微积分与概率论基础

3月机器学习在线班　邹博

2015年3月7日

回忆知识

- 求S的值：

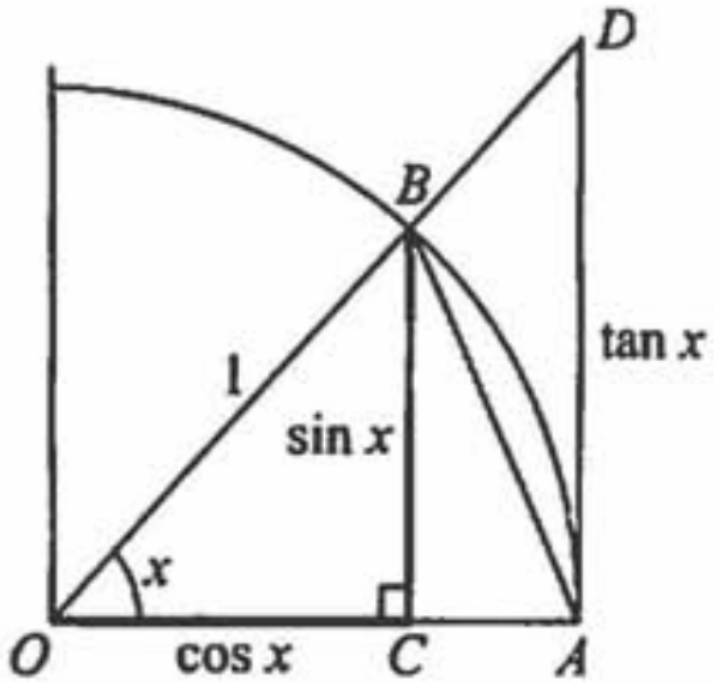
$$S = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!} + \cdots$$

复习微积分：两边夹定理

- 当$x \in U(x_0,r)$时，有$g(x) \leqslant f(x) \leqslant h(x)$成立，并且$\lim\limits_{x \to x_0} g(x) = A$，$\lim\limits_{x \to x_0} h(x) = A$，那么

$$\lim_{x \to x_0} f(x) = A$$

极限



- 由右图：$\sin x < x < \tan x$，

  $x \in U(0, \varepsilon)$
- 从而：$1 < x/\sin x < 1/\cos x$
- 即：$\cos x < \sin x/x < 1$
- 因为：$\lim_{x \to 0} \cos x = \cos 0 = 1$
- 从而：

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

- 该式将三角函数和多项式建立了极限关系

思考

- 该式的极限是多少？

$$\lim_{x \to 0} \frac{\sin^2 x}{x^2}$$

复习微积分：极限存在定理

- 单调有界数列必有极限
  - 单增数列有上界，则其必有极限

构造数列$\{x_n\}$

$$x_n = \left(1+\frac{1}{n}\right)^n$$

$$= 1 + C_n^1 \frac{1}{n} + C_n^2 \frac{1}{n^2} + C_n^3 \frac{1}{n^3} + \cdots + C_n^n \frac{1}{n^n}$$

$$= 1 + n \cdot \frac{1}{n} + \frac{n(n-1)}{2!} \cdot \frac{1}{n^2} + \frac{n(n-1)(n-2)}{3!} \cdot \frac{1}{n^3} + \cdots + \frac{n(n-1)(n-2)\cdots 1}{n!} \cdot \frac{1}{n^n}$$

$$= 1 + 1 + \frac{1}{2!} \cdot \left(1-\frac{1}{n}\right) + \frac{1}{3!} \cdot \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) + \cdots + \frac{1}{n!} \cdot \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{n-1}{n}\right)$$

$$< 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!}$$

$$< 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}}$$

$$= 3 - \frac{1}{2^{n-1}}$$

$$< 3$$

自然常数

- 根据 $\left(1+\frac{1}{n+1}\right)^{n}<\left(1+\frac{1}{x}\right)^{x}<\left(1+\frac{1}{n}\right)^{n+1}$

- 从而公式 $\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x}$ 的极限存在，定义为e。

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

导数

- 简单的说，导数就是曲线的斜率，是曲线变化快慢的反应
- 二阶导数是斜率变化快慢的反应，表征曲线的凸凹性
  - 在GIS中，往往一条二阶导数连续的曲线，我们称之为“光顺”的。
  - 还记得高中物理老师时常念叨的吗：加速度的方向总是指向轨迹曲线凹的一侧

常用函数的导数

（1）$C'=0$（C 为常数）；

（2）$(x^n)'=nx^{n-1}$（$n\in Q$）；

（3）$(\sin x)'=\cos x$；

（4）$(\cos x)'=-\sin x$；

（5）$(a^x)'=a^x\ln a$；

（6）$(e^x)'=e^x$；

（7）$(\log_a x)'=\frac{1}{x}\log_a e$；

（8）$(\ln x)'=\frac{1}{x}$.

$$(u+v)'=u'+v'$$

$$(uv)'=u'v+uv'$$

应用

- 已知函数$f(x)=x^x$，$x>0$
- 求$f(x)$的最小值
  - 领会幂指函数的一般处理套路
- 附：$N^{\frac{1}{\log N}}=?$
  - 在计算机算法跳跃表Skip List的分析中，用到了该常数。
  - 背景：跳表是支持增删改查的动态数据结构，能够达到与平衡二叉树、红黑树近似的效率，而代码实现简单。

求解x^x

$$t = x^x$$

$$\rightarrow \ln t = x \ln x$$

$$\xrightarrow{\text{两边对x求导}} \frac{1}{t} t' = \ln x + 1$$

$$\xrightarrow{\text{令}t'=0} \ln x + 1 = 0$$

$$\rightarrow x = e^{-1}$$

$$\rightarrow t = e^{-\frac{1}{e}}$$

Taylor公式 – Maclaurin公式

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$$

$$f(x)=f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n+o(x^n)$$

Taylor公式的应用

- 数值计算：初等函数值的计算(在原点展开)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \cdots + (-1)^{m-1}\frac{x^{2m-1}}{(2m-1)!} + R_{2m}$$

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

- 在实践中，往往需要做一定程度的变换。

计算e^x

- 给定正实数x，计算e^x=?
- 一种可行的思路：
- 求整数k和小数r，使得
  - x = k*ln2 + r, |r| <= 0.5*ln2
- 从而：

$$e^x = e^{k\cdot\ln 2+r}$$

$$= e^{k\cdot\ln 2}\cdot e^r$$

$$= 2^k\cdot e^r$$

Taylor公式的应用

- 考察基尼指数的图像、熵、分类误差率三者之间的关系
  - 将f(x)=-lnx在x0=1处一阶展开，忽略高阶无穷小，得到f(x)≈1-x
  - 从而：

$$H(X) = -\sum_{k=1}^{K} p_k \ln p_k$$

$$\approx \sum_{k=1}^{K} p_k (1 - p_k)$$

  - 上述结论，在决策树章节中会进一步讨论

方向导数

- 如果函数z=f(x,y)在点P(x,y)是可微分的，那么，函数在该点沿任一方向L的方向导数都存在，且有：

$$\frac{\partial f}{\partial l}=\frac{\partial f}{\partial x}\cos\varphi+\frac{\partial f}{\partial y}\sin\varphi$$

- 其中，φ为x轴到方向L的转角。

梯度

- 设函数z=f(x,y)在平面区域D内具有一阶连续偏导数，则对于每一个点P(x,y)∈D，向量

$$\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$$

为函数z=f(x,y)在点P的梯度，记做**grad**f(x,y)

- 梯度的方向是函数在该点变化最快的方向
  - 考虑一座解析式为H(x,y)的山。在(x0,y0)点的梯度是在该点坡度最陡的方向。
- 梯度下降法
  - 思考：如果下山方向和梯度呈θ夹角，下降速度是多少？

凸函数

- f(x)在区间I上连续，如果对I上任意两点x1，x2，恒有f((x1+x2)/2)<(f(x1)+f(x2))/2，则称f(x)在I上是凸的。

- 注：中国大陆数学界某些机构关于函数凹凸性定义和国外的定义是相反的。Convex Function在某些中国大陆的数学书中指凹函数。Concave Function指凸函数。但在中国大陆涉及经济学的很多书中，凹凸性的提法和其他国家的提法是一致的，也就是和数学教材是反的。举个例子，同济大学高等数学教材对函数的凹凸性定义与习惯定义正好相反。另外，也有些教材会把凸定义为上凸，凹定义为下凸。

凸函数的判定

- 定理：f(x)在区间[a,b]上连续，在(a,b)内二阶可导，那么：
  - 若f''(x)>0，则f(x)是凸的；
  - 若f''(x)<0，则f(x)是凹的

- 即：一元二阶可微的函数在区间上是凸的，当且仅当它的二阶导数是非负的

凸函数

- 凸函数更一般的表述

$$f\left(\lambda x_1+\left(1-\lambda\right)x_2\right)\le\lambda f\left(x_1\right)+\left(1-\lambda\right)f\left(x_2\right)$$

f为凸函数，则有：

$$f(\theta_1 x_1+\ldots+\theta_n x_n)\le\theta_1 f(x_1)+\ldots+\theta_n f(x_n)$$

其中$0\le\theta_i\le 1,\theta_1+\ldots+\theta_n=1.$

- 意义：可以在确定函数的凸凹性之后，对函数进行不等式替换。

凸性质的应用

- 设p(x)、q(x)是在X中取值的两个概率分布，给定如下定义式：

$$D(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\log\frac{p(x)}{q(x)}$$

- 试证明：D(p||q) ⩾0
  - 上式在最大熵模型等内容中会详细讨论。

注意到y=logx在定义域上是凹函数

$$
\begin{aligned}
& D(p \| q) \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum_x p(x) \log \frac{q(x)}{p(x)} \\
&\geq -\log \sum_x \left(p(x) \cdot \frac{q(x)}{p(x)} \right) \\
&= -\log \sum_x q(x) \\
&= -\log 1 \\
&= 0
\end{aligned}
$$

概率论

- 对概率的认识：$P \in [0,1]$
  - P=0
    - 事件出现的概率为0→事件不会发生？
  - 将位于[0,1]的函数y=f(x)看成x对应y事件的概率
    - 要求f(x)在定义域[0,1]的积分为1
- 古典概型
  - 排列组合
- 概率密度函数Probability Density Function
- 累计分布函数

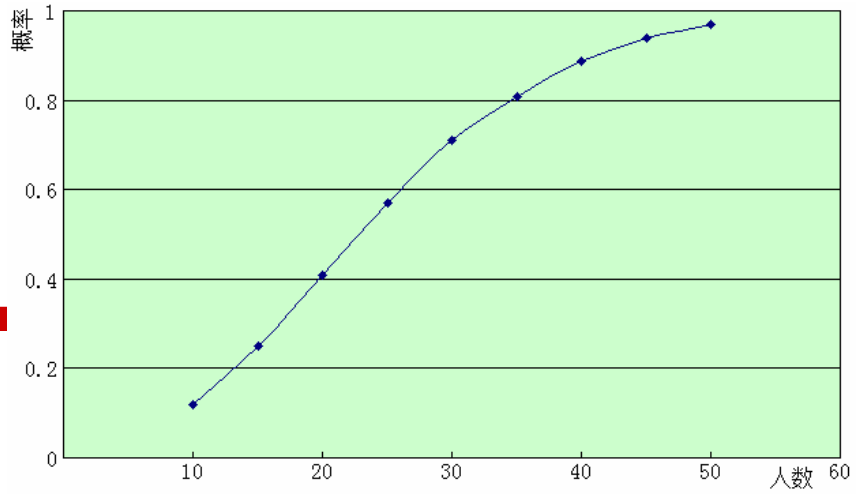
古典概型

- 举例：将n个不同的球放入N(N≥n)个盒子中，假设盒子容量无限，求事件A={每个盒子至多有1个球}的概率。

解 $\mathrm{P(A)}=\frac{P_N^n}{N^n}$

- 基本事件总数：
  - 第1个球，有N种放法；
  - 第2个球，有N种放法；
  - ……
  - 共：N^n种放法。
- 每个盒子至多放1个球的事件数：
  - 第1个球，有N种放法；
  - 第2个球，有N-1种放法；
  - 第3个球，有N-2种放法；
  - ……
  - 共：$\mathrm{N(N-1)(N-2)\cdots(N-n+1)}=P_N^n$

实际问题



- 某班上有50位同学，至少有2人生日相同的概率是多少？

n	10	15	20	25	30	35	40	45	50
P	0.12	0.25	0.41	0.57	0.71	0.81	0.89	0.94	0.97

装箱问题

- 将12件正品和3件次品随机装在3个箱子中。每箱中恰有1件次品的概率是多少？

解

- 将15件产品装入3个箱子，每箱装5件，共有15!/(5!5!5!)种装法；
- 先把3件次品放入3个箱子，有3!种装法。对于这样的每一种装法，把其余12件产品装入3个箱子，每箱装4件，共有12!/(4!4!4!)种装法；
- P(A)= (3!*12!/(4!4!4!)) / (15!/(5!5!5!)) = 25/91

与组合数的关系

- 把n个物品分成k组，使得每组物品的个数分别为n1,n2…nk，(n=n1+n2+…+nk)，则不同的分组方法有$\frac{n!}{n_1!n_2!n_3!\cdots n_k!}$种。

- 上述问题的简化版本，即n个物品分成2组，第一组m个，第二组n-m个，则分组方法有$\frac{n!}{m!(n-m)!}$，即：C_n^m。

概率

- 条件概率：

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

- 全概率公式：

$$P(A)=\sum_i P(A|B_i)P(B_i)$$

- 贝叶斯(Bayes)公式：

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

思考题

- 8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取一支射击，结果中靶。求该枪是已校准过的概率。

分布

- 复习各种常见分布本身的统计量
- 在复习各种分布的同时，重温积分、Taylor展式等前序知识
- 常见分布是可以完美统一为一类分布

两点分布

0—1分布

已知随机变量 X 的分布律为

X	1	0
p	p	$1-p$

则有 $E(X)=1\cdot p+0\cdot q=p,$

$$D(X)=E(X^2)-[E(X)]^2$$

$$=1^2\cdot p+0^2\cdot(1-p)-p^2=pq.$$

二项分布 Bernoulli distribution

设随机变量 X 服从参数为 n, p 二项分布，

(法一) 设X_i为第i 次试验中事件 A 发生的次数，$i=1,2,\cdots,n$

则

$$X=\sum_{i=1}^{n} X_i$$

显然，X_i 相互独立均服从参数为 p 的0—1分布，

$$\text{所以}\quad E(X)=\sum_{i=1}^{n} E(X_i)=np.$$

$$D(X)=\sum_{i=1}^{n} D(X_i)=np(1-p).$$

二项分布

(法二) X 的分布律为

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\cdots,n),$$

则有 $$E(X)=\sum_{k=0}^{n}k\cdot P\{X=k\}=\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k}$$

$$=\sum_{k=0}^{n}\frac{kn!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$=\sum_{k=1}^{n}\frac{np(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}$$

$$=np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}$$

$$=np[p+(1-p)]^{n-1}=np$$

二项分布

$$E(X^2) = E[X(X-1)+X] \quad = E[X(X-1)] + E(X)$$

$$= \sum_{k=0}^{n} k(k-1)\binom{k}{n} p^k (1-p)^{n-k} + np$$

$$= \sum_{k=0}^{n} \frac{k(k-1)n!}{k!(n-k)!} p^k (1-p)^{n-k} + np$$

$$= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(n-k)!(k-2)!} p^{k-2} (1-p)^{(n-2)-(k-2)} + np$$

$$= n(n-1)p^2 [p+(1-p)]^{n-2} + np = (n^2-n)p^2 + np.$$

$$D(X) = E(X^2) - [E(X)]^2 = (n^2-n)p^2 + np - (np)^2$$

$$= np(1-p)$$

考察Taylor展式

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

$$1 = 1 \cdot e^{-x} + x \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \frac{x^3}{3!} \cdot e^{-x} + \cdots + \frac{x^k}{k!} \cdot e^{-x} + R_n \cdot e^{-x}$$

$$\frac{x^k}{k!} \cdot e^{-x} \longrightarrow \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

泊松分布

设 $X \sim \pi(\lambda)$, 且分布律为

$$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}, \quad k=0,1,2,\cdots, \quad \lambda>0.$$

则有

$$E(X)=\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\cdot\lambda$$

$$=\lambda e^{-\lambda}\cdot e^{\lambda}=\lambda$$

泊松分布Poisson distribution

- 在实际事例中，当一个随机事件，以固定的平均瞬时速率λ(或称密度)随机且独立地出现时，那么这个事件在单位时间(面积或体积)内出现的次数或个数就近似地服从泊松分布P(λ)。
  - 某一服务设施在一定时间内到达的人数
  - 电话交换机接到呼叫的次数
  - 汽车站台的候客人数
  - 机器出现的故障数
  - 自然灾害发生的次数
  - 一块产品上的缺陷数
  - 显微镜下单位分区内的细菌分布数
  - 某放射性物质单位时间发射出的粒子数

泊松分布

$$E(X^2)=E[X(X-1)+X]$$

$$=E[X(X-1)]+E(X)$$

$$=\sum_{k=0}^{+\infty}k(k-1)\cdot\frac{\lambda^k}{k!}e^{-\lambda}+\lambda$$

$$=\lambda^2e^{-\lambda}\sum_{k=2}^{+\infty}\cdot\frac{\lambda^{k-2}}{(k-2)!}+\lambda=\lambda^2e^{-\lambda}e^{\lambda}+\lambda=\lambda^2+\lambda.$$

所以 $D(X)=E(X^2)-[E(X)]^2=\lambda^2+\lambda-\lambda^2=\lambda$

泊松分布的期望和方差都等于参数λ.

均匀分布

设 $X \sim U(a,b)$, 其概率密度为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b,\\ 0, & \text{其他}.\end{cases}$$

则有 $E(X)=\int_{-\infty}^{\infty} xf(x)\mathrm{d}x=\int_a^b \dfrac{1}{b-a}x\mathrm{d}x=\dfrac{1}{2}(a+b).$

$$D(X)=E(X^2)-[E(X)]^2$$

$$=\int_a^b x^2\frac{1}{b-a}\mathrm{d}x-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}$$

指数分布

设随机变量 X 服从指数分布，其概率密度为

$$f(x)=\begin{cases}\dfrac{1}{\theta}e^{-x/\theta}, & x>0,\\ 0, & x\le 0.\end{cases}\quad \text{其中}\ \theta>0.$$

则有

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\,dx=\int_{0}^{+\infty}x\cdot\frac{1}{\theta}e^{-x/\theta}\,dx$$

$$=-xe^{-x/\theta}\Big|_{0}^{+\infty}+\int_{0}^{+\infty}e^{-x/\theta}\,dx\quad=\theta$$

$$D(X)=E(X^2)-[E(X)]^2=\int_{0}^{+\infty}x^2\cdot\frac{1}{\theta}e^{-x/\theta}\,dx-\theta^2$$

$$=2\theta^2-\theta^2\quad=\theta^2$$

指数分布

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\le 0\end{cases}$$

- 其中λ＞0是分布的一个参数，常被称为率参数(rate parameter)。即每单位时间内发生某事件的次数。指数分布的区间是[0,∞)。如果一个随机变量X呈指数分布，则可以写作：X~ Exponential(λ)。
- 指数分布可以用来表示独立随机事件发生的时间间隔，比如旅客进机场的时间间隔、软件更新的时间间隔等等。
- 许多电子产品的寿命分布一般服从指数分布。有的系统的寿命分布也可用指数分布来近似。它在可靠性研究中是最常用的一种分布形式。

指数分布的无记忆性

- 指数函数的一个重要特征是无记忆性(遗失记忆性，Memoryless Property)。
  - 如果一个随机变量呈指数分布，当s,t≥0时有：

$$P(x > s+t \mid x > s) = P(x > t)$$

  - 即，如果x是某一元件的寿命，已知元件使用了s小时，它总共使用至少s+t小时的条件概率，与从开始使用时算起它使用至少t小时的概率相等。

正态分布

设 $X \sim N(\mu,\sigma^2)$, 其概率密度为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad \sigma>0,\quad -\infty<x<+\infty.$$

则有 $$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$$

$$=\int_{-\infty}^{+\infty}x\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x.$$

令 $$\frac{x-\mu}{\sigma}=t \Rightarrow x=\mu+\sigma t,$$

正态分布

$$E(X)=\int_{-\infty}^{+\infty}x\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(\mu+\sigma t)e^{-\frac{t^2}{2}}dt$$

$$=\mu\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}te^{-\frac{t^2}{2}}dt$$

$$=\mu.$$

正态分布

$$D(X)=\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)\,\mathrm{d}x$$

$$=\int_{-\infty}^{+\infty}(x-\mu)^2\cdot\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x.$$

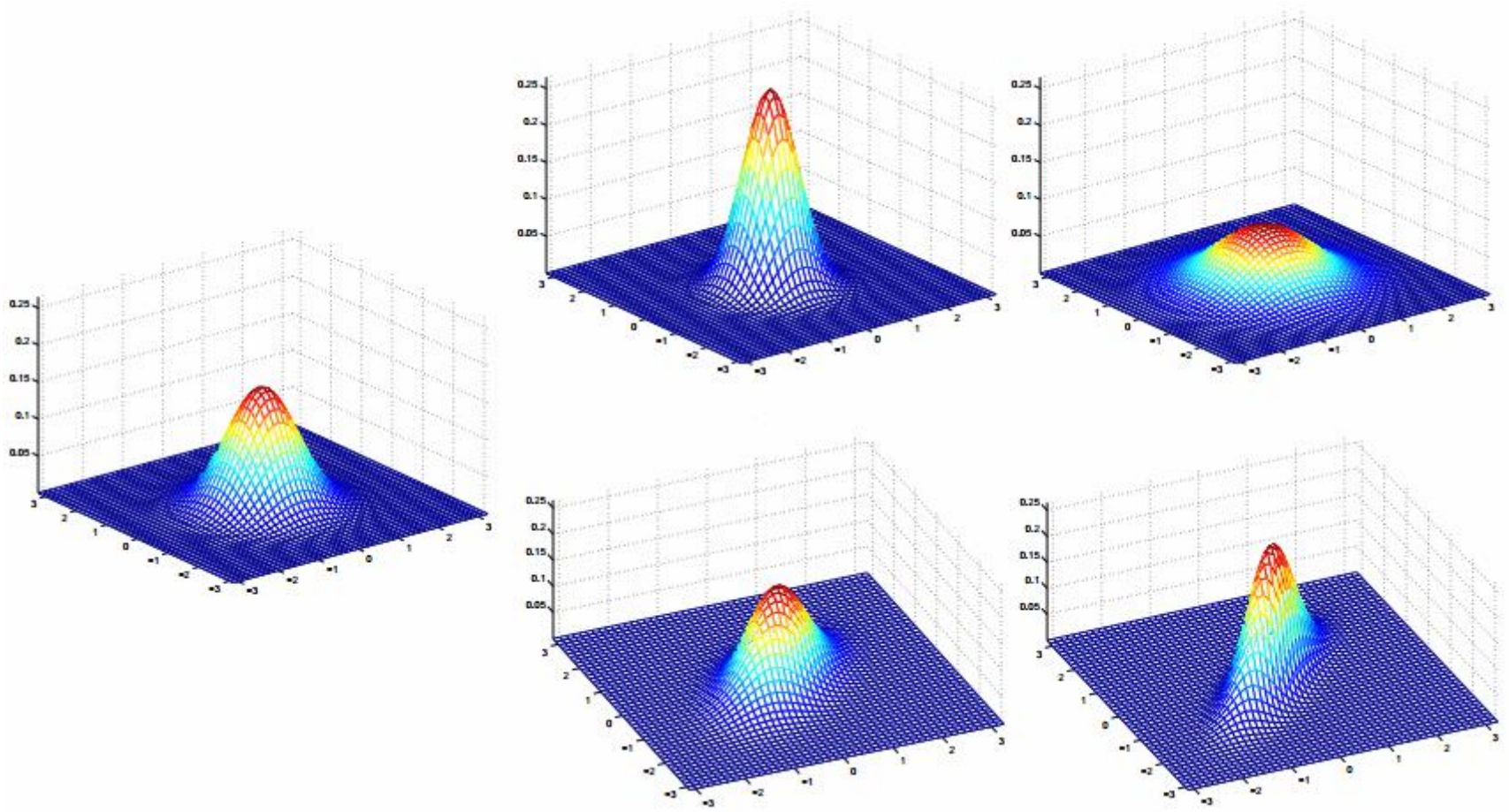
令 $\dfrac{x-\mu}{\sigma}=t$,得

$$D(X)=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t^2\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\left(-t\mathrm{e}^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\right)$$

$$=0+\frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi}=\sigma^2.$$

二元正态分布



总结

分　　布	参数	数学期望	方差
两点分布	$0<p<1$	p	$p(1-p)$
二项分布	$n\geq 1,$ $0<p<1$	np	$np(1-p)$
泊松分布	$\lambda>0$	λ	λ
均匀分布	$a<b$	$(a+b)/2$	$(b-a)^2/12$
指数分布	$\theta>0$	θ	θ^2
正态分布	$\mu,\sigma>0$	μ	σ^2

指数族

The exponential family

To work our way up to GLMs, we will begin by defining exponential family distributions. We say that a class of distributions is in the exponential family if it can be written in the form

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta)) \tag{6}$$

Here, η is called the **natural parameter** (also called the **canonical parameter**) of the distribution; $T(y)$ is the **sufficient statistic** (for the distributions we consider, it will often be the case that $T(y) = y$); and $a(\eta)$ is the **log partition function**. The quantity $e^{-a(\eta)}$ essentially plays the role of a normalization constant, that makes sure the distribution $p(y;\eta)$ sums/integrates over y to 1.

A fixed choice of T, a and b defines a *family* (or set) of distributions that is parameterized by η; as we vary η, we then get different distributions within this family.

指数族

- 指数族概念的目的，是为了说明广义线性模型Generalized Linear Models
  - 凡是符合指数族分布的随机变量，都可以用GLM回归分析

如：Bernoulli分布和高斯分布

We now show that the Bernoulli and the Gaussian distributions are examples of exponential family distributions. The Bernoulli distribution with mean ϕ, written Bernoulli(ϕ), specifies a distribution over $y \in \{0, 1\}$, so that $p(y = 1; \phi) = \phi$; $p(y = 0; \phi) = 1 - \phi$. As we varying ϕ, we obtain Bernoulli distributions with different means. We now show that this class of Bernoulli distributions, ones obtained by varying ϕ, is in the exponential family; i.e., that there is a choice of T, a and b so that Equation (6) becomes exactly the class of Bernoulli distributions.

Bernoulli分布属于指数族

We write the Bernoulli distribution as:

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right). \end{aligned}$$

Thus, the natural parameter is given by $\eta = \log(\phi/(1-\phi))$. Interestingly, if we invert this definition for η by solving for ϕ in terms of η, we obtain $\phi = 1/(1+e^{-\eta})$. This is the familiar sigmoid function! This will come up again when we derive logistic regression as a GLM. To complete the formulation of the Bernoulli distribution as an exponential family distribution, we also have

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

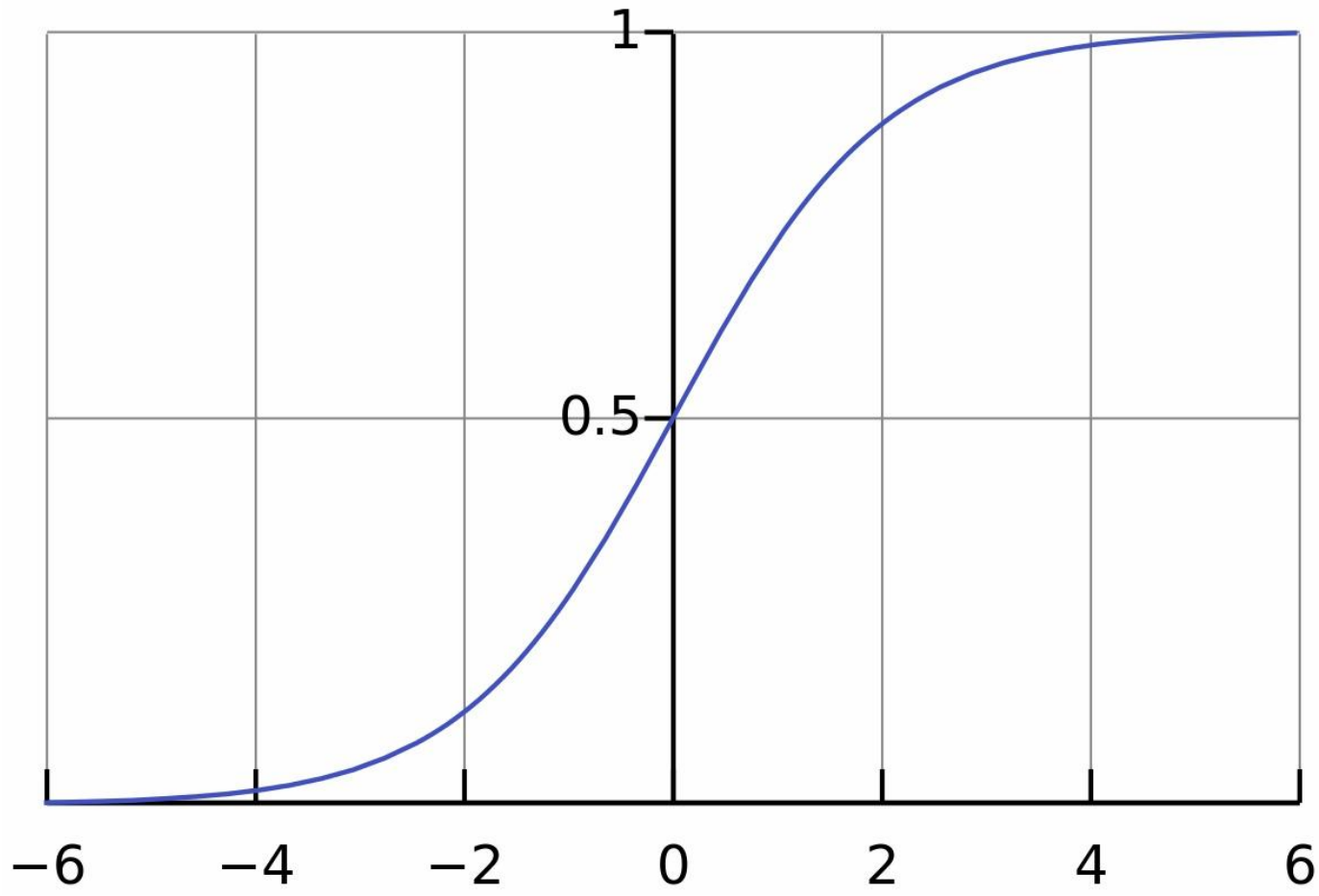
考察参数Φ

- 注意在推导过程中，出现了Logistic方程。

$$\Phi = \frac{1}{1+e^{-\eta}}$$

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

Logistic函数



Logistic函数的导数 $f(x)=\frac{1}{1+e^{-x}}$

$$f'(x)=\left(\frac{1}{1+e^{-x}}\right)'$$

$$=\frac{e^{-x}}{\left(1+e^{-x}\right)^2}$$

$$=\frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}$$

$$=\frac{1}{1+e^{-x}}\cdot\left(1-\frac{1}{1+e^{-x}}\right)$$

$$=f(x)\cdot(1-f(x))$$

- 该结论后面会用到

Gaussian分布也属于指数族分布

Lets now move on to consider the Gaussian distribution. Recall that, when deriving linear regression, the value of σ^2 had no effect on our final choice of θ and $h_\theta(x)$. Thus, we can choose an arbitrary value for σ^2 without changing anything. To simplify the derivation below, lets set $\sigma^2 = 1$. We then have:

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \cdot \exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 \\ &= \eta^2/2 \\ b(y) &= (1/\sqrt{2\pi}) \exp(-y^2/2) \end{aligned}$$

参考文献

- Prof. Andrew Ng, Machine Learning, Stanford University
- 同济大学数学教研室 主编，高等数学，高等教育出版社，1996
- 王松桂，程维虎，高旅端编，概率论与数理统计，科学出版社，2000

感谢大家！

恳请大家批评指正！

三月机器学习在线班第一课——微积分与概率论基础

本课程的 1-4 节课都是复习机器学习相关的数学知识部分，这节课主要讲了微积分与概率论部分的数学知识。

函数与极限部分

这部分介绍了极限存在的两个准则，并通过准则推导得到了极为常用的两个重要极限。

1. 两边夹定理：

$$\text{当 } x \in U(x_0, r) \text{ 时，有 } g(x) \leqslant f(x) \leqslant h(x) \text{ 成立，}$$
$$\text{并且 } \lim_{x \to x_0} g(x) = A,\ \lim_{x \to x_0} h(x) = A,\ \text{那么}$$

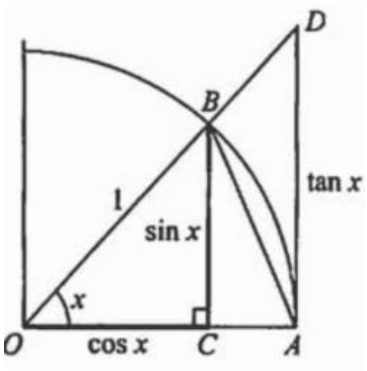
$$\lim_{x \to x_0} f(x) = A$$

这个定理很熟悉吧，高数学过的。该定理表明，当函数的表达式较为复杂，难以判定其收敛性的时候后，可以考虑将其适当放大或者缩小，只要放大与缩小后的两个函数同时收敛于同一个数值 A，那么该函数的极限存在且等于 A。

那么它有什么用呢？其中一个用途就是推出一个重要的极限：

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

我们使用几何方法+两边夹定理来推得这个极限，下面是具体过程（高数课本中可找到）：



在左图所示的单位圆中，x 表示圆心角的弧度（$\mathbf{0 < x < \pi / 2}$）。过点 A 做圆的切线与 OB 延长线交于点 D。可以知道 ΔAOB 面积< 扇形 AOB< ΔAOD 面积，有：$\frac{1}{2}\sin x < \frac{1}{2}x < \frac{1}{2}\tan x$，那么可得：$\sin x < x < \tan x$。从而有：$1 < x/\sin x < 1/\cos x$，即 $\cos x < \frac{\sin x}{x} < 1 (0 < x < \frac{\pi}{2})$。

在证明 $x \to 0$ 的极限时，必须同时证明 $x \to 0^+$ 与 $x \to 0^-$ 时候的极限。不等式 $\cos x < \frac{\sin x}{x} < 1 (0 < x < \frac{\pi}{2})$ 中，当 x 用 $-x$ 代替时，$\cos x$ 与 $\frac{\sin x}{x}$ 都不变，那么对于开区间 $(-\frac{\pi}{2}, 0)$ 内的一切 x 也是成立的，也就有：

$$\cos x < \frac{\sin x}{x} < 1 (0 < |x| < \frac{\pi}{2})$$

又有 $\lim_{x \to 0} \cos x = 1$ 和 $\lim_{x \to 0} 1 = 1$，那么根据两边夹定理可以知道 $\lim_{x \to 0} \frac{\sin x}{x} = 1$。

上面的极限将三角函数与多项式建立了极限关系，非常重要！

2. 极限存在定理：

若数列单调递增（减）且有上（下）界，则数列必有极限。（注意：这里写的是数列）。

那么该定理有啥用呢？我们来看看下面的问题吧：

证明：当$n\to\infty$时，数列$(1+\frac{1}{n})^n$有极限。

首先，让我们来证明该数列是有界的，然后再来证明它的单调性，最后求出极限。

证明数列有界的过程：

$$x_n=\left(1+\frac{1}{n}\right)^n$$

二项式定理展开

$$=1+C_n^1\frac{1}{n}+C_n^2\frac{1}{n^2}+C_n^3\frac{1}{n^3}+\cdots+C_n^n\frac{1}{n^n}$$

$$=1+n\cdot\frac{1}{n}+\frac{n(n-1)}{2!}\cdot\frac{1}{n^2}+\frac{n(n-1)(n-2)}{3!}\cdot\frac{1}{n^3}+\cdots+\frac{n(n-1)(n-2)\cdots1}{n!}\cdot\frac{1}{n^n}$$

$$=1+1+\frac{1}{2!}\cdot\left(1-\frac{1}{n}\right)+\frac{1}{3!}\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)+\cdots+\frac{1}{n!}\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{n-1}{n}\right)$$

$$<1+1+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}$$

这两处地方都选择了适当放大

$$<1+1+\frac{1}{2}+\frac{1}{2^2}+\cdots+\frac{1}{2^{n-1}}$$

$$=3-\frac{1}{2^{n-1}}$$

$$<3$$

那么我们看到，该数列是有界的，它小于 3。接下来，让我们来证明该数列的单调性（有上界，则需证明它是递增的）。可分别将$(1+\frac{1}{n})^n$和$(1+\frac{1}{n+1})^{n+1}$按二项式定理展开：

$$(1+\frac{1}{n})^n=1+1+\sum_{k=2}^{n}\frac{1}{k!}(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})$$

$$(1+\frac{1}{n+1})^{n+1}=1+1+\sum_{k=2}^{n}\frac{1}{k!}(1-\frac{1}{n+1})\cdots(1-\frac{k-1}{n+1})+(\frac{1}{n+1})^{n+1}$$

比较上面两个展开式右端对应项，显然前者较小；而且后者多出了一正数项，于是可以得到：

$$(1+\frac{1}{n})^n<(1+\frac{1}{n+1})^{n+1}$$

由此说明此数列是单调递增的。

综上所述，可以知道数列$(1+\frac{1}{n})^n$单调递增且有上界，根据单调有界准则可以知道该数列的极限必然存在。通常人们将这个数列的极限值记为 e，由大数学家欧拉首次引进这个数。

呼~ 现在是总算知道 e 是怎么来的了。让我们写下该极限：

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

3. **重要极限**：$\lim_{x\to\infty}(1+\frac{1}{x})^x=e$

这个极限很重要，我们想要证明它。需要先区分的是它与上面得到的数列极限的小小区别：一个是有理数x，一个是整数n。

证：先考虑$x\to+\infty$的情况。

对任意正数x（可设$x\geq1$），总存在唯一的正整数 n，使得$n\leq x<n+1$，于是$\frac{1}{n+1}<\frac{1}{x}\leq\frac{1}{n}$，从而有$(1+\frac{1}{n+1})<(1+\frac{1}{x})\leq(1+\frac{1}{n})$，可以得到不等式：

$$(1+\frac{1}{n+1})^n<(1+\frac{1}{x})^x\leq(1+\frac{1}{n})^{n+1}$$

又由前面数列的极限可得：

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

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

且当$n\to\infty$时候，必然有$x\to+\infty$，那么由双边夹定理，可得

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

再来考虑$x\to-\infty$的情况。思路是利用已经得到的$x\to+\infty$的定理求解。

令$x+1=-t$，当$x\to-\infty$时，有$t\to+\infty$，于是有：

$$\lim_{x\to-\infty}(1+\frac{1}{x})^x=\lim_{t\to+\infty}(1-\frac{1}{t+1})^{-(t+1)}=\lim_{t\to+\infty}(\frac{t}{t+1})^{-(t+1)}$$

$$=\lim_{t\to+\infty}(\frac{t+1}{t})^{t+1}=\lim_{t\to+\infty}(1+\frac{1}{t})^t(1+\frac{1}{t})=e$$

综上所述，可得$\lim_{x\to\infty}(1+\frac{1}{x})^x=e$。

导数与微分部分

1. 导数

简单地说，一阶导数就是曲线的斜率，是曲线变化快慢的反应。而二阶导数是斜率变化快慢的反应，表征曲线的凹凸性（二阶导数与后面要解释的凸函数有关联）。

说起导数，我们都并不陌生，就让我们来回忆一下一些常用函数的导数：

（1）$C'=0$（C 为常数）；　（2）$(x^n)'=nx^{n-1}$（$n\in Q$）；

（3）$(\sin x)'=\cos x$；　（4）$(\cos x)'=-\sin x$；

（5）$(a^x)'=a^x\ln a$；　（6）$(e^x)'=e^x$；

（7）$(\log_a x)'=\frac{1}{x}\log_a e$；　（8）$(\ln x)'=\frac{1}{x}$.

同样不应该忘记的还有几个求导常用的关系式：

$$(u+v)'=u'+v'$$

$$(uv)'=u'v+uv'$$

2. 幂指函数的一般处理策略

问题：已知幂指函数 $f(x)=x^x, x>0$，求 $f(x)$ 的最小值。

解：

$$t=x^x$$

$$\rightarrow \ln t=x\ln x$$

步骤 1：两边取对数

$$\xrightarrow{\text{两边对x求导}} \frac{1}{t}t'=\ln x+1$$

步骤 2：两边对 x 求导

$$\xrightarrow{\text{令}t'=0} \ln x+1=0$$

$$\rightarrow x=e^{-1}$$

步骤 3：求极值

P.S. 本应该先判断二阶导从而判断导数的极大极小，此处忽略

$$\rightarrow t=e^{-\frac{1}{e}}$$

X 在 x0 处的 n 阶段无穷小

3. Taylor 公式 – Maclaurin 公式

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$$

$$f(x)=f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n+o(x^n)$$

Taylor 公式想必大家都听过，但是你会使用吗？反正我自己在之前已经忘了差不多了（本人研二，数学渣渣）。重新听讲一遍非常有必要呀~

Taylor 公式表征的是函数的一阶导数、二阶导数以及各高阶导数与函数本身的关系。Taylor 公式如何得到的我们暂且不究，来看看 Taylor 公式的实际应用吧。邹博老师给出了一个问题：如何运用 Taylor 公式求得：给定正实数 x 下的 e^x 的值？

老师提供的一种思路是这样的：求出最大整数 k 和小数 r 的值，使得

$$x=k*\ln 2+r,|r|<=0.5*\ln 2$$

从而得到：$e^x=e^{k\bullet\ln 2+r}=e^{k\bullet\ln 2}\cdot e^r=2^k\cdot e^r$。由于 2^k 非常容易求解得到，那么我们只需要求解 e^r 即可，对于 e^r，我们可以使用 Taylor 公式在原点处的展开式求得近似值（为什么是在原点展开？因为希望取得最大整数 k 下对应的小数 r，同时 r 有范围限制（在原点附件））。

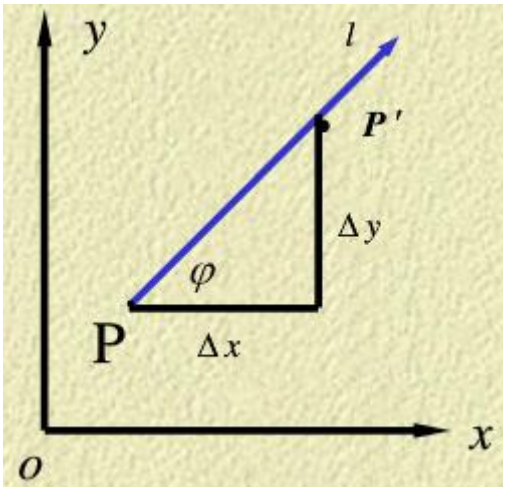
答疑：同学们会好奇的一点：为什么不直接在 e^x 中使用 Taylor 公式在 x 处的展开式然后取前面 n 项去估计，而选择在一个更小的 e^r 上使用 Taylor 公式在原点的展开式呢？

邹博老师的解释是：直接在 e^x 中使用 Taylor 公式在 x 处的展开式误差比较大（如直接将 x=100 代入到 e^x 的 Taylor 展开公式中取前面几项，与真实值比较相差较大）。而在原点处展开的误差会小一些，估计地更精准。

遗留问题：在函数估计的时候，使用 Taylor 公式在何处展开是否有研究？或者说，试验 Taylor 公式展开时候是否需要注意在何处展开呢？

4. 方向导数与梯度

聊完了一元函数的导数，让我们来看二元函数的导数相关的知识。首先引出的是方向导数。方向导数谈论的是函数 $z=f(x,y)$ 在一点 P 上沿着某一方向的变化率问题。为了说明该问题，让我们来看看右图的例子：设函数 $z=f(x,y)$ 在点 $p(x,y)$ 的某一领域 $U(p)$ 内有定义，自点 p 引射线 l。假设 x 轴正方向到射线 l 的转角为 φ，并假设 $p'(x+\Delta x,y+\Delta y)$ 为射线 l 上的另一个点，且 $p'\in U(p)$。



由此得：$|pp'|=\rho=\sqrt{(\Delta x)^2+(\Delta y)^2}$

且设 $\Delta z=f(x+\Delta x,y+\Delta y)-f(x,y)$

需要考虑的一点是：当 p' 沿着 l 趋近于 p 时，极限 $\lim\limits_{\rho\to 0}\dfrac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}$ 是否存在？

如果此极限存在，我们就称该极限为函数 $z=f(x,y)$ 在点 p 沿方向 l 的方向导数。记为：

$$\frac{\partial f}{\partial l}=\lim_{\rho\to 0}\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}$$

关于方向导数的定理如下：

如果函数函数 $z=f(x,y)$ 在点 p 是可微分的，那么函数在该点沿任意方向 l 的方向导数都存在，而且有：

$$\frac{\partial f}{\partial l}=\frac{\partial f}{\partial x}\cos\varphi+\frac{\partial f}{\partial y}\sin\varphi$$

其中 φ 为 x 轴正方向到射线 l 的转角（如上图所示）。

此式子的证明不难，见下面的证明部分：

证明：由于函数可微，则增量可以表示为：

$$f(x+\Delta x,y+\Delta y)-f(x,y)=\frac{\partial f}{\partial x}\Delta x+\frac{\partial f}{\partial y}\Delta y+o(\rho)$$

两边同时除以 ρ，可得：

$$\frac{f(x+\Delta x,y+\Delta y)-f(x,y)}{\rho}=\frac{\partial f}{\partial x}\frac{\Delta x}{\rho}+\frac{\partial f}{\partial y}\frac{\Delta y}{\rho}+\frac{o(\rho)}{\rho}$$

看上面的图像可以知道：$\frac{\Delta x}{\rho}=\cos\varphi,\frac{\Delta y}{\rho}=\sin\varphi$，由此可以得到定理中的公式。得证。

有了方向导数之后，我们好奇的是：函数 $z=f(x,y)$ 在点 p 沿那一个方向增加的速度最快呢？我们令方向增加最快的方向为函数在该点的梯度。由此引出了梯度的定义：

设函数 $z=f(x,y)$ 在平面区域 D 内具有一阶连续偏导数，则对于每一点 $p(x,y)\in D$，向量

$$(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y})$$

为函数 $z=f(x,y)$ 在点 P 处的梯度，记为 $grad\ \mathrm{f}(x,y)$。

梯度的方向就是函数在该点变化最快的方向，确切的说，应该是梯度的方向及其反方向是函数在该点变换最快的方向（一个增加最快，一个减小最快）。

梯度相关的概念在我们今后的学习中会用到，其中梯度下降法就是基于上面的知识来讲解的。

---

5. 凸函数

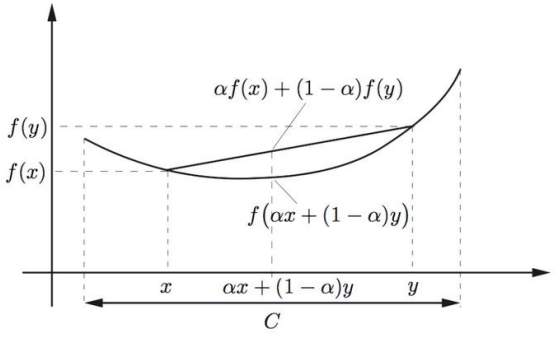
在机器学习的很多情形里（logistic 回归，SVM 等），我们希望求得某些函数的最优值。但是通常情况下找到函数的全局最优是不容易的。有一类特殊的函数——凸函数，我们可以通过相应的研究进而非常高效地找到该类函数的最优值。下面仅对凸函数做简单介绍，更详细的部分会在接下来课程凸优化部分进行详细讲解。

凸函数的定义：

函数 $f(x)$ 在区间 I，如果对 I 上的任意两点 x,y，恒有一下不等式成立：

$$f(\theta x+(1-\theta)y)\le\theta f(x)+(1-\theta)f(y),\text{其中}\theta\in[0,1]$$

则称 $f(x)$ 在区间 I 上是凸的。



直观上理解的话，可以看看右边的图。方便记忆，群友提到了一个方面识别凸函数的方式：一个正立着的碗的曲线就是凸函数，一个倒立着的碗的曲线就是凹函数。

凸函数的判定：如果 $f(x)$ 在区间[a,b]上连续，在（a,b）内二阶可导，那么：

如果 $f''(x)>0$，则 $f(x)$ 是凸的；

如果 $f''(x)<0$，则 $f(x)$ 是凹的；

假如 $f(x)$ 为凸函数，还可对上面不等式进行拓展：

$$f(\theta_1 x_1 + \cdots + \theta_n x_n) \le \theta_1 f(x_1) + \cdots + \theta_n f(x_n),$$

其中 $\theta_i \in [0,1], \theta_1 + \cdots + \theta_n = 1$。

实际上，上面的不等式可以拓展到无限项求和或者积分的情况，我们可以把该不等式写成积分的形式就是：

$$f(\int p(x)x dx) \le \int p(x) f(x) dx, \text{其中} \int f(x)dx = 1, \text{且} p(x) \ge 0, \forall x.$$

上式可以改写成：

$$f(E(x)) \le E[f(x)]$$

这个式子就是 Jensen 不等式，前提是函数 $f(x)$ 为凸函数。在接下来的课程上我们会遇到该不等式并需要对它加以应用。

凸函数的应用：

1. 假设 $p(x)$、$q(x)$ 是在随机变量 X 上取值的两个概率分布，定义 $D(p \| q)$ 的表达式如下：

$$D(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

试证明：$D(p \| q) \ge 0$。

证明： $D(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_x p(x) \log \frac{q(x)}{p(x)}$

我们使用凸函数本身性质的拓展不等式来求解，令 $\theta_i = p(x_i), f(t_i) = \log t_i, \text{其中} t_i = \frac{q(x)}{p(x)}$。

则上面的推导继续：$-\sum_x p(x) \log \frac{q(x)}{p(x)} \ge -\log \sum_x p(x) \frac{q(x)}{p(x)} = -\log \sum_x q(x) = -\log 1 = 0$。

问题得证。

概率论部分

作为引子，邹博老师首先提出了一个问题：如果事件出现的概率为0，那么是否说明该事件一定不会发生呢？

答案是否定的。老师的给出了一个解释：一个桌面，假设其面积为1，我有一根针，针头是一个良好定义的点，那么我把针头扎向桌面任意一个点，这件事发生的概率为0，原因是点的面积为0。但是实际上我将针头扎向桌面，肯定会扎到某一个点。

这个答案我一开始也没想通，后来慢慢才通过了。有没有更好解释的例子呢，查了网上的答案，知乎上有这个问题的讨论：http://www.zhihu.com/question/20208198。大家可以看看其中的内容。

首先我们来看待一个概率论中的简单部分——古典概型。在这个模型下，随机实验所有可能的结果是有限的，并且每个基本结果发生的概率是相同的（即等概率事件）。古典概型主要涉及到排列组合方面的知识。

举例：将 n 个不同的球放在 $N(N \geq n)$ 个盒子里，假设何止的容量无限，求事件 A ={每个盒子至多一个球}的概率。

解：其实这一类问题有一个统一的解法：首先求基本事件的总数，然后求相应事件 A 的事件数。我们可以做如下考虑：

基本事件总数：

- 第1个球，有N种放法；
- 第2个球，有N种放法；
- ……
- 共：N^n种放法。

每个盒子至多放1个球的事件数：

- 第1个球，有N种放法；
- 第2个球，有N-1种放法；
- 第3个球，有N-2种放法；
- ……
- 共：$N(N-1)(N-2)\cdots(N-n+1)=P_N^n$

于是该问题的解就是：$P(A)=\dfrac{P_N^n}{N^n}$。

另一个相似的问题是这样的：某班上有50位同学，求至少有2个生日相同的概率是多少呢？

对待这个问题，我们可以先求没有两个人生日相同的概率，然后用1减去该概率即可求得问题的答案。实际上求解没有两个人生日相同的概率，就相当于上题中每个盒子至多有一个球的概率，只不过此时我们的盒子是365天。而我们需要把50位同学塞进这365个盒子里，并满足每个盒子的人数不超过一个。依据上面的公式求得这个概率为 $\dfrac{P_{365}^{50}}{365^{50}}$。而

$1-\frac{P_{365}^{50}}{365^{50}}$ 就是问题的答案。

再来一类古典概型的例子：装箱问题。问题如下：

将 12 个正品和 3 个次品随机装在 3 个箱子中（每个箱子都有 5 个）。每个箱子恰有一个次品的概率是多少？

解：首先，我们来看看总事件数：将 15 个产品装进 3 个箱子，每箱装 5 个，总共有 $C_{15}^5 C_{10}^5 C_5^5$ 种装法。对于事件 A ={每个箱子恰有一个次品}而言，我们这样考虑：先把 3 个次品放入 3 个箱子，共有 $3!$ 中装法。对于这样的每种装法，把剩余 12 件产品装入 3 个箱子，每箱装 4 件，可以得到 $C_{12}^4 C_8^4 C_4^4$ 种装法。那么 $P(A)=(3!*C_{12}^4 C_8^4 C_4^4)/(C_{15}^5 C_{10}^5 C_5^5)$ =25/91。

虽然古典概型在机器学习中应用不多，但是这是基本的知识，还是有必要掌握的~~

----------我是分割线 QAQ----------

下面我们来讲讲概率论当中比较重要的知识点：条件概率、全概率公式以及贝叶斯公式。这部分内容大家都比较熟悉，所以就只是贴上数学教科书上面的文字了，权当复习吧。但是相关的应用还是必须要会的，接下来会有对应的应用题。

条件概率：

定义 设 A,B 是两个事件，且 $P(A)>0$，称

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

为在事件 A 发生的条件下事件 B 发生的**条件概率**.

乘法定理：

（二）乘法定理

由条件概率的定义 (5.2)，立即可得下述定理.

乘法定理 设 $P(A)>0$，则有

$$P(AB)=P(B|A)P(A). \tag{5.3}$$

(5.3)式称为**乘法公式**.

(5.3) 式容易推广到多个事件的积事件的情况. 例如，设 A,B,C 为事件，且 $P(AB)>0$，则有

$$P(ABC)=P(C|AB)P(B|A)P(A). \tag{5.4}$$

在这里，注意到由假设 $P(AB)>0$ 可推得 $P(A)\geqslant P(AB)>0$.

一般，设 $A_1,A_2,\cdots,A_n$ 为 n 个事件，$n\geqslant 2$，且 $P(A_1A_2\cdots A_{n-1})>0$，则有

$$P(A_1A_2\cdots A_n)=P(A_n|A_1A_2\cdots A_{n-1})P(A_{n-1}|A_1A_2\cdots A_{n-2})\cdots P(A_2|A_1)P(A_1).$$

全概率公式和贝叶斯公式：

定义 设 S 为试验 E 的样本空间，$B_1, B_2, \cdots, B_n$ 为 E 的一组事件. 若

(i) $B_iB_j=\varnothing, i\neq j, i,j=1,2,\cdots,n$;

(ii) $B_1 \cup B_2 \cup \cdots \cup B_n = S$,

则称 $B_1, B_2, \cdots, B_n$ 为样本空间 S 的一个**划分**.

若 $B_1, B_2, \cdots, B_n$ 是样本空间的一个划分，那么，对每次试验，事件 $B_1, B_2, \cdots, B_n$ 中必有一个且仅有一个发生.

定理 设试验 E 的样本空间为 S，A 为 E 的事件，$B_1, B_2, \cdots, B_n$ 为 S 的一个划分，且 $P(B_i)>0\ (i=1,2,\cdots,n)$，则

$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\cdots+P(A|B_n)P(B_n). \tag{5.6}$$

(5.6) 式称为**全概率公式**.

该公式适用于某些问题中 P(A)不容易直接求得，但是却容易找到 S 的一个划分 B_1，B_2，…，且 $P(B_i)$和 $P(A|B_i)$为已知或者容易求得的情况。

另一个重要的公式就是下面的贝叶斯公式：

定理 设试验 E 的样本空间为 S. A 为 E 的事件，$B_1, B_2, \cdots, B_n$ 为 S 的一个划分，且 $P(A)>0, P(B_i)>0\ (i=1,2,\cdots,n)$，则

$$P(B_i \mid A)=\frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n}P(A \mid B_j)P(B_j)}, \quad i=1,2,\cdots,n. \tag{5.7}$$

(5.7) 式称为**贝叶斯(Bayes)公式**①.

证 由条件概率的定义及全概率公式即得

$$P(B_i \mid A)=\frac{P(B_iA)}{P(A)}=\frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n}P(A \mid B_j)P(B_j)}, i=1,2,\cdots,n. \qquad \square$$

思考题：8 支步枪中有 5 支已经校准过，3 支未校准。一名射手拿校准过的枪射击，中靶概率为 0.8；用未校准的枪射击，中靶概率为 0.3；现在从 8 支枪当中随机取出一支射击，结果中靶，求该枪是已经校准过的概率。

解：令 A 表示校准的枪支，B 表示中靶。那么我们可以得到：

$$P(A)=5/8, P(\bar{A})=3/8, P(B|A)=0.8, P(B|\bar{A})=0.3$$

现在要求的是 $P(A|B)$ 的概率。由贝叶斯公式可得：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}=81.63\%$$

接下来复习的是各种常见分布本身的统计量（期望、方差）

常见分布包括：

离散型概率分布：两点分布、二项分布、泊松分布；

连续型概率分布：均匀分布、指数分布、高斯分布。

两点分布（该实验也称为伯努利实验）：

已知随机变量 X 的分布律为：

X	1	0
p	p	$1-p$

那么我们可以求得该离散分布的期望为：$E(X)=1\cdot p+0\cdot(1-p)=p$

方差为：$D(X)=E(X^2)-[E(X)]^2=1^2\cdot p+0^2\cdot(1-p)-p^2=pq$ 。

二项分布：

将伯努利实验单独重复地进行 n 次，称这一串重复的独立实验为 n 重伯努利实验。这里的“重复”是指在每次试验中 $P(A)=p$ 保持不变；“独立”指的是每次试验结果互不影响。在抛 n 次硬币的试验中，令随机变量 X 表示人头（好吧，在我国这里是玫瑰花）朝上的次数，那么令 $f(x)=\mathrm{P}(X=x)$ 为概率质量函数（相对的，连续型的称为概率密度函数），则：

$$f(x)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x}, x=0,\cdots n\\ 0, els\end{cases}$$

如果某一个随机变量的概率质量函数如上式所述，那么我们称这类随机变量为二项分布随机变量，可以表示为 $X\sim Binomial(n,p)$，即随机变量 X 服从参数为 (n,p) 二项分布。现在我们来求解二项分布的期望和方差。

$$E(X)=\sum_{k=0}^{n}k\cdot p(X=x)=\sum_{k=0}^{n}k\cdot\binom{n}{k}p^k(1-p)^{n-k}$$

K=0 项为 0，所以忽略

$$=\sum_{k=0}^{n}k\cdot\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}=np\sum_{k=1}^{n}\cdot\frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k}$$

设 $a=k-1, b=n-1$，那么

所有二项分布概率质量函数之和为 1

$$=np\sum_{a=0}^{b}\frac{b!}{a!(b-a)!}p^a(1-p)^{b-a}=np\sum_{a=0}^{b}\binom{b}{a}p^a(1-p)^{b-a}=np$$

泊松分布：

此处老师推导泊松分布的期望的过程比较有趣，他是从 Taylor 公式展开来推导的，我们来具体看看：

将函数 e^x 在 x 点处的 Taylor 公式展开：

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

两侧同时乘以 e^{-x}，可以得到：

$$1 = 1 \cdot e^{-x} + x \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \cdots + \frac{x^k}{k!} \cdot e^{-x} + R_k \cdot e^{-x}$$

我们可以看到，上面展开式的每一项 $\frac{x^k}{k!} \cdot e^{-x}$ 就是泊松分布的概率质量函数，只不过此处的 $x = \lambda$。那么我们尝试使用上面的结论来推导泊松分布的期望。

设 $X \sim \pi(\lambda)$，且分布律为：

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, k = 0, 1, 2, \cdots, \lambda > 0$$

则有：

K=0 项为 0，所以忽略　　使用上面的推到结果

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \left(\frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \right) = \lambda \cdot 1 = \lambda$$

在课上，当老师推导出来的时候，我们都觉得这个推导太流弊了，真的是学习了!!

接下来推导泊松分布的方差：

$$E(X^2) = E[X(X-1) + X]$$

$$= E[X(X-1)] + E(X)$$

$$= \sum_{k=0}^{+\infty} k(k-1) \cdot \frac{\lambda^k}{k!} e^{-\lambda} + \lambda$$

$$= \lambda^2 e^{-\lambda} \sum_{k=2}^{+\infty} \cdot \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda.$$

所以 $D(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

我们可以看到，泊松分布的期望和方差都是参数 λ。

课后拓展：

泊松定理（用泊松分布来逼近二项分布的定理）：

设 $\lambda>0$ 是一个常数， n 是任意一个正整数，设 $np_n=\lambda$ （在二项分布（n,p）中 p 即为 p_n ），则对于任一固定的非负整数 k ，有：

$$\lim_{n\to\infty}\binom{n}{k}p_n{}^k(1-p_n)^{n-k}=\frac{\lambda^k}{k!}\cdot e^{-\lambda}$$

定理的条件 $np_n=\lambda$ （常数）意味着当 n 很大时 p_n 必然很小，因此，上述定理表明当 n 很大且 p 很小时有以下近似式：

$$\binom{n}{k}p^{\,k}(1-p\,)^{n-k}\approx\frac{\lambda^k}{k!}\cdot e^{-\lambda}(\text{其中}\lambda=np)$$

也就是说以 n,p 为参数的二项分布的概率值可以由参数为 $\lambda=np$ 的泊松分布的概率值来近似。当然这必须满足一定的条件： n 很大且 p 很小.

具体证明过程：

证 由 $p_n=\frac{\lambda}{n}$，有

$$\binom{n}{k}p_n^k(1-p_k)^{n-k}=\frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k(1-\frac{\lambda}{n})^{n-k}$$

$$=\frac{\lambda^k}{k!}[1\cdot(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})](1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k}.$$

对于任意固定的 k，当 $n\to\infty$ 时

$$1\cdot(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})\to1,(1-\frac{\lambda}{n})^n\to e^{-\lambda},(1-\frac{\lambda}{n})^{-k}\to1.$$

故有 $$\lim_{n\to\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k}=\frac{\lambda^k e^{-\lambda}}{k!}.$$

均匀分布：

设 $X \sim U(a,b)$,其概率密度为：

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b,\\ 0, & \text{其他.}\end{cases}$$

那么可以得到：

$$E(X)=\int_{-\infty}^{+\infty} xf(x)dx=\int_a^b x\cdot\frac{1}{b-a}dx=\frac{1}{2}(a+b)$$

$$D(X)=E(X^2)-[E(X)]^2=\int_a^b x^2\cdot\frac{1}{b-a}dx-[\frac{1}{2}(a+b)]^2=\frac{(b-a)^2}{12}$$

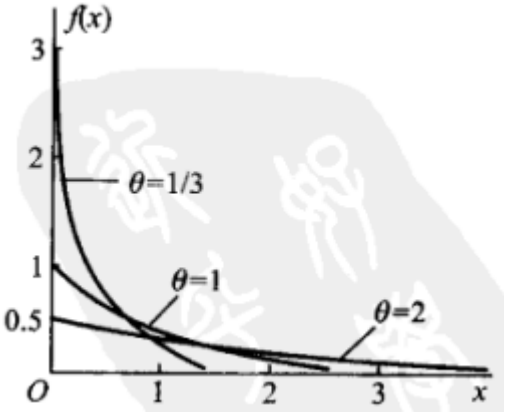
（额，这个没啥有趣的，仅仅写下来罢了，为了齐全~）

指数分布：

假设随机变量 X 服从指数分布，它的概率密度函数可以写成两个形式：

$$f(x)=\begin{cases}\dfrac{1}{\theta}e^{-x/\theta}, x>0, \text{其中}\theta>0\\ 0, x\le 0.\end{cases} \quad \text{and} \quad f(x)=\begin{cases}\lambda e^{-\lambda x}, x>0, \text{其中}\lambda>0\\ 0, x\le 0.\end{cases}$$

其中 $\lambda>0$ 被称为率参数，即每单位时间内发生某事件的次数。简单记为：$X \sim Exponential(\lambda)$。可以比较容易得到其概率密度函数 $f(x)$ 的几个不同参数图形，如右图所示。



指数分布一个重要特征是无记忆性。如果一个随机变量 X 服从指数分布，当 $s,t\ge 0$ 有：

$$p(x>s+t\,|\,x>s)=p(x>t)$$

也就是说，如果 X 表示的是某一元器件（老师课堂上举例为手电筒 额...）的寿命，已知该手电筒使用了 s 个小时，它在这个基础上还可以使用至少 t 小时（即总共使用 s+t 小时）的条件概率，与从开始使用时候算起它使用至少 t 小时的概率是相等的。挺神奇的！

我们来计算指数分布的期望和方差，具体过程如下：

$$E(X)=\int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x=\int_0^{+\infty} x\cdot\frac{1}{\theta}\mathrm{e}^{-x/\theta}\,\mathrm{d}x$$

$$=-x\mathrm{e}^{-x/\theta}\Big|_0^{+\infty}+\int_0^{+\infty}\mathrm{e}^{-x/\theta}\,\mathrm{d}x \quad =\theta$$

$$D(X)=E(X^2)-[E(X)]^2=\int_0^{+\infty}x^2\cdot\frac{1}{\theta}e^{-x/\theta}\,dx-\theta^2$$
$$=2\theta^2-\theta^2\quad=\theta^2$$

正态分布：

好了，终于到了非常非常重要的角色要出场了~~对了，就是正态分布，也称为高斯分布。在机器学习领域正态分布的使用真的是非常广泛。

%>_<%，我竟然不知道如何写知识点了，好吧，太宽泛了的感觉，就空着吧 O__O"…

----------我是分割线啊 QAQ ----------

指数分布族与广义线性模型

实际上，之前我们所描述的两点分布、高斯分布等，都是一种更广泛的算法的特例——指数分布族。

首先我们给出指数分布族的概率表达式：

$$p(y;\eta)=b(y)\exp(\eta^T T(y)-a(\eta))$$

这里，η 表示自然参数（natural parameter），$T(y)$ 为充分统计量（sufficient statistic，通常情况 $T(y)=y$ ）．$a(\eta)$ 为 log partition function(不会翻译，也没找到相关材料额)。在给定 a,b,T 的前提下，这个公式就定义了一个以 η 为参数的概率分布集合，当我们对 η 取不同的值时，得到的概率分布集合 $p(y;\eta)$ 是不同的。

我们来证明高斯分布和两点分布其实都是指数分布族的成员，即可以通过转化找到符合以上表达式格式的 a,b,T 。

两点分布：

$$P(y;\phi)=\phi^y(1-\phi)^{1-y}$$
$$=\exp(y\log\phi+(1-y)\log(1-\phi))$$
$$=\exp((\log\frac{\phi}{1-\phi})y+\log(1-\phi))$$

对比一下，我们得到：$\eta=\log\frac{\phi}{1-\phi},T(y)=y,a(\eta)=-\log(1-\phi),b(y)=1$ 。有趣的一点是：在 $\eta=\log\frac{\phi}{1-\phi}$ 表达式中，我们转换为用 η 表示 ϕ 的方式，可得 $\phi=\frac{1}{1-e^{-\eta}}$；该表

达式和 logistic 函数很相似不是么？到底有什么玄机呢，我们接着往下看吧。将 $\phi = \frac{1}{1-e^{-\eta}}$ 代入 $a(\eta) = -\log(1-\phi)$ 中，可得 $a(\eta) = \log(1+e^{\eta})$，那么最终的指数分布族参数为：$T(y) = y, a(\eta) = \log(1+e^{\eta}), b(y) = 1$。

高斯分布：

为简单化，我们令 $\sigma^2 = 1$，此时的概率密度函数

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y-\mu)^2)$$
$$= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) \cdot \exp(\mu y - \frac{1}{2}\mu^2)$$

对比一下，得到：$\eta = \mu, T(y) = y, a(\eta) = \frac{1}{2}\mu^2 = \frac{1}{2}\eta^2, b(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)$。

实际上，我们可以知道，指数分布族的成员远不仅仅这两个，还包括：泊松分布、多元正态分布、gamma 分布、Dirichlet 分布等。有了指数分布族的各个参数表示，实际上我们就可以推出广义线性模型。

在机器学习的问题中，通常会遇到如下问题（分类问题或回归问题）：已知独立变量 y 是 x 的函数，我们只有 x 的样本，如何估计 y 的值呢？实际上广义线性模型可以解决该类问题。当然，前提是我们的问题符合广义线性模型的三个假设：

（1） $y \mid x;\theta \sim ExponentialFamily(\eta)$。即在给定 x 的前提下（θ 为参数），随机变量 y 服从某一指数分布族分布（参数为 η）。

（2）给定 x，我们的目标是预测 $E[T(y) \mid x]$，也就是说预测输出为 $h(x) = E[T(y) \mid x]$。（之前说过，通常意义下 $T(y) = y$）

（3）指数分布族的参数 η 与输入 x 之间存在线性关系，即 $\eta = \theta^T x$。

我们先来看看二值分类问题（满足两点分布）如何通过广义线性模型得到预测输出。在二值分类问题上，有 $y \in \{0,1\}$。我们知道其满足两点分布，同时 $y \mid x;\theta \sim Bernoulli(\phi)$，即有 $E[y \mid x;\theta] = \phi$。还记得在由两点分布推导指数分布族的过程中，我们有 $\phi = \frac{1}{1-e^{-\eta}}$，那么我们可以得到：

$$h(x)=E[T(y)\mid x;\theta]=E[y\mid x;\theta]=\phi$$

$$\frac{1}{1-e^{-\eta}}=\frac{1}{1-e^{-\theta^T x}}$$

第一个等式是源于广义线性模型的第二个假设；第二个等式是两点分布转化为指数分布族中所得结论$T(y)=y$；第三个等式是两点分布自身性质所致；第四个等式也是转化为指数分布族所得结论$\phi=\frac{1}{1-e^{-\eta}}$；第四个等式是广义线性模型的第三个假设。

好了，为什么解决二值分类问题使用 logistic 模型，可以从这里找到根据。哈哈，有点神奇是么？

接下来再来尝试使用广义线性模型来解决房屋价格预测问题（网易公开课《机器学习》第一课时提到的问题，当时 Ng 老师直接使用了线性模型来解决）。在房价问题上，用 y 来表示房屋价格，而输入 x 表示的是房屋面积（最简单的输入了哈）；假设影响房价的原因是多方面的，根据中心极限定理，可以得到$y\mid x;\theta\sim \mathrm{N}(\mu,\sigma^2)$。而之前我们就得到正态分布是指数分布族的成员，并且已经知道其对应的参数$\eta=\mu,T(y)=y$。那么我们可以得到：

$$h(x)=E[T(y)\mid x;\theta]=E[y\mid x;\theta]$$
$$=\mu=\eta=\theta^T x$$

第一个等式是源于广义线性模型的第二个假设；第二个等式是正态分布转化为指数分布族中所得结论$T(y)=y$；第三个等式是正态分布自身性质所致；第四个等式也是转化为指数分布族所得结论$\eta=\mu$；第四个等式是广义线性模型的第三个假设。

那么，也就可以通过广义线性模型来解释为什么使用线性模型来解决房价问题是合理的了。

广义线性模型的建立更像是搭建了一个系统，一个黑箱，直接给出一个预测输出 h(x)模型，进而我们可以通过各种方法求解问题的想要参数。