需要熟练掌握的算法理论基础

Hailiang Zhao @ ZJU.CS.CCNT
http://hliangzhao.me

Current Version: Mar 30, 2021

一 概率论、数理统计与信息论

1 正确区分概率空间、样本点、事件、事件的组合、随机变量等概念。

（1） （定义）离散概率空间$(\Omega, \Pr)$由(1)一个非空可数集合Ω和(2)定义在其上的概率质量函数（pmf）$\Pr: \Omega \to \mathrm{R}$组成。要求$\forall \omega \in \Omega, \Pr[\omega] \geq 0$且$\sum_{\omega \in \Omega} \Pr[\omega] = 1$。每一个$\omega \in \Omega$称为样本点。
同理可定义连续概率空间。如无特别指出，下文的内容定义在离散概率空间上。

（2） （定义）事件是非空可数集合Ω的一个子集（因此必然也是可数的），事件A发生的概率定义为：

$$\Pr[A] \coloneqq \sum_{\omega \in A} \Pr[\omega]$$

（3） 事件的组合通常用逻辑符号（而非集合运算符号）来描述，如$A \wedge B$、$A \vee B$、$\neg A$等。
（定义）如果$A \wedge B = \emptyset$，那么我们称事件A和B是不相交的（disjoint as sets）；
（定义）事件A和B是独立的（independent）当且仅当$\Pr[A \wedge B] = \Pr[A] \cdot \Pr[B]$；
（定义）事件的集合$\{A_i | i \in I\}$是相互独立的（fully/mutually independent）当且仅当

$$\Pr\left[\bigwedge_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} \Pr[A_i]$$

（定义）事件的集合$\{A_i | i \in I\}$是 k-wise independent 如果$\{A_i | i \in I\}$的每一个由k个事件组成的子集都是相互独立的；当$k = 2$时，我们称$\{A_i | i \in I\}$是成对独立的（pairwise independent）。若对于任意的k均有$\{A_i | i \in I\}$是k-wise independent，则$\{A_i | i \in I\}$是相互独立的。
（定义）对于任意两个事件A和B且$\Pr[B] > 0$，则在B发生的条件下A发生的概率为

$$\Pr[A|B] = \frac{\Pr[A \wedge B]}{\Pr[B]}$$

（性质）组合事件的一些 identities：

- Union bound：对于任意事件$A_1, A_2, \dots, A_n$有

$$\Pr\left[\bigwedge_{i=1}^{n} A_i\right] \leq \prod_{i=1}^{n} \Pr[A_i]$$

- Disjoint Union：如果$\{A_1, A_2, \dots, A_n\}$是成对独立的，则

$$\Pr\left[\bigvee_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \Pr[A_i]$$

- The principle of inclusion-exclusion：对于任意有穷事件集合$\{A_1, A_2, \dots, A_n\}$，有

$$\Pr\left[\bigvee_{i=1}^{n} A_i\right] = 1 - \sum_{I \subseteq \{A_1, A_2, \dots, A_n\}} (-1)^{|I|} \Pr\left[\bigwedge_{i=1}^{n} A_i\right]$$

- Independent union：如果$\{A_1, A_2, \dots, A_n\}$是相互独立的，则

$$\Pr\left[\bigvee_{i=1}^{n} A_i\right] = 1 - \prod_{i=1}^{n}(1 - \Pr[A_i])$$

- Bayes' Theorem：对于任意事件A和B，如果$\Pr[A] \neq 0, \Pr[B] \neq 0$，则根据条件概率的定义可得

$$\Pr[A|B]\Pr[B] = \Pr[A \wedge B] = \Pr[B|A]\Pr[A]$$

- Law of total probability：设$\{B_n : n = 1, 2, \dots\}$是概率空间$\Omega$的有限或可数无限的分割，且每个集合可数，则对任意事件A有：

$$\Pr[A] = \sum_{n} \Pr[A \wedge B_n] = \sum_{n} \Pr[A|B_n]\Pr[B_n]$$

（4） （定义）随机变量X是一个从Ω到某个数值集合V的一个映射/函数（$X(\omega \in \Omega) = x \in V$）。随机变量既不随机，也不是一个变量。显然$\Pr[X = x] = \Pr[A \coloneqq \{\omega \in \Omega | X(\omega) = x\}]$。

（5） 定义概率累计函数 CDF 如下：

$$\mathrm{CDF}_X(X \leq x) = \Pr[X \leq x] = \Pr[A \coloneqq \{\omega \in \Omega | X(\omega) \leq x\}]$$

（6） （定义）期望是定义在随机变量上的：

$$E[X] = \sum_{x} x \cdot \Pr[X = x]$$

若X为任意整数随机变量，则

$$E[X] = \sum_{x \geq 1} (\Pr[X \geq x] - \Pr[X \leq -x])$$

（通过等价代换$\Pr[X = x] = \Pr[X \geq x] - \Pr[X \geq x + 1]$证明）

（定义）事件A发生的情况下随机变量X的条件期望：

$$E[X|A] \coloneqq \sum_{x} x \cdot \Pr[X = x|A] = \sum_{x} \frac{x \cdot \Pr[X = x \wedge A]}{\Pr[A]} = \sum_{x} \frac{x \cdot \Pr[\{\omega \in A | X(\omega) = x\}]}{\Pr[A]}$$

（7） （定义）如果对于任意的x和y，事件$\{\omega | X(\omega) = x\}$和事件$\{\omega | Y(\omega) = y\}$是独立的，则随机变量$X$和$Y$是独立的。同理可定义一组随机变量$X_1, X_2, \dots, X_n$相互独立。

（性质）如果定义在实数域的随机变量X和Y是独立的，则

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

反之不一定成立。

（性质）如果定义在实数域的一组随机变量$X_1, X_2, \dots, X_n$相互独立，则

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]$$

反之不一定成立。

（8） （性质）不论$X_1, X_2, .., X_n$是否独立（any kinds of independent），总有

$$E\left[\sum_{i=1}^{n} (\alpha_i \cdot X_i)\right] = \sum_{i=1}^{n} (\alpha_i \cdot E[X_i])$$

2 再谈随机变量的独立性：

（1） （定义）对于两个随机变量，若知道其中一个变量的取值不会对知道另一个变量的分布有任何帮助，那么就称这两个随机变量相互独立。形式化地，若对于任意的x和y，$\Pr[X = x \wedge Y = y] =$

$\Pr[X=x]\cdot\Pr[Y=y]$或$\Pr[X=x|Y=y]=\Pr[X=x]$（其实是一-1-(7)中第一个公式按定义展开），则X和Y独立。

注意此处$X=x$这种写法是不严谨的，实际上应该是$\{\omega\in\Omega|X(\omega)=x\}$。

（2） （性质）如果随机变量X和Y独立，则$E[X\cdot Y]=E[X]\cdot E[Y]$；且对于任意给定函数$f$，都有$f(X)$和$f(Y)$独立。

（3） （定义）一组随机变量$X_1,X_2,\dots,X_n$相互独立当且仅当对于任意的$x_1,x_2,\dots,x_n$有

$$\Pr\left[\bigwedge_{i=1}^{n}(X_i=x_i)\right]=\prod_{i=1}^{n}\Pr\left[X_i=x_i\right]$$

上式其实是一-1-(7)中第二个公式按定义展开。且对于任意给定函数f，都有$\{f(X_1),f(X_2),\dots,f(X_n)\}$相互独立。

3 一种常见的问题是：已知随机变量X的 pdf 且$Y=f(X)$，求解随机变量Y的 pdf。利用 pdf 和 CDF 之间的转换关系以及$Y=f(X)$这个等价代换即可求解。

4 需要记住一些常用的概率分布的随机变量的定义、pmf/pdf 以及期望的计算。包括均匀分布、伯努利分布、二项分布、几何分布、泊松分布、指数分布。

5 推导泊松分布和指数分布的 pdf：

（1） 泊松分布：将一个时间段切分成等大小的时间片，每个时间片内事件只有“发生”和“不发生”两种情形（伯努利分布），泊松分布的随机变量就是该二项分布在时间片个数趋于无穷时事件发生的个数；

将该随机变量记为X，事件发生的概率为p，则$P(X=k)=lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{1-k}$。因为二项分布的期望值$E[X]=np$，不妨记为$\mu$，则$p=\frac{\mu}{n}$，带回上式可得$P(X=k)=lim_{n\to\infty}\frac{\mu^k}{k!}e^{-\mu}$。

（2） 指数分布：若某事件在给定时间段内发生的次数服从泊松分布，那么指数分布的随机变量就是该事件前后两次发生的时间间隔。

引入时间参数t：令$\mu=\lambda t$，当$t=1$时表示一个单位时间段内事件发生的次数，$t=\frac{1}{2}$则是半个单位时间段内事件发生的次数。此时$P(X=k,t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}$。令$Y=$两个事件发生的时间间隔，则$Y>t$的概率等价于$t$时间内事件发生 0 次的概率。所以

$$P_Y(Y\le t)=1-P_Y(Y>t)=1-P(X=0,t)=1-e^{-\lambda t}$$

因为t是定义在非负数上的，所以$t<0$时$P_Y(Y\le t)=0$。基于此 CDF 可以写出 pdf。

6 指数分布（连续）和几何分布（离散）具有无记忆性：

$$P(X=b|X>a)=P(X=b-a)$$

事件发生的概率并不会伴随着等待的时间变长而增加，事件第$n+1$次发生与第1次发生概率一样，不会因为已经发生了n次而改变。这一结论与人类的本能“赌徒心理”正好相反。

7 使用贝叶斯定理求解问题的关键：正确地识别问题中的条件概率。

8 学会使用概率论解决实际问题：

（1） 对应任意硬币（可能被做假，不妨假设正面落下的概率为p），通过如下方法总能产生公平的抛硬币效果：

```
VonNeumannCoin():
  x ← BiasedCoin()
  y ← BiasedCoin()
```

```
if x ≠ y
    return x
else
    return VonNeumannCoin()
```

Proof: 若停机，则必然有$x \neq y$，所以$\Pr[x = 0 \wedge y = 1 | x \neq y] = \Pr[x = 1 \wedge y = 0 | x \neq y] = \frac{pq}{2pq} = \frac{1}{2}$。令随机变量$T$为 *BiasedCoin*()被调用的次数，则

$$E[T] = E[T|x \neq y] \cdot \Pr[x \neq y] + [T|x = y] \cdot \Pr[x = y]$$

其中$\Pr[x \neq y] = 2pq$，$E[T|x \neq y] = 2$（因为结束了），$\Pr[x = y] = 1 - 2pq$，$E[T|x = y] = 2 + E[T]$（递归调用），所以可求得$E[T] = \frac{1}{pq}$。

（2） 假设有n种不同的卡，每次买一包干脆面可以抽出一种卡。假设买了n包干脆面，可以抽出多少张不同的卡？为了收集到全部种类的卡，至少要买几包干脆面？

对应第i种卡，定义随机变量$X_i \in \{0, 1\}$表示我们是否拥有它。定义随机变量X表示我们拥有的、不同种类的卡的个数，显然有$X = \sum_{i=1}^{n} X_i$，所以有$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \Pr[X_i = 1]$。在$n$次购买中，我们没有抽到第$i$种卡的概率为$\left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e}$（因为每次购买抽出某种卡的概率服从 iid），所以$E[X] = \frac{n}{e} \approx 0.632n$。

设$T(n)$为收集了n种卡所购买的干脆面的包数。将购买行为划分为n个阶段，第i个阶段在抽到第i种卡之后停止。用$T_i(n)$表示第i个阶段购买的干脆面的包数。则$T(n) = \sum_{i=1}^{n} T_i(n)$。在第$i$个阶段，抽到一张新卡的概率为$\frac{n-(i-1)}{n}$，此时$T_i(n)$实际上对应的几何分布的随机变量，根据几何分布的期望可知$E[T_i(n)] = \frac{n}{n-i+1}$，所以

$$E[T(n)] = \sum_{j=1}^{n} \frac{n}{j} = n(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n})$$

后面这个几何级数满足$\ln(n+1) \leq \sum_{j=1}^{n} \frac{1}{j} \leq \ln(n) + 1$（借助每段函数$\frac{1}{j}$的上下界函数的积分来计算，具体步骤参见二-6），所以至少要买$\Theta\big(n\log(n)\big)$包。

9 对于给定的随机变量，和期望值相差给定距离的取值发生的概率是多少？

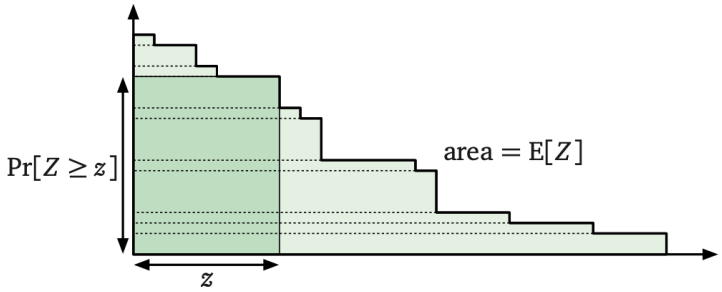
随着对随机变量的独立性的要求越来越高，接下来给出的估计逐步精确。

（1） 马尔科夫不等式（Markov's Inequality）：

令X是非负、整数随机变量，则

$$\forall a \in R^+ : P(X \geq a) \leq \frac{E(X)}{a}$$

Proof 1:



根据上图显然有$z \cdot \Pr[Z \geq z] \leq E[Z]$。

Proof 2：

因为Z的任意取值$z \geq 0$且$a > 0$，所以$z \neq 0$时有$\frac{z}{a} \geq 1$。所以$\Pr[Z \geq a] = \int_a^\infty f(z)dz \leq \int_a^\infty \frac{z}{a} f(z)dz$。

所以$\frac{E[Z]}{a} = E\left[\frac{Z}{a}\right] = \int_0^\infty \frac{z}{a} f(z)dz = \int_0^a \frac{z}{a} f(z)dz + \int_a^\infty \frac{z}{a} f(z)dz \geq \int_a^\infty \frac{z}{a} f(z)dz = \Pr[Z \geq a]$。Q.e.d。

（推论）基于马尔科夫不等式，可以得到距离期望一定距离的随机变量的取值发生的概率的上限：

$$\Pr\left[X \geq (1+\delta)E[X]\right] \leq \frac{1}{1+\delta}$$

（2） 切比雪夫不等式（Chebyshev's Inequality）：

对于一组指示（随机）变量$X_1, X_2, \ldots, X_n \in \{0,1\}$，令$p_i = E[X_i] = \Pr[X_i = 1]$。定义$X = \sum_i X_i$，且令$\mu = E[X] = \sum_i p_i$。如果$X_1, X_2, \ldots, X_n$成对独立，则

$$\Pr[(X-\mu)^2 \geq z] < \frac{\mu}{z}$$

Proof:

$\forall i$定义$Y_i = X_i - p_i$，定义$Y = \sum_i Y_i = X - \mu$。则$E[Y^2] = E\left[\sum_{i,j} Y_i \cdot Y_j\right] = \sum_i E\left[Y_i^2\right] + \sum_{i \neq j} Y_i \cdot Y_j = \sum_i E\left[Y_i^2\right] + 0$（因为 pairwise independent）。对$\sum_i E\left[Y_i^2\right]$展开可得$\sum_i E\left[Y_i^2\right] = \sum_i (1-p_i)^2 \cdot p_i + (-p_i)^2 \cdot (1-p_i) = \sum_i p_i(1-p_i) < \sum_i p_i = \mu$，所以$E[Y^2] < \mu$。代入马尔科夫不等式即可得$\Pr[Y^2 \geq z] < \frac{E[Y^2]}{z} < \frac{\mu}{z}$。Q.e.d。

（推论）$\forall \Delta, \delta \in \mathrm{R}^+$，有

$$\Pr[X \geq \mu + \Delta] < \frac{\mu}{\Delta^2} \qquad \Pr[X \geq (1+\delta)\mu] < \frac{1}{\delta^2 \mu}$$

$$\Pr[X \leq \mu - \Delta] < \frac{\mu}{\Delta^2} \qquad \Pr[X \leq (1-\delta)\mu] < \frac{1}{\delta^2 \mu}$$

（令$z = \Delta^2$、令$\Delta = \delta\mu$）

（推论）将随机变量$X_1, X_2, \ldots, X_n$解读为n次重复独立试验，即$\Pr[X_i = 1] = p$，此时$\mu = np$。令$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$为前$n$次重复独立试验的样本均值，则可以得到弱大数定理：

$$\lim_{n \to \infty} \Pr\left[\left|\overline{X}_n - p\right| \geq \epsilon\right] = 0$$

（另一种形式）：对于随机变量X，定义$u = E(X), \sigma = \sqrt{Var(X)}$，则

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}, k > 0$$

Proof:

定义$Y = |X - \mu| > 0$，则由马尔科夫不等式可得$\Pr[|X - \mu| \geq a] < \frac{E[|X-\mu|]}{a}$。两边同时平方一下得

$$\Pr[|X - \mu| \geq a] \cdot \Pr[|X - \mu| \geq a] < \frac{E[|X - \mu|] \cdot E[|X - \mu|]}{a^2}$$

$X - \mu$的本质是样本的测量误差，每次测量都服从独立同分布（其实是正态分布，一-18 给出了解释），对于给定的$f(\cdot) = |\cdot|$仍然有$|X - \mu|$服从独立同分布。所以$E[|X - \mu|] \cdot E[|X - \mu|] = E[(X - \mu)^2] = \sigma^2$（随机变量独立的性质）。对于前后两次测量，事件$\{\omega \in \Omega | |X(\omega) - \mu| \geq a\}$与事件$\{\omega \in \Omega | |X(\omega) - \mu| \geq a\}$独立，所以$\Pr[|X - \mu| \geq a] \cdot \Pr[|X - \mu| \geq a] = \Pr[|X - \mu| \geq a \wedge |X - \mu| \geq a] = \Pr[|X - \mu| \geq a]$。所以

$$\Pr[|X - \mu| \geq a] < \frac{\sigma^2}{a^2}$$

令$a = \frac{k}{\sigma}$即可得到结论。Q.e.d。

切比雪夫不等式比马尔科夫不等式更加精确。

（3） Higher Moment Inequalities：

切比雪夫不等式成立的条件是要求$X_1, X_2, \ldots, X_n$成对独立。如果$X_1, X_2, \ldots, X_n$满足$2k$-wise independent，则

$$\Pr[(X - \mu)^k \geq z] = O\left(\frac{\mu^k}{z}\right)$$

$O(\cdot)$中的隐藏参数与k有关。

（推论）$\forall \Delta, \delta \in \mathrm{R}^+$，有

$$\Pr[X \geq \mu + \Delta] = O\left(\left(\frac{\mu}{\Delta^2}\right)^k\right) \qquad \Pr[X \geq (1 + \delta)\mu] = O\left(\left(\frac{1}{\delta^2\mu}\right)^k\right)$$

$$\Pr[X \leq \mu - \Delta] = O\left(\left(\frac{\mu}{\Delta^2}\right)^k\right) \qquad \Pr[X \leq (1 - \delta)\mu] = O\left(\left(\frac{1}{\delta^2\mu}\right)^k\right)$$

（令$z = \Delta^2$、令$\Delta = \delta\mu$）

（4） 切尔诺夫界（Chernoff Bounds）：

（引理）如果$X_1, X_2, \ldots, X_n$相互独立，则$\forall \alpha \geq 1, E[\alpha^X] \leq e^{(\alpha-1)\mu}$。

Proof:

显然$E[\alpha^{X_i}] = p_i\alpha + (1 - p_i) = (\alpha - 1)p_i + 1$，因为$1 + t \leq e^t$恒成立，所以$E[\alpha^{X_i}] \leq e^{(\alpha-1)p_i}$。根据随机变量$X_1, X_2, \ldots, X_n$相互独立的性质可以得到

$$E[\alpha^X] = E[\alpha^{X_1} \cdot \ldots \cdot \alpha^{X_n}] = \sum_i E[\alpha^{X_i}] \leq e^{(\alpha-1)\mu}$$

（切尔诺夫界-upper tail）如果$X_1, X_2, \dots, X_n$相互独立，则$\forall x \geq \mu, \Pr[X \geq x] \leq e^{x-\mu}\left(\frac{\mu}{x}\right)^x$。

Proof:

当固定住x且满足$\frac{x}{\mu} \geq 1$时，函数$t \to \left(\frac{x}{\mu}\right)^t$单调增，所以$\Pr[X \geq x] = \Pr\left[\left(\frac{x}{\mu}\right)^X \geq \left(\frac{x}{\mu}\right)^x\right]$。根据马尔科夫不等式有$\Pr\left[\left(\frac{x}{\mu}\right)^X \geq \left(\frac{x}{\mu}\right)^x\right] \leq \frac{E\left[\left(\frac{x}{\mu}\right)^X\right]}{\left(\frac{x}{\mu}\right)^x}$。根据引理可得$E\left[\left(\frac{x}{\mu}\right)^X\right] \leq e^{x-\mu}$，所以$\Pr[X \geq x] \leq e^{x-\mu}\left(\frac{\mu}{x}\right)^x$。

（切尔诺夫界-lower tail）如果$X_1, X_2, \dots, X_n$相互独立，则$\forall x \leq \mu, \Pr[X \leq x] \leq e^{x-\mu}\left(\frac{\mu}{x}\right)^x$。

Proof:

固定住x且满足$\frac{x}{\mu} \leq 1$，则$\Pr[X \leq x] = \Pr\left[\left(\frac{x}{\mu}\right)^X \geq \left(\frac{x}{\mu}\right)^x\right]$。剩余证明和上文一样。

（推论 1）将$\frac{\mu}{x}$替换为α则有

$$\forall \alpha > 1\text{: } \Pr[X \geq x] \leq \frac{e^{(\alpha-1)\mu}}{\alpha^x}$$
$$\forall \alpha < 1\text{: } \Pr[X \leq x] \leq \frac{e^{(\alpha-1)\mu}}{\alpha^x}$$

可验证上述二式在$\alpha = \frac{\mu}{x}$时取等。

（推论 2）$\forall \Delta, \delta \in \mathrm{R}^+$，有

$$\Pr[X \geq \mu + \Delta] \leq e^{\Delta}\left(\frac{\mu}{\mu+\Delta}\right)^{\mu+\Delta} \qquad \Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

$$\Pr[X \leq \mu - \Delta] \leq e^{-\Delta}\left(\frac{\mu}{\mu-\Delta}\right)^{\mu-\Delta} \qquad \Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

（分别令$x = \mu + \Delta$和$\mu - \Delta$、令$\Delta = \delta\mu$）

（推论 3）进一步地，还有更为松散的上界：

$$\Pr[X \geq (1+\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{3}\right)$$
$$\Pr[X \leq (1-\delta)\mu] \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$$

Proof:

对函数$f(x) = \ln(x+1), x \geq -1$在$x = 0$处展开，得到$f(x) \geq f(0) + f'(0)(x-0) + \frac{f''(0)}{2}(x-0)^2$，

代入可得$\ln(1+x) \geq x - \frac{1}{2}x^2$。将$x$替换为$-\delta$并限定$\delta \in [0,1]$得到$\ln(1-\delta) \geq -\delta + \frac{\delta^2}{2}$。代回推论2中右下方公式即可得本推论的公式2。此外，又因为$\frac{1}{\ln(1+x)} \leq \frac{1}{x\left(1-\frac{1}{2}x\right)} = \frac{1}{x} + \frac{1}{2-x}$，将$x$替换为$\delta$并限定$\delta \in [0,1]$可得$\frac{1}{\delta} + \frac{1}{2-\delta} \leq \frac{1}{\delta} + \frac{1}{2}$，所以得到$\frac{1}{\ln(1+\delta)} \leq \frac{1}{2} + \frac{1}{\delta}$。代回推论2中右上方公式即可得本推论的公式1。

10 理解多元概率分布：

（1） 理解排列数和组合数：

n个不同元素的排列数：$P_n^k = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$；

n个不同元素的组合数：在排列数的基础上除去同一个数字重复出现的排列数：$C_n^k = \frac{P_n^k}{k!} = \frac{n!}{k! \cdot (n-k)!}$。

（2） 多元概率分布就是随机变量变成了随机向量。下面谈一谈典型的多元概率分布：多项分布和多元正态分布。

多项分布：就是二项分布推广到多元的情形。

设一个袋子中很多球，一共有K种颜色。从袋子中有放回地取出n个球（保证每次取样服从独立同分布），令$X = (X_1, \ldots, X_k)$，其中X_k表示颜色为k的球的个数，那么X服从多项分布：

$$P((X_1, \ldots, X_K) = (x_1, \ldots, x_K) | \mu_1, \ldots, \mu_K) = \frac{n!}{x_1! \ldots x_K!} \mu_1^{x_1} \ldots \mu_K^{x_K}$$

其中μ_k为每次抽取的球的颜色为k的概率，$\sum_{k=1}^K x_k = n$。$\frac{n!}{x_1! \ldots x_K!}$是因为要从$n$个球的排列数中去除同一个颜色的球重复的排列数。

Gamma 函数被定义为$\Gamma(z) = \int_0^\infty \frac{t^{z-1}}{\exp(t)} dt$，是阶乘扩展到实数域的版本（可通过$\Gamma(z+1) = (z+1) \cdot \Gamma(z)$验证）。将$\frac{n!}{x_1! \ldots x_K!}$替换为 Gamma 函数则有

$$P((X_1, \ldots, X_K) = (x_1, \ldots, x_K) | \mu_1, \ldots, \mu_K) = \frac{\Gamma\left(\sum_{k=1}^K (x_k + 1)\right)}{\sum_{k=1}^K \Gamma(x_K + 1)} \mu_1^{x_1} \ldots \mu_K^{x_K}$$

多元正态分布：

$$P((X_1, \ldots, X_K) = (x_1, \ldots, x_K) | \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right)$$

其中$\mu = (\mu_1, \ldots, \mu_K)$为各个随机变量的均值，$\Sigma$为协方差矩阵。$\Sigma$作为对称阵是一定可以正交对角化的，故$\Sigma^{-1}$必然存在（后面会详细谈到如何理解正交对角化）。若$\Sigma = \sigma^2 I$（各维随机变量相互独立且方差相同），则是各向同性多元正态分布。

11 从极大似然估计的角度理解贝叶斯公式：

$$P(A|B) = P(A) \frac{P(B|A)}{P(B)}$$

$P(A)$是原本A发生的概率（先验），$\frac{P(B|A)}{P(B)}$是新信息带来的调整（其中$P(B|A)$是似然；$P(B)$是标准化项，让$P(A|B)$仍在0到1之间），$P(A|B)$是调整后A发生的概率（后验）。

在一-1-(3)最后我提到了全概率公式$\Pr[A]=\sum_n \Pr[A\wedge B_n]=\sum_n \Pr[A|B_n]\Pr[B_n]$。该公式还可以表示为

$$\Pr[A]=E[\Pr[A|N]]$$

此处N是任意随机变量，该式在连续情况下也成立。同样地，该式可以理解为

“A的先验概率等于A的后验概率的事前期望值”

12 理解最大似然估计（MLE）和贝叶斯推断：

（1） 将某一事件发生的概率称为这个事件的参数，通过证据（已经发生了的观测数据），对该参数进行推断的过程就是似然，通过“最大化证据发生的概率”找到最有可能的参数，就是最大似然估计。

（2） 为什么“最大化证据发生的概率”是合理的？这是因为我们朴素地认为：

一个合理的参数应该尽可能地让真实发生的事件（观测数据）的概率最大。

（3） 参数本身也可以看成是一个随机变量，它也具备自己的 pdf。在没有证据之前，假设参数服从某一个分布，我们将这个分布称为该参数的先验分布。经过观测数据的修正得到的新的分布，称之为该参数的后验分布。贝叶斯推断就是“先验分布 + 观测数据 → 后验分布”这个过程。

人们为了简化贝叶斯推断过程，喜欢让参数的先验分布和后验分布是同一个分布（这个分布当然是手工构造出来的），该分布的系数经过观测数据的洗礼得到了修正。满足这样条件的先验分布和后验分布被称为共轭分布。例如，第 7 条提到的多项分布，其参数的共轭分布是 Dirichlet 分布：$P(u_1,\dots,u_K|\alpha)=\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\alpha_1)\cdot\dots\cdot\Gamma(\alpha_K)}\mu_1^{\alpha_1}\dots\mu_K^{\alpha_K}$。为什么参数的共轭分布和原始随机变量（向量）的分布如此相似？这是因为参数的共轭分布的参数α就是原分布的随机变量（向量）。

（4） 观测数据越多（证据越多），参数的分布就越集中，修正的效果就越好。这个结论可在 Höffding's inequality 上观察到：

$$\forall\varepsilon\geq 0, P\left(\left|\hat{\theta}-\theta^*\right|\geq\varepsilon\right)\leq 2\exp\{-2N\varepsilon^2\}$$

该不等式常用于执行 Probably Approximate Correct（即 PAC Learning，用于计算满足给定条件的最少样本数）。该不等式其实就是一-9-(4)-推论三。

（5） 以抛硬币为例，我们很自然地会假设正面落下的概率为$\frac{1}{2}$，这就是先验。我们之所以会瞬间想到这个先验是因为我们作为现代人接受了各式基础教育。这个先验在几万年前的原始人脑海里是绝不存在的。有了这个先验，我们不断地抛硬币获取观测数据，根据这些观测数据来修正先验。如果硬币没有造假，那么在观测数据量极为庞大的情况下，后验无限逼近$\frac{1}{2}$。如果硬币造假了，正面落下的次数远多于反面，那么后验就会被修正到一个远小于$\frac{1}{2}$的位置。以上这个例子中，参数的分布是连续分布（Beta 分布）。如果举的例子中参数的分布是离散分布（比如分类问题）可能会更好理解。在观测数据足够多的情况下，再差的先验都可以被很好地修正，先验的存在感微乎其微。传统的统计学习算法强烈依赖于假设空间（先验）的选取，正是因为观测样本无法给出足够合理的修正，因而它们又被称为“小样本学习”。

13 执行贝叶斯推断（先验 → 最大似然估计 → 最大后验近似）：

（1） 二项分布：以抛硬币为例，设参数为θ，表示正面落下的概率（Head）。

先验：

假设θ服从如下先验分布（我们称之为 Beta 分布）

$$P(\theta)=\frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)}\sim Beta(\theta|\beta_H,\beta_T)=\frac{\Gamma(\beta_H+\beta_T)}{\Gamma(\beta_H)\Gamma(\beta_T)}\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}$$

其中β_H,β_T是先验分布的参数。

似然：

抛硬币数次，得到了一组观测数据D（e.g. $\{H,T,H,H,T,\dots\}$），其中 Head 个数为α_H，Tail 个数为α_T。则可以得到似然

$$P(D|\theta)=\theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

之所以是累乘是因为我们假设每次采样都服从 i.i.d.。（根据 MLE 可以得到在观测数据D的修正下，$\hat{\theta}=\frac{\alpha_H}{\alpha_H+\alpha_T}$。）

后验：

根据贝叶斯公式，后验$P(\theta|D)$满足

$$P(\theta|D)\propto P(\theta)P(D|\theta)\sim Beta(\theta|\beta_H+\alpha_H,\beta_T+\alpha_T)$$

所以后验仍然服从 Beta 分布，只是 Beta 分布参数发生了变化。因此先验分布和后验分布共轭。

此时对真实参数θ^*的计算应该是这样的：

$$\theta^*=E[\theta]=\int_0^1 P(\theta|D)\theta d\theta$$

参数θ的每个取值发生的概率选取的是后验概率。

如果$P(\theta|D)$的分布已知（本案例中是已知的，就是 Beta 分布），那么硬着头皮计算这个期望就完事了。但是如果$P(\theta|D)$的分布未知，我们可以用最大化后验的方式去估计后验，即

$$\theta^*\approx \text{argmax}_\theta P(\theta|D)$$

这种估计方法就是最大后验近似（Maximum A Posteriori approximation，MAP）。对于本问题，可以验证，MAP 的结果是

$$\text{argmax}_\theta P(\theta|D)=\frac{\alpha_H+\beta_H-1}{\alpha_T+\beta_T+\alpha_H+\beta_H-2}$$

当观测数量$N=\alpha_T+\alpha_H\to\infty$时，$\text{argmax}_\theta P(\theta|D)\to\frac{\alpha_H}{\alpha_H+\alpha_T}=\text{argmax}_\theta P(D|\theta)$，而后者就是最大似然。这就验证了上文陈述过的结论：当观测样本量足够庞大的时候，最大似然估计≈最大后验近似，先验的作用微乎其微。

（2） 多项分布：

二项分布推广到多元随机变量就得到了多项分布，先验、似然、后验都长得很像。多项分布非常强大，可用于描述所有离散随机事件。

先验（我们称之为 Dirichlet 分布）：

$$P(\theta)\sim Dir(\{\theta_1,\dots,\theta_K\}|\{\beta_1,\dots,\beta_K\})=\frac{\Gamma\left(\sum_{k=1}^K\beta_k\right)}{\prod_{k=1}^K\Gamma(\beta_k)}\prod_{k=1}^K\theta_k^{\beta_k-1}$$

似然：

根据观测数据$D=\{X_1,\dots,X_N\}=\{N_1,\dots,N_K\}$，其中$N_k$表示第$k$个颜色的球出现的个数。则似然为

$$P(D|\theta)=\prod_{k=1}^K\theta_k^{N_k}$$

（根据 MLE 可以得到在观测数据D的修正下，$\widehat{\theta_k}=\frac{N_k}{\sum_{k=1}^K N_k}$。）

后验：

$$P(\theta|D)\propto P(\theta)P(D|\theta)\sim Dir(\{\theta_1,\dots,\theta_K\}|\{\beta_1+N_1,\dots,\beta_K+N_K\})$$

最大后验近似：

$$\forall k = 1, \dots, K\colon\ \theta_k^* = \int \theta_k Dir\left(\{\theta_1, \dots, \theta_K\} | \{\beta_1 + N_1, \dots, \beta_K + N_K\}\right) d\theta_k \approx \mathrm{argmax}_{\theta_k} P(\theta_k | D)$$

$$= \frac{\beta_k + N_k}{\sum_{k=1}^{K} \beta_k + \sum_{k=1}^{K} N_k}$$

同样可验证当$\sum_{k=1}^{K} N_k \to \infty$时$\mathrm{argmax}_{\theta_k} P(\theta_k | D) \to \mathrm{argmax}_{\theta_k} P(D | \theta_k)$。

（3） 正态分布：可按照同样的套路描述，此处不再展开。

对正态分布的一组观测数据做 MLE 可得$\mu = \sum_{i=1}^{N} x_i$，$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$。这验证了为什么用样本均值估计期望是可行的，同时也验证了为什么可以用$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$近似$\sigma^2$。我们后面还会再谈及这个问题。

（4） 因为每次采样要求服从独立同分布，所以似然永远都是每次采样概率的累乘。这就是为什么在做 MLE 的时候总会引入对数，就是为了将累乘转化为累加，从而更方便地执行求导运算。

14 样本方差的计算公式为$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$，为什么分母是$n-1$?

（1） 中心极限定理：从均值为μ，方差为σ^2的任意一个总体中抽取样本量为n的样本，当n充分大的时候，样本均值$\bar{X}$的抽样分布近似服从均值为μ，方差为$\frac{\sigma^2}{n}$的正态分布。

（2） 问题 1：已知随机变量X的期望为μ，则方差定义为$\sigma^2 = E[(X-\mu)^2]$。当方差σ^2未知时，为什么可以用$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$近似$\sigma^2$?

求解思路：证明$E[S^2] = \sigma^2$。借助中心极限定理，该式表明随机变量$Y = S^2$服从均值为σ^2的正态分布，因此选择$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$来近似$\sigma^2$是合理的。

（3） 问题 2：当期望μ也未知时，用均值$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$来近似$\mu$，为什么$S^2$的计算公式变成了

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)^2$$

i.e. 分母变成了$n-1$?

求解思路：作差$\sum_{i=1}^{n}(X_i - \bar{X})^2 - \sum_{i=1}^{n}(X_i - \mu)^2)$，可发现二者（期望上）相差$\frac{1}{n}\sigma^2$。

15 理解协方差与相关系数的定义由来：

（1） 如何描述两个随机变量X和Y之间的相关性？观察围起来的矩形面积差的符号！
设随机变量X和Y的n个样本为$(x_1, y_1), \dots, (x_n, y_n)$，则面积差之和为$A = \sum_{i \le j}(x_i - y_i)(x_j - y_j)$，一共有$C_n^2$个面积差。若$A < 0$则负相关，这表明二者的增长方向是相反的。同理可判断$A = 0$则不相关，$A > 0$则正相关。

（2） 计算面积和太麻烦！直接使用期望——

$$A' = \sum_{i=1}^{n}(x_i - \mu_X)(x_i - \mu_Y)$$

但是每个样本点发生的概率不一样，最好赋予不同的权重，所以还要修改。干脆直接定义在期望上好了：

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

这就是随机变量X和Y的协方差。对Cov的符号判断和对A的一致。当样本个数给定时（例如n个），则$\Sigma_{n\times n}$是协方差矩阵，且$\Sigma_{n\times n}(X_i,Y_j) = (X_i-\mu_X)(Y_j-\mu_Y)$。

（3） 和方差一样，协方差有单位，且无法比较X和Y哪一个和Z更相关。因此引入相关系数去掉量纲：

$$\rho_{XY} = \frac{Cov(X,Y)}{Var(X)\times Var(Y)}$$

这就是随机变量X和Y的协方差系数。

16 点估计与区间估计（以正态分布为例，假设随机变量$X\sim N(\mu,\sigma^2)$）

（1） 点估计：做一次抽样$X_1,\dots,X_n$，取样本均值$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$就是我们对期望的一次点估计；

（2） 区间估计：点估计最大的问题在于我们没法判断哪一次估计$\hat{\mu}_1,\dots,\hat{\mu}_m$更好。把点替换为区间会怎样？尝试分析一下——

设随机变量$M = \frac{1}{n}\sum_{i=1}^{n} X_i$表示每次大小为$n$的抽样的均值，则根据中心极限定理可知$M\sim N(\mu,\frac{\sigma^2}{n})$。

可以算出，以μ为中心，面积占比 95%的区间为$P(\mu-1.96\frac{\sigma}{n}\le M\le\mu+1.96\frac{\sigma}{n})$。也就是说，对于一次采样均值$\hat{\mu}$，如果我们以$1.96\frac{\sigma}{n}$为半径作区间，就有 95%的概率包含$\mu$（可做图验证）。我们称 95%为置信水平。显然，构造区间估计需要随机变量的方差已知。

17 理解假设检验：

以抛硬币为例，假设硬币上没有作假，“抛几次，现实世界碰一碰？”就是检验该假设。定义随机变量X=“抛n次，其中正面落下的次数”。例如$n=10$，正面落下 9 次，则 P 值$P(X\ge 9)\approx 0.01$（亦可检测双侧 P 值$P(X\ge 9\wedge X\le 1)$）。如将显著水平定为 0.05，则$P(X\ge 9)<0.05$，说明这个硬币很明显造假了（i.e.假设显著不成立）。

18 为什么用算术平均值作为测量/观测的结果是合理的？

（1） 这个问题触及了现代概率论的核心。它和以下几个问题等价：

“最小二乘法的理论依据是什么？”

“为什么测量误差符合正态分布？”

“如何推导正态分布的 pdf？”

（2） 让我们层层递进。

对于真值y^*，将我们对其的估计记为y。假设测量了n次，每次观测结果记为$y_1,\dots,y_n$，则总的误差平方为

$$\epsilon = \sum_{i=1}^{n}(y-y_i)^2$$

如果误差是随机的，那么就应该围绕着真值上下波动，有正有负（假设我们的观测是无偏的）。[这其实就是大数定理——服从独立同分布的样本均值收敛于期望值]那么能够让ϵ取值最小的y就应该是最接近y^*的值。求解该最优化问题（令$\frac{d\epsilon}{dy}=0$即可），可以得到

$$y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

这就是观测的算术平均值。这个求解思路正是最小二乘法（MSE 最小化）。用最小二乘法做多项

式拟合只需要把上述y替换成$f(x)=\sum_{i=1}^{n+1}a_i x^{n+1-i}$并依次对$a_i$求偏导即可。

（3） 如果最小二乘法是对的，那么用算术平均值作为测量/观测的结果就是合理的。可是，为什么最小二乘法就对呢？

我们可以用 MLE 来尝试回答这一问题。设每次一观测值x_i和真值x之间存在一个误差ϵ，假设误差服从概率分布$p(\epsilon)$, 则 n次观测会得到一个联合概率分布，记为

$$L(x)=p(x-x_1)\cdot p(x-x_2)\cdot \ldots \cdot p(x-x_n)$$

根据 MLE，$x=\text{argmax}_x L(x)\leftrightarrow \frac{dL(x)}{dx}=0$，又因为最小二乘法认为$x=\frac{1}{n}\sum_{i=1}^{n}x_i$，联立

$$\begin{cases}\dfrac{dL(x)}{dx}=0\\ x=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i\end{cases}$$

解得$p(\epsilon)=\frac{1}{\sqrt{(2\pi)}\sigma}\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$（这个结果是高斯最先发现的。这个过程很有趣，仅需要高中数学知识，用到的技巧很有全国高中数学联赛那味儿，务必自己推导）。这意味着测量误差符合正态分布！（其实那个时候人们还不知道这个 pdf 对后世有着怎样巨大的影响。后来人们直接把 Gaussian 作为 Normal Distribution 的同义词，用于纪念高斯大神的巨大贡献。我这里的因果顺序和历史上知识发现的顺序是反的）

（4） 如果测量误差符合正态分布(也就是概率分布$p(\epsilon)$)是可接受的，那么最小二乘法就是可接受的。可是，为什么测量误差符合正态分布呢？？（套娃中）

后来，拉普拉斯发现，若误差可视为微小量的叠加，则误差服从正态分布。这就是现代概率论历史上最重要的结论——中心极限定理的前身。（套娃结束）

（5） 中心极限定理的雏形是人们为了求解“二项分布中距离期望一定距离的事件发生的概率近似为多少”这一问题而发现的。后来又被推广到任意分布。而观测误差和最小二乘法的相关研究是从天文学引出的问题。后世的统计学三大分布无不依赖于正态分布而构造。

19 理解线性回归的本质：

（1） 回归就是根据输入的观测数据$(x_i,y_i)|_{i=1,\ldots,n}$（$x_i$可以是向量），通过最小二乘法学习目标函数$f(x)$的参数。如果目标函数$f(x)$来自如下假设空间：$f(x)=\sum_{j=1}^{k}\omega_j\cdot h_j(x)$，其中$\omega_j$和$h_j$均可以是向量（但是需要维度相同，这样点乘才合法），那么就是线性回归。$\omega_j|_{j=1,\ldots,k}$就是目标函数$f(x)$的参数，而$h_j(x)|_{j=1,\ldots,k}$则是基函数。上文已经详细描述了最小二乘法的合理性，此处不在展开。

当输入$x_i\in \mathrm{R}$，$h_i(x)=x^i$时，假设空间就是全体多项式的集合。此时线性回归问题变成了多项式拟合问题。基函数的选取是多种多样的。后面我们会看到，傅里叶级数也是单位正交基的线性组合。

（2） 线性回归的重要应用是主成分分析（PCA）。和一般的回归问题不同的是，PCA 不仅需要学习参数，还学要学习基函数。PCA 的目标是用更少的基函数来尽可能地保留原数据最多的信息量。有许多方法可以求解 PCA，以后会撰写新的文档深入分析。

20 理解熵及相关的概念：

（1） 熵被拿来衡量一个随机事件的不确定性。

接下来的内容将针对离散随机变量展开描述。

对一个随机变量进行编码，自信息$I(x)$是$X=x$的信息量/编码长度（选择对数函数只是为了描述

“事件发生的概率越大，则编码越容易、所包含的信息量越少”这一现象。你当然可以选择别的，只要大家认同即可）：

$$I(x) = -\log\big(P(X=x)\big) = -\log P(x)$$

而熵则是平均编码长度（也是最优编码长度）：

$$H(X) = E[I(X=x)] = -\sum_x P(x)\log P(x)$$

一个确定的事件，对其编码熵为零；一个不确定的事件（即随机事件），不确定的程度越大熵就越大。显然，均匀分布的熵最大。

（2） 由此可定义联合熵和条件熵：

$$\text{联合熵：}\ H(X,Y) = -\sum_x\sum_y P(x,y)\log P(x,y)$$

$$\text{条件熵：}\ H(X|Y) = -\sum_x\sum_y P(x,y)\log P(x|y)$$

（3） 互信息描述了“当一个随机变量已知时，另一个随机变量的不确定性减少了多少”：

$$I(X:Y) = I(Y;X) = \sum_x\sum_y P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

显然$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$。

（4） 交叉熵与 KL 散度：对于分布为p的随机变量，$H(p) = H(X)$是其最优编码长度，而交叉熵则是“按照概率分布为q的最优编码对真实分布为p的信息进行编码”的长度：

$$H(p,q) = E_p\big[-\log\big(q(x)\big)\big] = -\sum_x P(x)\log\big(q(x)\big)$$

相应地，概率分布q来近似p进行编码所造成的信息损失量就是 KL 散度：

$$D_{KL}(p||q) = H(p,q) - H(p) = -\sum_x P(x)\log\left(\frac{p(x)}{q(x)}\right) \geq 0$$

二 微积分、复数与优化

1 理解微积分的本质：

（1） 现在，不要再将导数理解为“瞬时变化率”。取而代之地，请严格记住其定义：

$$\frac{df(x)}{dt} = \lim_{dt\to 0}\frac{f(x+dt)-f(x)}{dt}$$

这是因为dt虽然趋近于 0 但并非 0，我们乘dt为无穷小量，本身有着极为严格的定义。对无穷小的错误理解将会导致无数的悖论。

所有的求导规则均从上式推演而来。作为一种观念上的总结，它们可以让我们在计算导数时无需每次都按照定义对无穷小量进行操作。完整的微分列表参见

https://en.wikipedia.org/wiki/Differentiation_rules

（2） 微分的严格定义：

设函数$y = f(x)$在某区间内有定义，x_0及$x_0 + \Delta x$在此区间内，若函数增量$\Delta y = f(x_0 + \Delta x) - f(x_0)$可表示为$\Delta y = A\cdot\Delta x + o(\Delta x)$，其中$A$是不依赖于$\Delta x$的常数，则函数$y = f(x)$在$x_0$处可微。

$A \cdot \Delta x$是$f(x)$在x_0处相对于自变量增量Δx的微分，记作dy。

微分是对差分的近似，以直代曲。因此微分又被称为“线性近似”。

（3） 一个函数在某一点的全微分是该函数在该点附近关于其全部自变量的最佳线性近似，这是单变量函数的微分在多变量函数上的推广。

（4） 积分 $= \lim_{n\to\infty} \sum$微分。

2 理解切线和导数的本质：切线是割线组成的数列的极限，导数是割线斜率的极限。

3 掌握三大微分中值定理（注意条件：闭区间上连续，开区间上可导）。

4 区分插值与拟合：

（1） 插值：函数精确经过每一个数据点；

（2） 拟合：采用最小二乘法的策略，做误差平方和最小化。并不要求精确经过每一个点。

5 推导多项式插值与泰勒展开式：根据观测数据$(x_0, y_0), (x_1, y_1), \dots, (x_n, .y_n)$，确定$f(x) = \sum_{i=0}^{n} a_i x^i$的各个系数，保证$f(x)$经过每一个观测数据。常规的计算方法就是带入表达式，求解线性方程组。弊端是每新增一个数据点全部系数a_i都要重新算。

（1） 牛顿插值法：每新增一个观测数据，只需计算相关的部分“均差”即可，计算可以增量式进行，已经付出的算力没有浪费——

设$f_1(x)$经过$\big(x_0, f(x_0)\big)$和$\big(x_1, f(x_1)\big)$并长成这个样子：

$$f_1(x) = f(x_0) + b_1(x - x_0)$$

带入$\big(x_1, f(x_1)\big)$解得$b_1 = \frac{f(x_1)-f(x_0)}{x_1-x_0}$。从而得到

$$f_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

设$f_2(x)$经过$\big(x_i, f(x_i)\big)|_{i=0,1,2}$并长成这个样子：

$$f_2(x) = f_1(x) + b_2(x - x_0)(x - x_1)$$

带入$\big(x_2, f(x_2)\big)$解得$b_2 = \frac{\frac{f(x_2)-f(1)}{x_2-1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2-x_0}$。从而得到

$$f_2(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) + \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}(x - x_0)(x - x_1)$$

进一步得到$f_3(x)$，…，$f_n(x)$。最后的$f_n(x)$就是我们想要的$f(x)$。

定义均差简化描述：

$$\text{一阶均差：} f\big[x_i, x_j\big] = \frac{f(x_i) - f\big(x_j\big)}{x_i - x_j}$$

$$\text{二阶均差：} f\big[x_i, x_j, x_k\big] = \frac{f\big[x_i, x_j\big] - f[x_j, x_k]}{x_i - x_k}$$

…

从而得到

$$f(x) = f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \dots, x_n](x - x_0) \cdot \dots \cdot (x - x_{n-1})$$

（2） 拉格朗日插值法：用$n+1$个n次曲线的叠加得到$f(x) = \sum_{i=0}^{n} a_i x^i$。

定义$f_i(x) = \prod_{j\neq i}^{0\le i\le n} \frac{x - x_j}{x_i - x_j}$，表示只经过数据点$(x_i, 1)$的$n$次曲线，则

$$f(x)=\sum_{i=0}^{n}y_i f_i(x)=\sum_{i=0}^{n}y_i\prod_{j\neq i}^{0\le i\le n}\frac{x-x_j}{x_i-x_j}$$

（3） 两种插值方法得到的是同一根曲线吗？

因为这两种方法均是线性的（多项式基的线性组合），所以可以从范德蒙德行列式入手分析解的唯一性是否成立。$Ax=b$有唯一解的条件是矩阵A非奇异，即行列式不为零。代入本问题即可验证是同一根曲线：

$$|A|=\prod_{j\neq i}^{0\le i\le j\le n}\left(x_j-x_i\right),x_j\neq x_i\ \leftrightarrow |A|\neq 0$$

（4） 多项式插值的本质是利用多项式的强大解释性去逼近任意光滑函数。这一思想促使了泰勒展开式的诞生。

设$f(x)$在$x_0,x_0+\Delta x,\dots,x_0+n\Delta x$处的值已知。和牛顿插值相比，这里的数据点是等间隔的。定义一阶差分：

$$\Delta f(x_0)=f(x_0+\Delta x)-f(x_0)$$

…

$$\Delta f(x_0+(n-1)\Delta x)=f(x_0+n\Delta x)-f(x_0+(n-1)\Delta x)$$

那么牛顿插值法中的$f_1(x)$变成了

$$f_1(x)=f(x_0)+\frac{\Delta f(x_0)}{\Delta x}(x-x_0)$$

定义二阶差分：

$$\Delta^2 f(x_0)=\Delta f(x_0+\Delta x)-\Delta f(x_0)$$

…

$$\Delta^2 f(x_0+(n-2)\Delta x)=\Delta f(x_0+(n-1)\Delta x)-\Delta f(x_0+(n-2)\Delta x)$$

那么牛顿插值法中的$f_2(x)$变成了

$$f_2(x)=f(x_0)+\frac{\Delta f(x_0)}{\Delta x}(x-x_0)+\frac{\Delta^2 f(x_0)}{2\Delta x^2}(x-x_0)(x-x_1)$$

可以验证，最终将会得到

$$f(x)=f_n(x)=f(x_0)+\frac{\Delta f(x_0)}{\Delta x}(x-x_0)+\frac{\Delta^2 f(x_0)}{2\Delta x^2}(x-x_0)(x-x_1)+\cdots$$
$$+\frac{\Delta^n f(x_0)}{2\Delta x^n}(x-x_0)\dots(x-x_{n-1})$$

当$\Delta x\to 0$时（此时$x-x_0=\cdots=x-x_n$），泰勒预测

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+o((x-x_0)^n)$$

也就是$\frac{\Delta^n f(x_0)}{\Delta x^n}=f^{(n)}(x_0)$，这就是泰勒展开式。

泰勒展开的本质是充分利用任意曲线上某一点上的“光滑信息”，
用多项式插值这一段局部曲线。

据此，我们可以直接计算任意曲线h在点x的n阶多项式p的近似：

$$\begin{cases} h(x) = p(x) \\ h'(x) = p'(x) \\ \cdots \\ h^{(n)}(x) = p^{(n)}(x) \end{cases}$$

求解这个线性方程组即可得到p的各个常系数，从而得到h的n阶泰勒展开。

（5） 泰勒展开在计算机内部有许多应用。比如三角函数不方便计算，但是多项式就好算多了。用泰勒展开估计函数值的合理程度收到收敛圆半径（展开点x_0到最近的奇点的距离）的影响。显然，对于$x^{\frac{1}{3}}$，在$x_0 = 29$处展开远比在$x_0 = 7$处展开精确的多。

“收敛圆的存在表明实数是复数的一部分，即使只研究实数问题，也会被复数影响到。”

（6） 记住一些常用的麦克劳林级数，它们在计算一些数列上下界的时候可能会起到一定帮助：

https://en.wikipedia.org/wiki/Taylor_series

6 利用积分计算级数的上下界：以调和级数$H_n = \sum_{i=1}^{n} \frac{1}{i}$为例，其中的每一项可以统一被一个如下的分段函数来表示：$h(x) \coloneqq \frac{1}{i+1}, 0 \le i < x \le i+1$，可通过积分的方式来理解（小长方形的宽度被限定为 1）。可以给出$h(x)$的上下界：

$$\overline{h}(x) \coloneqq \begin{cases} 1, 0 < x < 1 \\ \frac{1}{x}, x \ge 1 \end{cases}, \quad \underline{h}(x) \coloneqq \frac{1}{x+1}$$

所以$\int_0^n \underline{h}(x)dx \le H_n \coloneqq \int_0^n h(x)dx \le \int_0^n \overline{h}(x)dx$，可得$H_n = \Theta(\log(n))$。对$\log(n!)$展开并利用上下界积分即可得斯特林公式。

7 理解连续和一致连续的概念：

（1） 连续：$\lim_{x \to c}\big(f(x) - f(c)\big) = 0$（点）

（2） 一致连续：$\lim_{x_1 \to x_2}\big(f(x_1) - f(x_2)\big) = 0$（区间）

8 理解卷积运算：$x + y$恒等于n，从图像上看像卷毛巾一样。

（1） 离散：$(f * g)(n) = \sum_{\tau=-\infty}^{\tau=\infty} f(\tau) \cdot g(n-\tau)$

（2） 连续：$(f * g)(n) = \int_{\tau=-\infty}^{\tau=\infty} f(\tau) \cdot g(n-\tau)\, d\tau$

在卷积神经网络中，卷积常用于平滑图像，即把高频信号和周围的平均一下。

9 用牛顿迭代法寻找高次多项式方程的解：

做任意点处的切线，和x轴的交点作为迭代得到的新点，继续做该点的切线并寻找和x轴的交点。反复执行直到达到收敛条件：

$$x_{n+1} \leftarrow x_n - \frac{f(x_n)}{f'(x_n)}$$

将牛顿迭代法推广到任意曲线需要该曲线在给定区间内存在一阶导。该方法不总是收敛，可能会距离解越来越远、反复震荡、漏解。

10 理解梯度的概念：梯度是超曲面的函数值增长最快的方向在自变量空间上的投影（这个说法并不准确，因为能够选择方向的就是自变量，“投影”一词的使用不严谨。理解我所表达的意思就好）。

（1） 以一元函数$y = f(x)$为例，自变量空间是一维空间（即x轴）。只有两个方向：x轴正方向和x轴负

方向。$f'(x)>0$表明往前走$f(x)$会增大，梯度方向是x轴正方向；$f'(x)<0$表明往后走$f(x)$会增大，梯度方向是x轴负方向。

（2） 以二元函数$z=f(x,y)$为例，自变量空间是二维空间（xy平面），方向有无数个。梯度为$\Delta z=\left(\frac{\partial z}{\partial x},\frac{\partial z}{\partial y}\right)^{\mathrm{T}}$，这个向量的方向就是$z$增长最快的方向。

（3） 梯度与等高线垂直。这是显然的，因为等高线上的函数值相同，只有这样才可以做到“不偏不倚”。

11 理解局部最优性的必要条件：

（1） $\forall x\in \mathrm{R}^n, y=f(x)\in \mathrm{R}$，若$x^*$为局部最优解且$f$在$x^*$邻域内一阶可导，则$\nabla f(x^*)=0$。
证明思路：将f在x^*处一阶泰勒展开$f(x)=f(x^*)+\Delta x^{\mathrm{T}}\nabla f(x^*)$，要在邻域内$f(x^*)$最小，则恒有$\Delta x^{\mathrm{T}}\nabla f(x^*)\geq 0$。因为$\Delta x$可正可负（左邻域和右邻域），所以恒有$f(x^*)=0$。

（2） $\forall x\in \mathrm{R}^n, y=f(x)\in \mathrm{R}$，若$x^*$为局部最优解且$f$在$x^*$邻域内二阶可导，则$\nabla f(x^*)=0$且$\nabla^2 f(x^*)$为半正定矩阵。

证明思路：将f在x^*处二阶泰勒展开$f(x)=f(x^*)+\frac{1}{2}\Delta x^{\mathrm{T}}\nabla^2 f(x^*)\Delta x$，要在邻域内$f(x^*)$最小，则恒有$\frac{1}{2}\Delta x^{\mathrm{T}}\nabla^2 f(x^*)\Delta x\geq 0$，所以$\nabla^2 f(x^*)$半正定。

12 掌握无约束非线性优化算法：

最常用的无约束非线性优化算法是一维搜索，其核心是找到合适的迭代方向和每次迭代的最优步长。如果没有曲线信息可供利用（一阶、二阶甚至更高阶的导数），那么可以用黄金分割法、斐波那契法等方法缩小搜索区间，也可以借助多项式的强大表示能力执行插值法（牛顿法就是二点二次插值）；如果有信息可用，那么通常借助泰勒展开式获得迭代方向。具体地，如果有一阶信息可供利用，那么可以采用梯度法（最速下降法）及其一系列的改进方法（泰勒一阶展开）；如果有二阶信息可供利用，那么可以采用牛顿法及其一系列的改进方法（泰勒二阶展开）。牛顿法的两大问题分别是Hessian矩阵不总正定和Hessian矩阵的逆太难算。对于前者，改进策略是提出各种修正方案保证迭代方向一定是函数下降的方向；对于后者，改进策略是各种拟牛顿法（用别的矩阵来代替 Hessian 矩阵的逆）。此外，收敛速度介于梯度法和牛顿法之间的重要方法是共轭方向法，其本质上是保证每一次迭代都是有效的，即同一个迭代方向不会多次出现。

另一类无约束非线性优化的算法是信赖域方法。本质上，信赖域方法是在局部用二次型插值目标曲线，仍就是借助多项式的强大表示能力。

13 理解拉格朗日乘子法（处理等式约束）：

设$x\in \mathrm{R}^n$，目标函数为$f(x)$，等式约束为$g(x)=0, h(x)=0, \ldots$。要求$\min\limits_x f(x)$。

（1） 有一个等式约束：在极值点目标函数曲线和约束函数曲线相切，根据梯度与等高线垂直这一结论，二者的梯度相互平行：即$\nabla f+\lambda\nabla g=0$。

（2） 有多个等式约束：在极值点目标函数曲线的梯度是多条约束函数曲线的梯度的线性组合：

$$\nabla f+(\lambda\nabla g+\mu\nabla h+\cdots)=0$$

（3） 因此，引入拉格朗日函数$L(x)=f(x)+\lambda g(x)+\mu h(x)+\cdots$，联立

$$\begin{cases}\nabla L(x)=0\\ g(x)=0\\ h(x)=0\\ \cdots\end{cases}$$

即可求解极值点和λ、μ等乘子的最优值。

14 理解 KKT 条件（可处理不等式约束）：

设$x \in \mathrm{R}^k$，目标函数为$f(x)$，等式约束为$g_i(x)=0|_{i=1,\dots,n}$，不等式约束为$h_j(x) \le 0|_{j=1,\dots,m}$。要求$\min\limits_x f(x)$。

（1） 不等式约束没起作用：拉格朗日乘子法即可解决。

（2） 不等式约束起作用了：只可能在约束边界取得，因此不等式约束实际上通过等式约束来体现。那么同样有：在极值点目标函数曲线的梯度是多条约束函数曲线的梯度的线性组合。但是需要主要的是：在极值点目标函数曲线的梯度必然和不等式约束函数曲线的梯度方向的线性组合相反，所以$\mu_j \ge 0|_{j=1,\dots,m}$。
可总结为

$$\begin{cases} \nabla f + \sum\limits_{i=1}^{n} \lambda_i \nabla g_i + \sum\limits_{j=1}^{m} \mu_j \nabla h_j = 0 \\ h_j(x) = 0|_{j=1,\dots,m} \\ \mu_j \ge 0|_{j=1,\dots,m} \end{cases}$$

（3） 将两种情形整合到一起就有了 KKT 条件：

$$\begin{cases} \nabla f + \sum\limits_{i=1}^{n} \lambda_i \nabla g_i + \sum\limits_{j=1}^{m} \mu_j \nabla h_j = 0 \\ g_i = 0|_{i=1,\dots,n} \\ h_j \le 0|_{j=1,\dots,m} \\ \mu_j \ge 0|_{j=1,\dots,m} \\ h_j \mu_j = 0|_{j=1,\dots,m} \end{cases}$$

最后一个是 KKT 互补条件。当$\mu_j = 0$时，h_j没有出现在拉格朗日函数中，说明这个不等式约束没起作用，即$h_j \le 0$恒成立；当$u_j > 0$时，$h_j = 0$，说明h_j起作用了，在边界上起的作用。正是这个条件将两种情形整合到了一起。

15 牢记矩阵微积分的运算规则：

（1） 矩阵微积分是用矩阵和向量表示因变量每个成分关于自变量每个成分的偏导数。技巧是观察偏导的维度构成。

（2） 向量→标量：$\forall x \in \mathrm{R}^p, \forall y = f(x) = f(x_1, \dots, x_p) \in \mathrm{R}$，$\frac{\partial y}{\partial x} = \left(\frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial x_p}\right)^{\mathrm{T}} \in \mathrm{R}^{p \times 1}$。

（3） 标量→向量：$\forall x \in \mathrm{R}, \forall y = f(x) \in \mathrm{R}^p$，$\frac{\partial y}{\partial x} = \left(\frac{\partial y_1}{\partial x}, \dots, \frac{\partial y_n}{\partial x}\right) \in \mathrm{R}^{1 \times p}$。

（4） 向量→向量：$\forall x \in \mathrm{R}^p, \forall y = f(x) = f(x_1, \dots, x_p) \in \mathrm{R}^q$，

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_q}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_p} & \cdots & \frac{\partial y_q}{\partial x_p} \end{bmatrix} \in \mathrm{R}^{p \times q}$$

（5） 加法法则：$\forall x \in \mathrm{R}^p, \forall y = f(x) \in \mathrm{R}^q, z = f(x) \in \mathrm{R}^q$，$\frac{\partial (y+z)}{\partial x} = \frac{\partial y}{\partial x} + \frac{\partial z}{\partial x} \in \mathrm{R}^{p \times q}$。

（6） 乘法法则：$\forall x \in \mathrm{R}^p, \forall y = f(x) \in \mathrm{R}^q, z = f(x) \in \mathrm{R}^q, \frac{\partial y^{\mathrm{T}} z}{\partial x} = \frac{\partial y}{\partial x} z + \frac{\partial z}{\partial x} y$。$(p \times 1 = (p \times q) \times (q \times 1))$

（7） $\forall x \in \mathrm{R}^p, \forall y = f(x) \in \mathrm{R}^s, z = f(x) \in \mathrm{R}^t, A \in \mathrm{R}^{s\times t}$，$\frac{\partial y^\top Az}{\partial x} = \frac{\partial y}{\partial x}Az + \frac{\partial z}{\partial x}A^\top y \in \mathrm{R}^{p\times 1}$。

（8） $\forall x \in \mathrm{R}^p, \forall y = f(x) \in \mathrm{R}, z = f(x) \in \mathrm{R}^q$，$\frac{\partial yz}{\partial x} = y\frac{\partial z}{\partial x} + \frac{\partial y}{\partial x}z^\top \in \mathrm{R}^{p\times q}$。

（9） $\forall x \in \mathrm{R}, u = u(x) \in \mathrm{R}^s, g = g(u) \in \mathrm{R}^t$，$\frac{\partial g}{\partial x} = \frac{\partial u}{\partial x}\cdot\frac{\partial g}{\partial u} \in \mathrm{R}^{1\times t}$。

（10） $\forall x \in \mathrm{R}^p, u = u(x) \in \mathrm{R}^s, g = g(u) \in \mathrm{R}^t$，$\frac{\partial g}{\partial x} = \frac{\partial u}{\partial x}\cdot\frac{\partial g}{\partial u} \in \mathrm{R}^{s\times t}$。

（11） $\forall X \in \mathrm{R}^{p\times q}, u = u(X) \in \mathrm{R}^s, g = g(u) \in \mathrm{R}$，$\frac{\partial g}{\partial X_{ij}} = \frac{\partial u}{\partial X_{ij}}\cdot\frac{\partial g}{\partial u} \in \mathrm{R}$。

（12） 将以上规则付诸实践，如：

$$\forall x \in \mathrm{R}^p, \frac{\partial x}{\partial x} = I \in \mathrm{R}^{p\times p};$$

$$\forall x \in \mathrm{R}^p, \forall A \in \mathrm{R}^{q\times p}, \frac{\partial Ax}{\partial x} = A^\top \in \mathrm{R}^{p\times q};$$

$$\forall x \in \mathrm{R}^p, \forall A \in \mathrm{R}^{p\times q}, \frac{\partial x^\top A}{\partial x} = A \in \mathrm{R}^{p\times q}。$$

（13） Logistic 函数$\sigma(x) = \frac{1}{1+e^{-x}}$：$\sigma'(x) = \sigma(x)\big(1-\sigma(x)\big)$。当输入为$x = (x_1, \dots, x_n)^\top$时，

$$\sigma'(x) = diag(\sigma(x)\odot(1-\sigma(x))) \in \mathrm{R}^{n\times n}$$

（14） SoftMax 函数（将多个输入的标量映射为一个离散概率分布）：

设输入为$x = (x_1, \dots, x_n)^\top$，输出$z = (z_1, \dots, z_n)^\top$定义为

$$z_i = \mathrm{softmax}(x_i) = \frac{e^{x_i}}{\sum_{k=1}^n e^{x_k}}$$

我们来认真算一算 SoftMax 函数的导数怎么求解。

首先，$z = \mathrm{softmax}(x) = \frac{1}{\sum_{k=1}^n e^{x_k}}(e^{x_1}, \dots, e^{x_n})^\top = \frac{\exp(x)}{\sum_{k=1}^n e^{x_k}} = \frac{\exp(x)}{I_n^\top \exp(x)} \in \mathrm{R}^n$，其中$I_n = (1, \dots, 1)^\top \in \mathrm{R}^n$。所以 $\frac{\partial z}{\partial x} = \frac{\partial\left(\frac{\exp(x)}{I_n^\top \exp(x)}\right)}{\partial x} = \frac{\partial\left(\exp(x)\cdot\frac{1}{I_n^\top \exp(x)}\right)}{\partial x} = \frac{\partial(\exp(x))}{\partial x}\cdot\frac{1}{I_n^\top \exp(x)} + \frac{\partial\left(\frac{1}{I_n^\top \exp(x)}\right)}{\partial x}\cdot[\exp(x)]^\top = \frac{diag(\exp(x))}{I_n^\top \exp(x)} - \left(\frac{1}{I_n^\top \exp(x)}\right)^2\cdot\frac{\partial(I_n^\top \exp(x))}{\partial x}\cdot[\exp(x)]^\top$（最后一步用（9）来做）。

紧接着，根据（6），$\frac{\partial(I_n^\top \exp(x))}{\partial x} = \frac{\partial(\exp(x))}{\partial x}\cdot I_n + \frac{\partial I_n^\top}{\partial x}\cdot[\exp(x)]^\top = diag(\exp(x))I_n = \exp(x)$。

所以，原式$= diag\left(\frac{\exp(x)}{I_n^\top \exp(x)}\right) - \frac{\exp(x)}{I_n^\top \exp(x)}\cdot\frac{[\exp(x)]^\top}{I_n^\top \exp(x)} =$

$$diag\big(\mathrm{softmax}(x)\big) - \mathrm{softmax}(x)\cdot[\mathrm{softmax}(x)]^\top$$

16 复数存在的意义是什么？

（1） 让我们从更高的维度去审视函数，从而获得更多未知的细节。复数虽然看起来和向量一样，但实际上实数域的扩充，在本质上更接近实数。（相比之下，向量的乘法运算点积、叉积与实数乘法大不相同。）复数在计算上兼容实数，有加减乘除乘方开根号取对数等运算。

（2） 重点关心复数乘法：设$z_1 = a + bi$，长度为$|z_1|$，幅角为$\arg(z_1)$；设$z_2 = c + di$，长度为$|z_2|$，幅角为$\arg(z_2)$。则$z_1 * z_2$的长度为$|z_1| * |z_2|$，幅角为$\arg(z_1) + \arg(z_2)$。

当$z_2 = i$时，其作用时让z_1逆时针旋转90°（缩放因子为 1，角度增大90°）。

17 理解欧拉公式$e^{i\theta} = cos\theta + i \cdot sin\theta$

（1） 掌握证明和验证欧拉公式的方法。利用e^x的三种定义方式：

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$$

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots$$

$$e^x = \frac{d}{dx}e^x$$

进行推导。下面这个页面有全部详尽的论证过程：

https://zh.wikipedia.org/wiki/%E6%AC%A7%E6%8B%89%E5%85%AC%E5%BC%8F

（2） 根据欧拉公式可知$e^{i\theta}$是复平面上长度为 1、幅角为θ的复数，$e^{i\theta}$的轨迹组成了复平面上的单位圆。任意复数可经由欧拉公式表达为$z = a + bi = re^{i\theta}$，且$|z| = r$，$\arg(z) = \theta$。

设$z_1 = r_1e^{i\theta_1}$，$z_2 = r_2e^{i\theta_2}$，则

$$z_1 * z_2 = r_1r_2e^{i(\theta_1+\theta_2)}$$

$$\frac{z_1}{z_2} = \frac{r_1}{r_2}e^{i(\theta_1-\theta_2)}$$

$$a^i = e^{i\ln a}$$

系数项乘（除）、指数相加（减），这种运算方式更接近实数，或许这才是复数的完整形态。

（3） 当$\theta = \pi$时，

$$e^{i\pi} + 1 = 0$$

这个公式中出现了零元、单位元、最基本的虚数i，以及两个最常用的自然数π和e。这就是“上帝的公式”。

18 理解傅里叶级数和傅立叶变换：

（1） 光作为一种波，其颜色由各个组成光的振幅（a_n）和频率（$\frac{n}{2\pi}$）决定。光的色散现象就是

$$f(x) = \sum_n a_n \sin(nx)$$

（2） 设$f(x)$是周期性函数且周期为T，满足傅里叶级数的收敛条件，则$f(x)$可写作

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n\cos\frac{2\pi nx}{T} + b_n\sin\frac{2\pi nx}{T}\right)$$

其中

$$\begin{cases} a_n = \dfrac{2}{T}\displaystyle\int_{x_0}^{x_0+T} f(x)\cdot\cos\frac{2\pi nx}{T}dx \\ b_n = \dfrac{2}{T}\displaystyle\int_{x_0}^{x_0+T} f(x)\cdot\sin\frac{2\pi nx}{T}dx \end{cases}$$

（3） 傅里叶级数的本质是将任意周期函数（曲线）视为无数旋转的叠加。

傅里叶级数可视为圆周运动的组合：不断以做圆周运动的点为新的圆周运动的圆心继续做圆周运动。套娃的次数越多就越逼近该周期函数，选取不同的点和不同的半径就是在逼近不同的周期函数。不断增加的x如同时间一样永不回头。

（4） 傅里叶级数表达式的推导源自于对问题“任意周期函数是否可写成三角函数之和”的求解，核心思路如下：

$f(x)=C$是常函数，也是周期函数 → 要有常数项；

$\sin x$和$\cos x$都要有，保证奇偶性；

需要$\sin\frac{2\pi nx}{T}$和$\sin\frac{2\pi nx}{T}$，保证T一定是其周期；

通过加减调节振幅。

请自行推导。此外，根据欧拉公式可得$\cos\theta=\frac{e^{i\theta}-e^{-i\theta}}{2}$，$\sin\theta=\frac{e^{i\theta}+e^{-i\theta}}{2}$，代入可得

$$f(x)=\sum_{n=-\infty}^{\infty}\left(c_n\cdot e^{i\frac{2\pi nx}{T}}\right)$$

其中$c_n=\frac{1}{T}\int_{x_0}^{x_0+T}f(x)\cdot e^{-i\frac{2\pi nx}{T}}dx$。

通过引入欧拉公式我们进入了复平面。这使得我们可以在时域上观察流逝的时间和运动得到的周期曲线；在频率上观察圆周运动的频率。傅立叶级数本质上是线性分量的叠加，$\left\{e^{i\frac{2\pi nx}{T}}\right\}|_{\forall n}$正是无限维向量空间中的一组单位正交基（可通过三角函数的点积为零验证），而c_n是基$e^{i\frac{2\pi nx}{T}}$上的系数。每一个基反映的是做不同圆周运动的旋转频率。实际上，系数a_n和b_n亦可通过在正交基上的投影得到。

（5） 什么是频域图？其实就是周期函数$f(x)$在不同频率上的基的振幅。以方波信号

$$f(x)=\begin{cases}2, & 2k\pi\le x\le(2k+1)\pi\\ 0, & (2k-1)\pi\le x\le 2k\pi\end{cases}$$

为例，用傅里叶级数

$$1+\frac{4}{\pi}\sin x+0\cdot\sin 2x+\frac{4}{3\pi}\sin 3x+0\cdot\sin 4x+\frac{4}{5\pi}\sin 5x$$

来逼近，那么基$\{1,\sin x,\dots,\sin 5x\}$上的系数向量为$\left\{1,\frac{4}{\pi},0,\frac{4}{3\pi},0,\frac{4}{5\pi}\right\}^{\mathrm{T}}$。在$x$轴上$x=0,\dots,5$的位置以系数向量对应的元素为高度绘制竖线，即可得到对应的频域图。从时域到频域的过程就是傅立叶变换，本质是基底的切换。

（6） 还可以用傅里叶级数去逼近任意非周期函数。如何逼近？令$T\to\infty$！T越大，频率越小，频段的划分越密集。当$T\to\infty$时，基会铺满整个频谱，频域图上各个竖线的最高点将会联结成一条连续的曲线。

三 线性代数与矩阵理论

1 理解线性代数中的基本概念：

（1） （定义）同维度的向量v和w的张成（span）是集合$\{av+bw|a,b\in R\}$。（仅通过向量加法与向量数乘，我们能得到的、所有的向量的集合）

（2） （定义）对于向量空间V中的一组向量$\{v_1,v_2,\dots,v_n\}$，它们是线性相关的 *iff* 存在非全为零的元素$a_1,a_2,\dots,a_n\in\mathrm{R}$满足

$$\sum_i a_i v_i = 0$$

它们是线性无关的 *iff* 不存在满足这样条件的$\{a_i\}_{i=1,2,\dots,n}$（即$a_i \equiv 0$）。

(3) （定义）向量空间的一组基是张成该空间的一个线性无关的向量集（相互不可替代）。基的个数就是空间的维数。对于n维线性空间V，

$$\left\{ v_i = \left[\underbrace{0, \dots, 1, \dots, 0}_{\text{仅第}i\text{个位置为 1}} \right]^{\mathrm{T}} \right\}_{i=1,\dots,n}$$

是一组单位正交基（关于正交的概念，后面会深入分析）。

(4) 理解线性变换（linear transformation）：

所谓“线性”，是指：

- 对于空间中的直线，变换后仍应当是直线；
- 直线的比例保持不变；
- 空间的原点必须保持固定。

满足第一、二点但是不满足第三点的，其实是仿射变换（affine transformation），即线性变换 + 平移。可以在高维度通过线性变换实现低维度的仿射变换。

本质上，线性变换是保持网格线平行且等距分布的变换。

严格意义上，若一个变换L满足可加性

$$L(u + v) = L(u) + L(v)$$

和成比例（一阶齐次）

$$L(cv) = cL(v), \forall c \in R$$

则称L是线性的。线性变换不仅可以作用于向量（当然，这取决于我们如何定义“向量”），也可以作用在函数上。例如，求导运算也是一种线性变换，这就是求导公式中可加性和成比例的由来：

$$\frac{\partial(f+g)}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$$

$$\frac{\partial(\alpha f)}{\partial x} = \alpha \frac{\partial f}{\partial x}$$

因此，我们最好把这种特征抽象出来，而不拘泥于具体的形式。感兴趣的同学可去了解线性空间（aka 向量空间）、度量空间、赋范空间以及内积空间。

(5) 那么问题来了，如何用数值去描述线性变换？

我们只需关注新的基的位置如何描述即可，而基前面数乘的系数在变换前后不会发生变换。默认情况下，旧的基为i和j，则对于变换前的向量$[x, y]^T$，有

$$\begin{bmatrix} x \\ y \end{bmatrix} = xj + yj \rightarrow xi' + yj' = \begin{bmatrix} xi'_1 + yj'_1 \\ xi'_2 + yj'_2 \end{bmatrix} = \begin{bmatrix} i'_1 & j'_1 \\ i'_2 & j'_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = [i', j'] \begin{bmatrix} x \\ y \end{bmatrix}$$

因此，我们发现，可以用矩阵来实施线性变换。一个矩阵和一个线性变换总是一一对应，且矩阵的每一列都是变换后的线性空间的一组“基”（打引号是因为此处允许i', j'线性相关）。上面的这个式子也引申出了矩阵-向量乘法的定义。注意，这里的新基其实仍然是用旧的坐标体系（旧基）来描述的。

(6) 线性变换作为一个函数，通过矩阵来实现可以写作$A(x)$，线性变换的复合可以写作$B\big(A(x)\big)$，去

掉这些括号，就可以得到矩阵乘法的运算，即复合变换。具体地，不妨设$A=[a_1,a_2]$，$B=[b_1,b_2]^T=\begin{bmatrix}b_{11} & b_{21}\\ b_{12} & b_{22}\end{bmatrix}$，则

$$BA=\begin{bmatrix}b_{11} & b_{21}\\ b_{12} & b_{22}\end{bmatrix}a_1+\begin{bmatrix}b_{11} & b_{21}\\ b_{12} & b_{22}\end{bmatrix}a_2\text{（分别作用在两个基向量上）}$$

矩阵运算被化归成了矩阵-乘法运算。我们就这样得到了矩阵乘法的计算方式。

思考复合变换的形成，我们可以很容易证明矩阵乘法满足结合律：$(AB)C=A(BC)$，但不满足交换律$AB\neq BA$。

(7) 线性变换前后空间内任意区域的缩放比例即行列式（determinant）。在二维线性空间上这对应着单位面积的缩放，在三维线性空间上这对应着单位体积的缩放。

- 如果一个矩阵的行列式为零，则说明该线性变换导致了空间的坍缩（新的线性空间的维度小于旧的线性空间的维度），这在矩阵的列向量（即新基）线性相关的时候发生；
- 如果一个矩阵的行列式为负，则说明该线性变换改变了空间的定向（orientation）。

总结一下：

$$\det(A)=|A|\begin{cases}>1\text{：对图形有放大作用}\\ =1\text{：图形大小不变}\\ 0<|A|<1\text{：对图形有缩小作用}\\ =0\text{：线性变换不可逆（“覆水难收”）}\\ <0\text{：改变左右手规则}\end{cases}$$

因为两次线性变换缩放的比例之积和一次复合变换的缩放比例相等，所以显然有

$$\det(A)\det(B)=\det(AB)$$

如果一个方阵和对角阵相似（即可对角化、可执行特征值分解），那么A的行列式等于所有特征值的积，后面会谈到这一点

(8) 从线性变换的角度出发理解线性方程组$Ax=b$：x的求解依赖于线性变换A对空间的操纵。

- 当$|A|\neq 0$时，空间没有发生坍缩，则这个线性变换是可逆的。我们总可以通过对b实施“A的逆变换”来得到x。这里也引申出了线性变换的逆（逆矩阵）的定义：

  A的逆是满足以下性质的唯一变换：$A^{-1}A=I$。

  其中I是“什么都不做”的变换。
- 当$|A|=0$时，空间发生坍缩，此时，我们无法通过一个线性变换再变换回去。从有损压缩的角度来理解，在没有任何场外信息介入的情况下，无法将一个点或一条线“解压缩”为一个平面，因为部分信息已经被丢失了。

如何描述空间是否发生了坍缩？我们需要一种建立在线性变换之上的概念，他需要表达出经过线性变换后，新空间的维度。这就是秩（rank）。

将x看成旧空间的任意未知向量，则所有Ax的取值组成了一个新的线性空间，我们称之为“A的列空间（column space）”，记为$range(A)$。这是因为A的每一列就是新空间的“基”（打引号是因为此处允许新基线性相关）。因此，A的秩等于A的列空间的维数。当秩等于新基的个数时，我们称之为“满秩”，即新基线性无关。

注意，零向量一定在列空间内，这是因为线性空间必然包含原点，而线性变换不会改变原点。不过，除了零向量，可能还会存在一些向量在经过线性变换A之后也成为了零向量，我们将所有的这些向量放入集合“A的零空间”内：

$$null(A) = \{x \in V | Ax = 0\}$$

$null(A)$又称为“A的核（kernel）”。$null(A)$的秩可以大于0。例如，当一条直线上的所有点经过线性变换后被压缩至原点，那么$rank\big(null(A)\big) = 1$。

如何理解矩阵的转置？已知$A \in \mathrm{R}^{m\times n}: V \to U$的每一列$A_{:,j}$是$A$对应的线性变换的新基在旧的线性空间$V$的坐标，每一行$A_{i,:}$其实是在以新基为坐标体系、旧基在新的线性空间$U$的坐标，即旧基在新基上的投影。

对于任意矩阵A，行秩=列秩。这是因为

$$n = rank\big(range(A)\big) + rank\big(null(A)\big) = rank\big(range(A^{\top})\big) + rank\big(null(A)\big)$$

此外还有$rank(A) = \min\{m, n\}$。

（9） 向量$a, b \in \mathrm{R}^n$的点积运算方式为$a \cdot b = \sum_{i=1}^{n} a_i b_i$。

从多维空间到一维空间的线性变换需要一个$1 \times n$的矩阵来实现。对于一个处于n维线性空间的向量b，当经过线性变换$A \in \mathrm{R}^{1\times n}$时，得到的就是其在新空间（一维空间，即数轴）上的坐标。这其实就是“投影”——将n维向量b投影到数轴上的任意向量上。因为$Ab = \sum_{i=1}^{n} A_{1i} b_i$，这就是向量点积的定义，所以我们就将投影和向量点积建立起了关联——

$a \cdot b$其实是a在b上的投影的长度、b的长度与符号的乘积。
当a在b上的投影和b方向相同时，乘积为正，否则为负。
$a \cdot b = b \cdot a$，即谁向谁投影是无所谓的。

从对偶性的角度来理解：

一个向量的对偶是由它所定义的线性变换；
一个多维空间到一维空间的对偶是多维空间中的某个特定向量。

（10） 首先给出向量叉乘的定义：对于三维空间中的向量v和w，$v \times w$仍然是一个三维向量，不妨记为p。p的长度为v和w所组成的平行四边形的面积，p的方向由右手定则确定（v对应食指，w对应中指，p对应拇指）。计算公式为

$$p = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \det\left(\begin{bmatrix} i & v_1 & w_1 \\ j & v_2 & w_2 \\ k & v_3 & w_3 \end{bmatrix}\right) \xleftrightarrow{\text{代数余子式变换}} \begin{vmatrix} v_2 & w_2 \\ v_3 & w_3 \end{vmatrix} i - \begin{vmatrix} v_1 & w_1 \\ v_3 & w_3 \end{vmatrix} j + \begin{vmatrix} v_1 & w_1 \\ v_2 & w_2 \end{vmatrix} k$$

其中i, j, k为三维空间的基向量。

如何从线性变换的角度理解叉积？
对于给定的三维空间中的向量v和w，不妨定义如下函数$f: \mathrm{R}^3 \to \mathrm{R}$：

$$f\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

该函数的几何意义是：输出是输入和v, w组成的平行六面体的有向体积。

理解一：根据行列式的性质（代数余子式的计算规则）可知，这是一个线性变换（证明“可加性”和“成比例”）。因此，必然存在一个1×3的矩阵P代表这个变换，根据第（9）点提及的对偶性，可将“P代表的变换对$[x,y,z]^T$ 的作用”理解为“向量$p=P^T$与$[x,y,z]^T$的点积”。代替上式的左边可得到

$$\begin{cases}p_1=v_2w_3-v_3w_2\\p_2=v_3w_1-v_1w_3\\p_3=v_1w_2-v_2w_1\end{cases}$$

所以向量$p=(v_2w_3-v_3w_2)i+(v_3w_1-v_1w_3)j+(v_1w_2-v_2w_1)k$。

理解二：如何计算$[x,y,z]^\top,v,w$组成的平行六面体的有向体积？小学数学告诉我们“底面积乘高”——不妨以v,w组成的平面为底，其面积为v,w组成的平行四边形的面积，而高垂直于该底面，且长度为$[x,y,z]^T$在该高上的投影。
因此，函数f的几何意义应当详细描述为：对于任意给定的三维向量$[x,y,z]^\top$，将其投影到与v,w组成的平面垂直的直线上，并且将投影长度与v,w组成的平行四边形的面积相乘。

但是，这和垂直于v,w组成的平面、且长度为与v,w组成的平行四边形的面积的向量与$[x,y,z]^\top$点乘是一回事（即“理解一”和“理解二”殊途同归）！将$[i,j,k]$带入$[x,y,z]^\top$可发现“理解一”中求得的p和通过叉乘定义得到的p是一回事，这就表明通过叉乘得到的p具备“理解二”中的几何意义。

上述内容阐述了叉乘的计算过程与其几何直观是如何关联的。

2 理解相似矩阵（$B=P^{-1}AP$）：

（1） 相似性只定义在方阵上。

（2） 线性变换是通过指定基下的矩阵来表达的，相似矩阵则是站在不同角度（另一组基）观察同一个线性变换。客观存在的一个点，不因为我们的观测角度发生了变化而变化，但是会有不同的坐标值。
对于$P^{-1}APx$，不妨设当前的线性空间代表了甲的视角，x是当前空间坐标体系下的一个点，P这个线性变换代表了乙的视角，则Px首先将x放到乙的视角下（即在乙的坐标体系下表示），然后在乙上对该点执行某个变换，最后再通过P^{-1}回到我们的视角下。如果存在一个$B=P^{-1}AP$，则Bx就是直接在我们的视角下对该点执行同一个变换。不管怎样，作用在该点上的变换是一样的。

（3） 为什么我们关心相似矩阵？相似矩阵的本质是坐标变换（切换观测角度），所以我们可以借助相似性将一个矩阵切换到更简单的表达方式上来理解。怎样的表达算简单呢？对角阵！这就是对角化。在这种思路的引导下，诞生了特征值分解。（以下两点涉及特征值与特征向量，建议先阅读三-3）

（4） 只要两个矩阵相似，它们的行列式就相同。这一点用$|B|=|P^{-1}|\cdot|A|\cdot|P|$很容易得到。行列式是相似矩阵之间的相似不变量。如果和对角阵相似（即可对角化、可执行特征值分解），那么A的行列式就十分容易计算，就是所有特征值的积。这是因为对角阵就是单纯地把各个维度拉伸指定数值的变换，缩放的比例因子显然等于各维度拉伸之积。因为对角阵不涉及旋转变换，因此实际上每个旧基向量都没有改变方向，仅仅是长度有所缩放，这符合特征向量的定义，所以对角阵的非零元素就是该矩阵的有所特征值。
因此，若矩阵A可对角化，则$|A|$就等于A所有特征值的乘积。

（5） 相似矩阵具有完全相同的特征值和特征向量。这是因为$|B-\lambda I|=|P^{-1}AP-P^{-1}\lambda IP|=|A-\lambda I|$。

3 理解特征值和特征向量：

（1） 观察$Ax=\lambda x$，可发现我们试图寻找的是经过线性变换A之后，在方向上保持不变（掉头不算）、仅仅是长度有所缩放的向量x以及缩放的比例λ。换句话说，特征向量x是经过线性变换A后仍然留在它所张成的线性空间中的向量。

（2） 将向量看成线性空间中的一个点，那么左乘矩阵可以看成是（通过操纵空间）对点施加的一种运动（使点发生位移）。特征值反映了运动的速度，而特征向量则反映了运动的方向。线性变换是各个方向上不同速度的运动的叠加。从这个理解出发，对于向量x，不断左乘A，可发现x会不断贴合到A最大特征值对应的特征空间上。实际上，这正是特征向量的一种迭代求法。从这个理解出发，也会明白为什么相似矩阵具有完全相同的特征值和特征向量（因为客观存在的运动方向和速度不会随着观测角度的变化而变化）。

（3） 特征空间是将同一个特征值所对应的一个（或数个）特征向量视为基，所张成的空间。特征空间中基的个数与对应的特征值重根的个数相等。

（4） 为什么对方阵A可以用$|A-\lambda I|=0$来求解其特征值和特征向量？

求证思路：根据特征值和特征向量的定义，即$Ax=\lambda x$恒成立，即$(A-\lambda I)x=0$。这是一个线性方程组，意味着将任意向量投射到零空间内，那么$A-\lambda I$作为一个线性变换所对应的伸缩因子必须恒为 0，即行列式为 0。

4 理解特征值分解：

（1） 特征值分解其实就是对角化。$\forall A\in \mathrm{R}^{n\times n}$，$A$可对角化（$A=PDP^{-1}$）当且仅当$A$有$n$个线性无关的特征向量。其中$P=(p_1,\dots,p_n)$是$n$个特征向量组成的矩阵，$D=diag(\lambda_1,\dots,\lambda_n)$是由特征值组成的对角阵。

证明思路：由$Ap_i=\lambda_i p_i|_{i=1,\dots,n}$可得$AP=A(p_1,\dots,p_n)=(Ap_1,\dots,Ap_n)=(\lambda_1 p_1,\dots,\lambda_n p_n)=(p_1,\dots,p_n)D=PD$，若$\{p_1,\dots,p_n\}$线性无关则$P$可逆，从而有$A=PDP^{-1}$。

这个结论就是特征值分解。

（2） 性质：$A^k=PD^kP^{-1}$。这个性质再一次验证了不断左乘A，x会不断贴合到最大特征值对应的特征空间上。因为特征值呈指数增长，很小的特征值将会非常小，相应的运动几乎可以忽略不计；而最大的特征值将会非常大，对应的“高速运动”将明显占据主导地位。

（3） 为什么我们关心对称矩阵（$A=A^{\mathrm{T}}$）？因为对称阵在执行特征值分解时具备许多优秀的性质，包括：

对称阵有n个实特征值（包括重根）[***TODO:*** 补充解释]；

对称阵的各个特征空间两两正交；

对称阵每个特征值的重根次数是该特征值对应的特征空间中线性无关的基底的个数；

对称阵可以正交对角化（正交的概念参见三-8）。

其中第二点的证明思路：

因为$\lambda_i p_i\cdot p_j=(\lambda_i p_i)^{\mathrm{T}}p_j=(Ap_i)^{\mathrm{T}}p_j=\left(p_i^{\mathrm{T}}A^{\mathrm{T}}\right)p_j=p_i^{\mathrm{T}}\left(A^{\mathrm{T}}p_j\right)=p_i^{\mathrm{T}}\left(\lambda_j p_j\right)=\lambda_j p_i\cdot p_j$且$\lambda_i\neq\lambda_j$，可得$p_i\cdot p_j=0\rightarrow p_i\perp p_j$。（$\lambda_i\neq\lambda_j$是因为我们考虑的是特征空间，重根的来自同一个特征空间）

所谓正交对角化是：$\forall A\in \mathrm{R}^{n\times n}$，$A$可正交对角化当且仅当存在正交阵$P$和对角阵$D$使得$A=PDP^{-1}=PDP^{\mathrm{T}}$。在这种情况下，$P$和$P^{-1}$作为正交阵对应的是方向相反的旋转变换，而$D$对应的则是伸缩变换。显然，旋转变换其实就是切换了观测角度而已，我们更关心伸缩变换。因此正交对角化可用于“规范化二次型”，相当于把矩阵“扶正”。我们后面还会谈到这一点。

（4） 对称阵不一定可逆。

（5） 方阵A可以正交对角化当且仅当A是对称阵。

（6） 根据对称阵可正交对角化这一性质，可得到谱分解定理（每个分量$\lambda_i u_i u_i^{\mathrm{T}}$是$A$的一个 spectrum）：

$$A=\lambda_1 u_1^{\mathrm{T}}u_1+\cdots+\lambda_n u_n^{\mathrm{T}}u_n$$

证明思路：令$D=diag(\lambda_1,\dots,\lambda_n)$，$P=(u_1,\dots,u_n)$，则$DP^\top=(\lambda_1u_1^\top,\dots,\lambda_nu_n^\top)^\top$。所以$PDP^\top=(u_1,\dots,u_n)(\lambda_1u_1^\top,\dots,\lambda_nu_n^\top)^\top=\sum_{i=1}^n u_i\lambda_iu_i^\top$（看成行向量和列向量的乘法）。

每个分量其实记录了一部分信息。如果将分量按照降序排列，则省略末位的一个或多个分量就实现了信息的压缩。（前20%的分量可能占据了90%的有效信息）

（7） 执行特征值分解的步骤：

Step 1：根据$|A-\lambda I|=0$找到所有特征值；

Step 2：对于每个特征值λ，根据$Ax=\lambda x$找到其对应的（一个或多个）特征向量（特征值重根的个数等于对应的特征空间中特征向量的个数）；

Step 3：将特征值按照降序排列组成D。对应地，排列特征向量组成P和$P^\top$。

5 $\forall A\in \mathrm{R}^{n\times n}$，理论上，如何求解线性方程组$Ax=b$？

将A写作$(a_1,\dots,a_n)^\top$，则$Ax=\sum_{i=1}^n x_ia_i$是以$\{a_1,\dots,a_n\}$为基的线性组合。若$b\in range(A)$则$\exists x$使得$\sum_{i=1}^n x_ia_i=b$。若$\{a_1,\dots,a_n\}$线性无关，则b只有一种表示方法，否则有无数种。这就对应着原方程组有唯一解和有无穷解。因此可总结如下：

（1） 若A为非奇异阵，则$x=A^{-1}b$。根据克拉默法则可知$x_i=\frac{|A_i|}{|A|}$。

（2） 若A为奇异阵且$b\in range(A)$，则$Ax=b$有无穷解。

（3） 若A为奇异阵且$b\notin range(A)$，则$Ax=b$无解。

其中，一个矩阵是奇异的当且仅当它至少有一个特征值为零。

可是，克拉默法则是怎样的到的？下面给出了一个有趣的几何学解释：

首先思考如下等价变换：

$$1\cdot y=\det\left(\begin{bmatrix}1&x\\0&y\end{bmatrix}\right),\qquad 1\cdot x=\det\left(\begin{bmatrix}x&0\\y&1\end{bmatrix}\right)$$

上述两个等式的左右分别是对同一个平行四边形的有向面积的不同计算方法。扩展到n维线性空间，即

$$x_i=\det(X)$$

其中$X_{:,j\neq i}=\left[\underbrace{0,\dots,1,\dots,0}_{\text{仅第}i\text{个位置为}1}\right]^\top$，$X_{:,i}=[x_1,\dots,x_n]^\top$。

对于向量$x=[x_1,\dots,x_n]^\top$，假设经过某个变换A，x变成了b，上面的等价变换会怎样？

首先，等式左边的x_i被缩放了$|A|$倍，而等式右边的$\det(X)$变成了$\det(X')$，其中$X'_{:,j\neq i}=A_{:,j\neq i}$，这是因为基从旧的变成了$\{A_{:,j}\}_{j=1,\dots,n}$，而$X'_{:,i}=[b_1,\dots,b_n]^\top$，变形即可得到$x_i=\frac{|A_i|}{|A|}$（这里$A_i$和$X'$是一回事）。

6 如何让计算机求解线性方程组$Ax=b$？

依据克拉默法则计算x的复杂度为$O(n^4)$，太慢！于是就诞生了高斯消去法——

（1） 如果A是下三角矩阵，则x非常容易计算。因为

$$x_i=\frac{b_i-\sum_{j=1}^{i-1}a_{ij}x_j}{a_{ii}}$$

依次求解$x_1,x_2,\dots,x_n$即可，这就是前向算法。这个过程要求$a_{ii}\neq 0|_{\forall i}$，复杂度为$O(n^2)$。

（2） 如果A是上三角矩阵，则有后向算法

$$x_i=\frac{b_i-\sum_{j=i+1}^{n}a_{ij}x_j}{a_{ii}}$$

（3） 如何将任意矩阵A转换为上/下三角矩阵？通过初等变换——

每实施一次初等变换，相当于左乘一个对应的初等矩阵。初等矩阵是可逆的，只需按照相反的初等变换变换回来即可。由此可以得到矩阵可逆的等价定义：

任意方阵A可逆 $\leftrightarrow$ 和单位矩阵I初等行变换等价。

并且：$(A|I) \approx (I|A^{-1})$，这就是初等变换法求矩阵的逆。

（4） 我们有无数种选择可以将A通过初等变换转换为上三角矩阵。高斯消去法则是固定了一个程式：

第i轮高斯消去是将从第$i+1$行到最后一行的第i个元素依次通过和第i行比较消为 0。

上述程式可以将A转换成下三角矩阵和上三角矩阵的乘积，请自行验证。这就是 LU 分解。由此得$LUx = b$，其中L为下三角矩阵，U为上三角矩阵，则先后使用前向算法和后向算法即可得到x。

7 如何通过迭代求解线性方程组$Ax = b$？

（1） 对于大规模稀疏矩阵而言，更好的求解策略是迭代。因为$x_i = \frac{b_i - \sum_{j \neq i} a_{ij} x_j}{a_{ii}}$，所以可以得到

$$\text{Jacobi 迭代方法：} x_i^{(k+1)} \leftarrow \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}}{a_{ii}}$$

每次用最新的$x_j|_{j=1,\dots,i-1}$，得到

$$\text{Gauss} - \text{Seidel 迭代方法：} x_i^{(k+1)} \leftarrow \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}}{a_{ii}}$$

（2） 采用矩阵分解的方式表达迭代方程：

令$D = diag(A)$，E是A的严格下三角部分的负，F是A的严格下上角部分的负，则$A = D - E - F$。

则对于 Jacobi 迭代方法有

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b$$

对于 Gauss-Seidel 迭代方法有

$$x^{(k+1)} = (D - E)^{-1}Fx^{(k)} + (D - E)^{-1}b$$

如果左右乘上松弛变量则有$\omega(D - E - F) = \omega x$，从而得到松弛法迭代公式

$$x^{(k+1)} = (D - \omega E)^{-1}(\omega F + (1 - \omega)D)x^{(k)} + (D - \omega E)^{-1}\omega b$$

8 深入理解正交性：

（1） 正交集（orthogonal set）：两两正交的向量的集合。如果一个正交集不含零向量，则这个正交集线性无关。（证明思路：$\forall i: 0 = c_1 v_1 + \cdots + c_p v_p \leftrightarrow 0 = 0 \cdot v_i = \left(c_1 v_1 + \cdots + c_p v_p\right) \cdot v_i$，由$v_i \cdot v_j = 0|_{i \neq j}$可知$c_i = 0$。）因此正交集可以作为向量空间$\mathrm{R}^p$的一组正交基。

进一步地，可定义单位正交集（orthonormal set）。

（2） 对任意向量$y \in \mathrm{R}^p$，如何求解y在一组正交基$\{v_1, \dots, v_p\}$各个基上的坐标（系数）？设$y = \sum_{i=1}^{p} c_i v_i$，则

$$c_i = \frac{y \cdot v_i}{v_i \cdot v_i}$$

这个结论亦可用于正交投影时各个分量上坐标的计算。

（3） 正交矩阵（orthogonal matrix）：拥有单位正交的列向量和行向量的方阵（因此更合理的名称应当是“单位正交矩阵（orthonormal matrix）”）。正交矩阵使单位正交基在正交变换之后仍然保持单位长度、相互垂直（刚体运动）。正交矩阵U满足$U^{-1} = U^{\mathsf{T}}$。这是一个非常优越的性质。因为很多运算涉及到求解矩阵的逆，计算机对一般的矩阵求逆的复杂度很高。如果能和正交矩阵扯上关系，就会方便处理很多。实际上 QR 分解和特征值分解就是这么干的。

证明思路：$U = (u_1, \dots, u_n)$，$U^{\mathsf{T}} = (u_1^{\mathsf{T}}, \dots, u_n^{\mathsf{T}})^{\mathsf{T}}$，乘一下即得证。

若U是列向量相互单位正交的矩阵且非方阵，虽然U上没有定义矩阵的逆，但仍然有$U^{\mathsf{T}}U = I$。

酉矩阵（Unitary Matrix）是正交阵在复空间上的推广。

（4） 左乘正交阵是在实施旋转变换，这是因为$|U^{\mathsf{T}}U| = 1$。Ux将x旋转过去，$U^{-1}Ux = U^{\mathsf{T}}Ux = Ix = x$又将$x$旋转回来了。

（5） 如何从任意一组基$\{x_1, \dots, x_n\}$中构造出一组正交基$\{q_1, \dots, q_n\}$？

$$q_1 = x_1$$

$$q_2 = x_2 - \frac{x_2 \cdot q_1}{q_1 \cdot q_1} q_1$$

$$q_3 = x_3 - \frac{x_3 \cdot q_1}{q_1 \cdot q_1} q_1 - - \frac{x_3 \cdot q_2}{q_2 \cdot q_2} q_2$$

$$\dots$$

$$q_n = x_p - \sum_{i=1}^{n-1} \frac{x_i \cdot q_i}{q_i \cdot q_i} q_i$$

如何理解呢？以x_2为例：将x_2正交分解，其中一个分量在q_1上，系数为$\frac{x_2 \cdot q_1}{q_1 \cdot q_1}$，另一个分量就是$q_2$。

后续的过程均是如此。每次得到的q_i是x_i在现有的正交基上分解完毕之后的补。若要求构造出一组单位正交基$\{q_1, \dots, q_n\}$，只需在上述每个步骤后面附上单位化的操作：

$$\forall j \in [1, \dots, n]: \ q_j = x_j - \sum_{i=1}^{j-1} (x_j \cdot q_i) q_i, \ \ q_j = \frac{q_j}{\|q_j\|}$$

仔细观察上式，若将$x_j \cdot q_i$记为r_{ij}（$1 \le i < q \le n$），将$\|q_j\|$记为r_{jj}，则$R = (r_{ij})_{n \times n}$是一个上三角矩阵。将$(q_1, \dots, q_n)$记为$Q$，则$X = QR$。也就是说，我们将任意矩阵$X$分解成了一个正交阵和上三角矩阵的乘积，这就是 QR 分解。类似于高斯消去法，这个分解也可以拿来求解线性方程组：

$$QRx = b, \text{令} Rx = y \rightarrow \ y = Q^{\mathsf{T}} b$$

$$Rx = y \rightarrow \text{前向算法}$$

9 理解二次型：

（1） 对于一个二次函数（方程），增加非二次项，不会改变曲线的形状，只会在xy轴上产生一些旋转、平移的效果。因此，研究二次函数（方程）的性质，只关心二次部分就够了。只包含二次部分的函数（方程）被称为二次齐次函数（方程）。

（2） 显然可以通过对称矩阵来描述二次齐次函数（方程）。例如

$$ax^2 + 2bxy + cy^2 \ \leftrightarrow \ [x \ y] \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \ \leftrightarrow \ x^{\mathsf{T}} A x$$

可以将对称阵A正交对角化$A = PDP^{\mathsf{T}}$，只保留伸缩变换的部分。在图像上，这就是将二次型对应的函数（方程）的曲线扶正。以圆锥曲线之一椭圆为例，这个操作相当于把一个斜着的椭圆摆正并平移到坐标轴的中心。

（3） 对于二次函数$f(x) = x^{\mathsf{T}} A x$，若$\forall x \in \mathrm{R} \wedge x \neq 0$，$f(x) > 0$，则$f(x)$是正定二次型，$A$是正定矩阵。同理可定义半正定矩阵、负定矩阵、半负定矩阵。正定阵必然可逆，但是不一定可对角化。

（4） 判断一个对称阵A是否是正定矩阵的重要方法是对其执行正交对角化$A = PDP^{\mathsf{T}}$，若D中每一个特征值都大于零，则正定。同理可判定是否为半正定矩阵、负定矩阵、半负定矩阵。为什么可以这样判断？这是因为$x^{\mathsf{T}} A x$和$x^{\mathsf{T}} D x$本质是一样的，后者只是移除了其中旋转的部分。将$x^{\mathsf{T}} D x$展开就得到了$\sum_{i=1}^{n} d_{ii} x_i^2$，只有当所有的$d_{ii}$都恒大于零时才可以保证$\sum_{i=1}^{n} d_{ii} x_i^2$恒大于零。也就是说，需要$A$的特征值都大于零。

（5） 正定/负定阵一定可逆。因为对应的特征值不可能为零。

（6） $f(x)=x^{\top}Ax$（$s.t.\left||x|\right|=1$）在何时取得极值？

当A为对称阵时，$f(x)$的最大值为A的最大特征值，$f(x)$的最小值为A的最小特征值。

这个结论可以通过拉格朗日乘子法得到。

证明思路：当A为对称阵时，对A执行正交对角化，则$f(x)=(P^{\top}x)^{\top}D(P^{\top}x)$，$P^{\top}$作为正交阵只对$x$产生旋转变换，不改变其二范数，故$f(x)=f(y)=y^{\top}Dy$，展开再使用拉格朗日乘子法即可。

10 如何理解奇异值分解（SVD）？

（1） 对称阵可以进行正交对角化。如何将这个优美的性质运用到任意矩阵A上？

（2） $\forall A\in R^{m\times n}$，因为$m\times n=(m\times m)\cdot(m\times n)\cdot(n\times n)$，不违背矩阵乘法规则，所以必然存在一个$m\times n$的矩阵$\Sigma$，一个$m\times m$的矩阵$U$，和一个$n\times n$的矩阵$V$使得$A=U\Sigma V^{\top}$。那么问题来了，$U$、$\Sigma$和$V$分别怎么算？

我们来看看$A^{\top}A$和$AA^{\top}$。因为$A^{\top}A=(U\Sigma V^{\top})^{\top}(U\Sigma V^{\top})=V(\Sigma^{\top}\Sigma)V^{\top}$且$A^{\top}A$是正交阵，所以$V$是$A^{\top}A$的特征向量组成的正交阵！$\Sigma^{\top}\Sigma$是$A^{\top}A$的特征值组成的对角阵。同理，因为$AA^{\top}=(U\Sigma V^{\top})(U\Sigma V^{\top})^{\top}=U(\Sigma\Sigma^{\top})U^{\top}$且$A^{\top}A$是正交阵，所以$U$是$AA^{\top}$的特征向量组成的正交阵，$\Sigma\Sigma^{\top}$是$AA^{\top}$的特征值组成的对角阵。以上这段分析就已经告诉我们$U$和$V$分别怎么算了。

那么Σ呢？要知道$\Sigma^{\top}\Sigma$是一个$n\times n$的对角阵而$\Sigma\Sigma^{\top}$是一个$m\times m$的对角阵，而m可以不等于n。尽管如此，设$k=\min\{m,n\}$，由$(A^{\top}A)^{\top}=AA^{\top}$可知二者前$k$个对角元素必然相等，不妨设为$\lambda_1,\dots,\lambda_k$。首先，必然有$\lambda_1\geq\cdots\geq\lambda_k\geq 0$，只有这样我们才可以对这些数开根号。证明思路：因为$\forall x\in \mathrm{R}^n$，$x^T(A^{\top}A)x=(Ax)^{\top}(Ax)=\|Ax\|_2\geq 0$，所以$A^{\top}A$是半正定矩阵。同理可证$A^{\top}A$也是半正定矩阵。根据第 11 点的第四条可知半正定矩阵的所有特征值必然均为非负数。

（3） 设$rank(A)=r$，则我们称$\sigma_i=\sqrt{\lambda_i}|_{i=1,\dots,r}$为$A$的奇异值（显然有$r\leq k$）。现在我们来说$\Sigma$怎么算。

其实，$\Sigma=\begin{bmatrix}\left(\sqrt{\Sigma^{\top}\Sigma}\right)_{r\times r} & 0\\ 0 & 0\end{bmatrix}=\begin{bmatrix}\left(\sqrt{\Sigma\Sigma^{\top}}\right)_{r\times r} & 0\\ 0 & 0\end{bmatrix}=\begin{bmatrix}diag(\sigma_1,\dots,\sigma_r) & 0\\ 0 & 0\end{bmatrix}$，其中$\left(\sqrt{\Sigma^{\top}\Sigma}\right)_{r\times r}$表示对$\Sigma^{\top}\Sigma$每个元素依次开根号并只保留前$r\times r$的部分。

（4） 其实，令$U=(u_1,\dots,u_m)$，由$A=U\Sigma V^{\top}$亦可知

$$\forall i=1,\dots,r:\ u_i=\frac{1}{\sigma_i}Av_i$$

至于$u_i|_{i=r+1,\dots,m}$，可用$\{u_1,\dots,u_m\}$相互单位正交得到。

（5） 执行奇异值分解的步骤：

Step 1：计算$rank(A)=r$；

Step 2：对$A^{\top}A$执行正交对角化$A^{\top}A=V(\Sigma^{\top}\Sigma)V^{\top}$，得到$V$和$\Sigma^{\top}\Sigma$；

Step 3：对$AA^{\top}$执行正交对角化$AA^{\top}=U(\Sigma\Sigma^{\top})U^{\top}$，得到$U$和$\Sigma\Sigma^{\top}$；

Step 4：得$\Sigma=\begin{bmatrix}\left(\sqrt{\Sigma^{\top}\Sigma}\right)_{r\times r} & 0\\ 0 & 0\end{bmatrix}$或$\Sigma=\begin{bmatrix}\left(\sqrt{\Sigma\Sigma^{\top}}\right)_{r\times r} & 0\\ 0 & 0\end{bmatrix}$。

（6） SVD 的应用：

类似普分解定理，$\forall A\in R^{m\times n}$有

$$A=\sum_{i=1}^{r}\sigma_i u_i v_i^{\top}$$

一共有r个分量，按照σ_i的大小降序排列。这里可以解释为什么Σ只选取$\sqrt{\Sigma^{\top}\Sigma}$的前$r\times r$的部分。这是因为秩代表了矩阵$A$线性无关的列向量的个数，从而决定了分量的个数。如果我们只保留这r个分量的前部分分量（例如前 20%的分量），就能够实现数据压缩。

11 矩阵的迹的本质什么？

（1） 迹是矩阵对角线元素之和。

（2） 如果两个矩阵相似，那么它们的迹相同。迹与观测空间无关，也是相似矩阵之间的相似不变量。如果和对角阵相似（即可对角化、可执行特征值分解），那么A的迹等于所有特征值的和。

（3） 行列式和迹均为相似变换中的不变量，是让相似矩阵拥有同一内核的本质原因。它们全部被特征值表达了出来。

最后

本文档参考了马同学的专栏（https://www.matongxue.com/columns/）、Jeff Erickson 的教材《Algorithms》（http://jeffe.cs.illinois.edu/teaching/algorithms/）、3Blue1Brown 的视频（https://www.3blue1brown.com/）以及 ZJU-CS 开设的课程《计算机应用数学》。

本文档会保持持续更新。最新版本可以在 http://hliangzhao.me/math/math.pdf 下载。有任何疑问或发现了错误请联系 hliangzhao@zju.edu.cn。