 是归一化因子，在模型训练过程中通过前向

后向算法进行求解。训练目标是最大化标注数据的条件似然：

$$L(\Lambda)=\sum_{m=1}^{M}\log(P(Y_m \mid X_m,\Lambda))+\log p(\Lambda) \qquad (5)$$

其中，$p(\Lambda)$ 是参数矩阵的先验概率。在测试阶段，该函数直接求解，无需计算归一化因子：

$$\begin{aligned}\arg\max_{Y} P(Y \mid X) &= \arg\max_{Y}\frac{1}{Z_X}\exp\left\{\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,X,t)\right\}\\ &= \arg\max_{Y}\left\{\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,X,t)\right)\right\} \qquad (6)\end{aligned}$$

模型在做预测时，最佳标签序列是通过动态规划算法（Viterbi 算法）搜索得到的，其中两个变量的计算公式如下：

$$\delta_t(y)=\max_{y^*}\left\{\delta_{t-1}(y^*)\exp\left(\sum_{k}\lambda_k f_k(y^*,y,X,t)\right)\right\} \qquad (7)$$

$$\varphi_t(y)=\arg\max_{y^*}\left\{\delta_{t-1}(y^*)\exp\left(\sum_{k}\lambda_k f_k(y^*,y,X,t)\right)\right\} \qquad (8)$$

程序代码文件: crf.py

3.6 实验 3：主题模型

主题模型可以从大规模语料库文本中生成表达这些文本的若干主题，将其表示为文本-主题分布，以此来达到聚类语料库信息的目的。早期的神经网络主题模型主要直接利用前馈神经网络构建主题模型。后续变分自编码器（Variational Auto-Encoder，VAE）被用于构建主题模型，目标是通过神经网络来推断主题模型中隐变量的后验分布。具体表示为对隐变量 z 的真实后验概率分布$p(z|x)$进行建模，通过贝叶斯法则可以表示为：$p(z|x)=\frac{p(x|z)p(z)}{p(x)}$。一般不直接求解真实后验分布$p(z|x)$，而是求解变分后验分布$q_\phi(z|x)$（$\varphi$为变分参数），并不断缩小$q_\phi(z|x)$与$p(z|x)$之间的差异来达到逼近$p(z|x)$的目的。$q_\phi(z|x)$通常选择易于计算的分布族，如正态分布（也叫高斯分布）。

这种方法本质上是将推断问题转换为优化问题，通过最小化变分分布$q_\phi(z|x)$和真实分布$p(z|x)$之间的 KL 散度来求解$p(z|x)$。它们间的 KL 散度定义为：

$$D_{\mathrm{KL}}\left[q_\phi(z|x)\parallel p(z|x)\right]=\sum_{z}q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z|x)}=\mathbb{E}_{q_\phi(z|x)}\left[\log\frac{q_\phi(z|x)}{p(z|x)}\right]$$

将上式中后验概率分布利用贝叶斯法则替换后得到：

$$
\begin{aligned}
D_{\mathrm{KL}}\left[q_\phi(z|x) \parallel p(z|x)\right] &= \mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x) - \log\frac{p(x|z)p(z)}{p(x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x) - \log p(x|z) - \log p(z) + \log p(x)\right] \\
&= \log p(x) - \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p(x,z) - \log q_\phi(z|x)\right]}_{\mathrm{ELBO}\ (\phi)}
\end{aligned}
$$

式中右侧括弧中部分称为$\log p(x)$的变分下界，记为 ELBO（Evidence Lower BOund）。在给定数据之后，x 的分布 p(x)可视作常数，因此最小化式左侧的 KL 散度的优化目标，等价于最大化 ELBO。使用插项的技巧，公式中 ELBO 可重写为：

$$
\begin{aligned}
\mathrm{ELBO}(\phi) &= \mathbb{E}\left[\log p(x,z) - \log p(z) + \log p(z) - \log q_\phi(z|x)\right] \\
&= \mathbb{E}[\log p(x|z)] - D_{\mathrm{KL}}\left[q_\phi(z|x) \parallel p(z)\right]
\end{aligned}
$$

因此，最大化 ELBO 等价于最大化式右侧，其中第一项为似然函数，将迫使解码器将生成的样本 $\hat{x}$ 尽可能还原为输入样本 x，通常采用交叉熵进行度量；第二项为关于主题 z 的分布的正则项，将迫使**变分后验分布**$q_\phi(z|x)$逼近**先验分布**$p(z)$。

假设将这两个分布都选定为多元正态分布，其中变分后验分布$q_\phi(z|x)$的均值和方差分别假定为$\mu(x)$和$\sigma^2(x)$，并假定其协方差阵$\Sigma(x)$为对角阵，先验分布$p(z)$则通常取标准正态分布$N(0,1)$。因此，实际的优化目标为：

$$
\begin{aligned}
D_{\mathrm{KL}}\left[q_\phi(z|x) \parallel p(z)\right] &= D_{\mathrm{KL}}[\mathcal{N}(\mu(x),\Sigma(x)) \parallel \mathcal{N}(0,1)] \\
&= \frac{1}{2}\left(\mathrm{tr}(\Sigma(x)) + \mu(x)^T\mu(x) - d - \log\det(\Sigma(x))\right)
\end{aligned}
$$

其中，d 为主题变量 z 的维数，tr()为矩阵迹运算，det()为矩阵的行列式值。

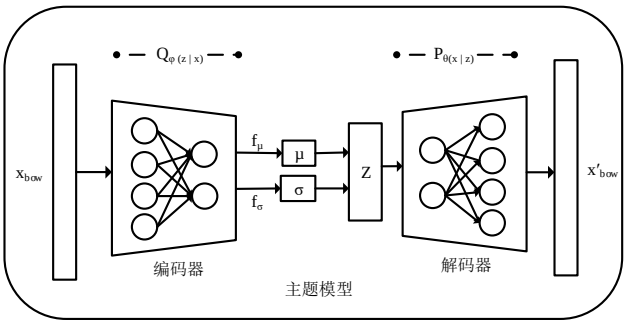


图 主题模型结构图

上图展示了基于 VAE 推断的主题模型架构，其中$q_\phi(z|x)$作为编码器，将**文档的词袋表示** BOW（Bag Of Word）x（是一个关于文档词表的多项式分布）映射为主题隐变量 z 服从的正态分布的均值$\mu(x)$和方差$\sigma^2(x)$，从该分布中采样得到$z\sim N\left(\mu(x),\sigma(x)\right)$。考虑到分布需满足规范性，因此从隐空间采样得到正态分布变量z后，还需要将z归一化才能作为主题分布，采取的方法是使z通过$Softmax$层，即：$\theta = soft\,max(f_\theta(z))$。最终则将$p_\theta(x|z)$用作解码器，通过主题隐变量 z 生成文档的词袋表示$\hat{x}$。

程序代码文件: ntm.py

小结

本讲介绍了机器学学、深度学习相关的线性代数、微积分、概率论、统计学、信息论、最优化理论等数学知识

实验作业

分组完成实验二和实验三的代码分析，撰写代码学习报告。学习报告按照以下提纲书写：包括：引言（问题背景描述）、模型及算法形式化描述、实现代码分析、测试结果展示、反思总结。