微积分与概率论基础

七月算法 邹博

2015年10月11日

主要内容

- 本课程示例概述
- 复习数学分析
- 概率论基础

什么是机器学习

- 对于某给定的任务T，在合理的性能度量方案P的前提下，某计算机程序可以自主学习任务T的经验E；随着提供合适、优质、大量的经验E，该程序对于任务T的性能逐步提高。
- 这里最重要的是机器学习的对象：
  - 任务Task,T，一个或者多个
  - 经验Experience,E
  - 性能Performance,P
- 即：随着任务的不断执行，经验的累积会带来计算机性能的提升。
  - McGraw-Hill, 1997

换个表述

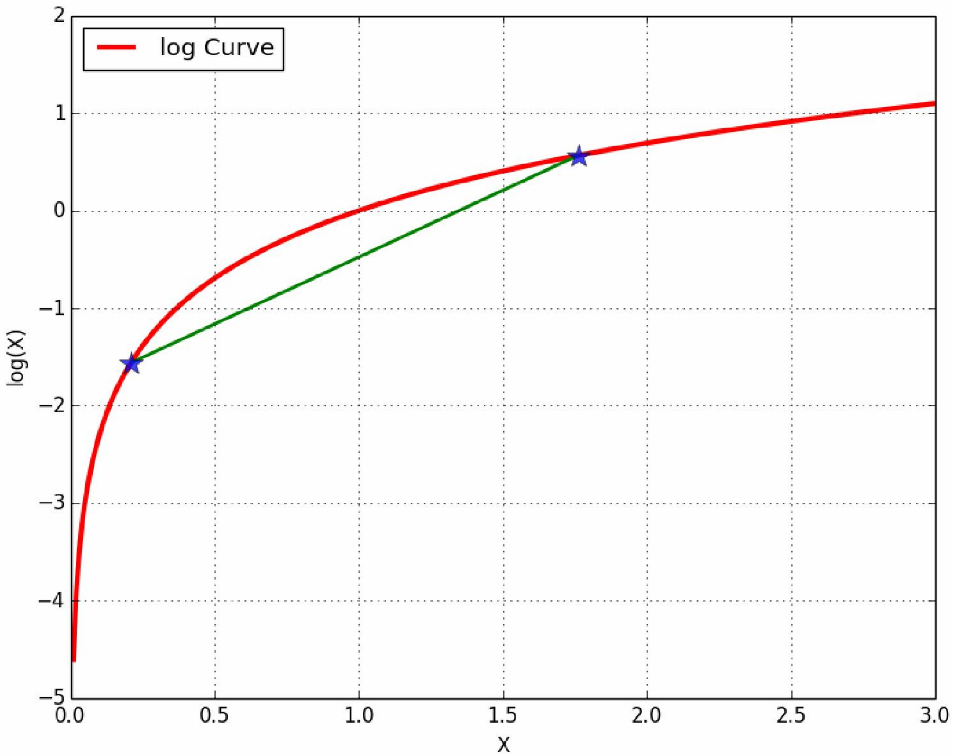
- 机器学习是人工智能的一个分支。我们使用计算机设计一个系统，使它能够根据提供的训练数据按照一定的方式来学习。随着训练次数的增加，这个系统可以在性能上不断学习和改进，通过优化该学习模型，能够基于先前学习得到的参数来预测相关问题的输出。
- 思考：
  - 如何设计无人驾驶机动车？

无人驾驶汽车

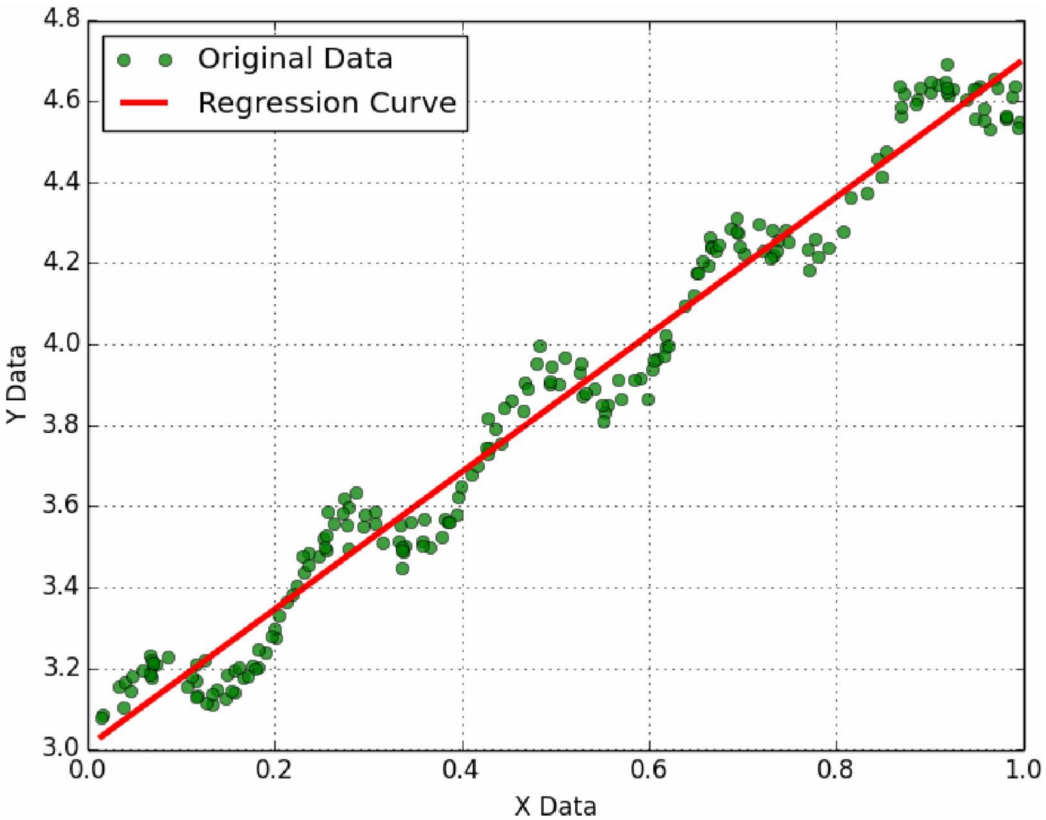
- 汽车的无人汽车模块已经成熟：全自动公共交通工具已经出现在了世界上的多个城市。
- Lutz探路者/CYCAB/Google

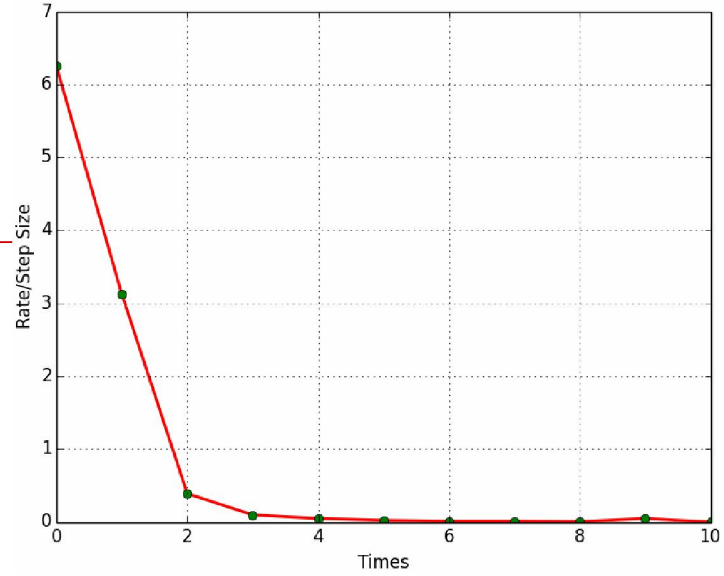
Python Code示例

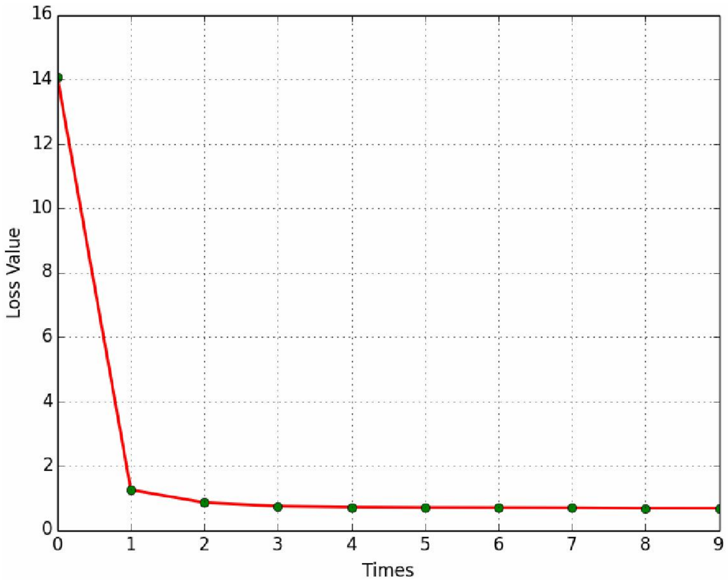
```python
# -*- coding:utf8 -*-

import math
import matplotlib.pyplot as plt

if __name__ == "__main__":
    x = [float(i)/100.0 for i in range(1,300)]
    y = [math.log(i) for i in x]
    plt.plot(x, y, 'r-', linewidth=3, label='log Curve')
    a = [x[20], x[175]]
    b = [y[20], y[175]]
    plt.plot(a, b, 'g-', linewidth=2)
    plt.plot(a, b, 'b*', markersize=15, alpha=0.75)
    plt.legend(loc='upper left')
    plt.grid(True)
    plt.xlabel('X')
    plt.ylabel('log(X)')
    plt.show()
```



线性回归、rate、Loss

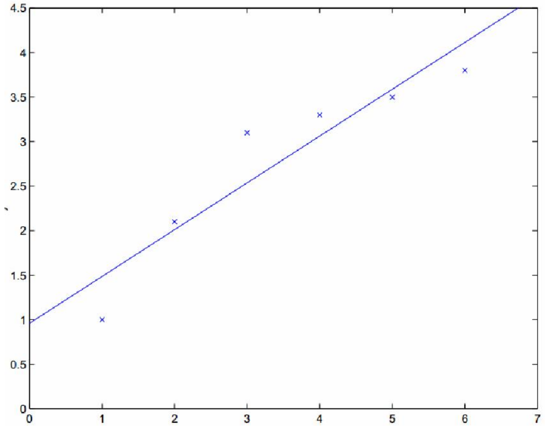


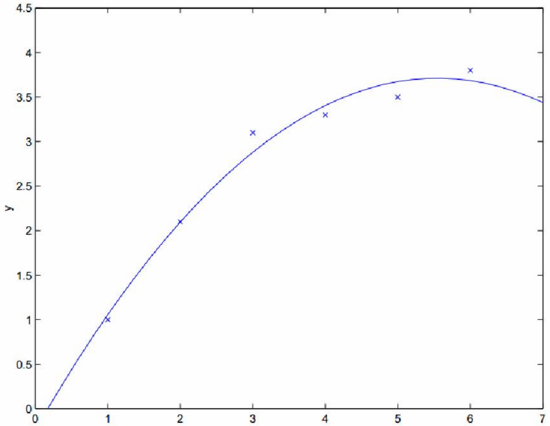


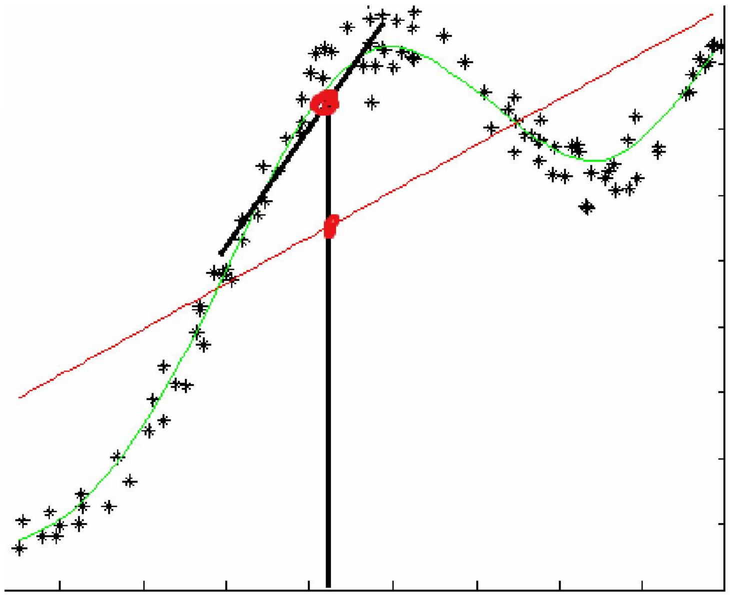


线性回归的进一步思考

- 误差假设：高斯分布、两点分布
- 线性的含义：对参数θ线性
- 局部加权回归

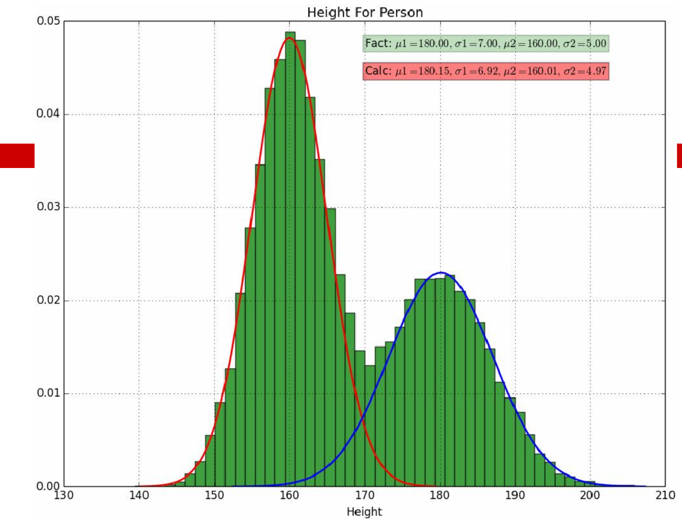


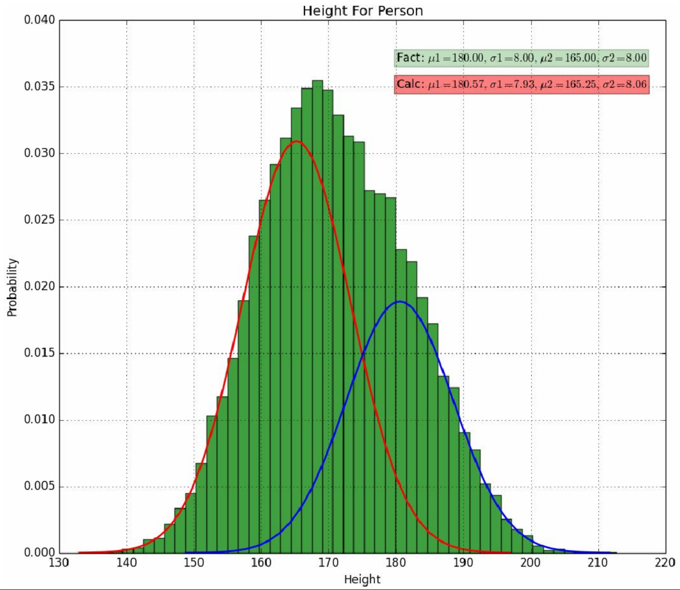




$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

EM Code

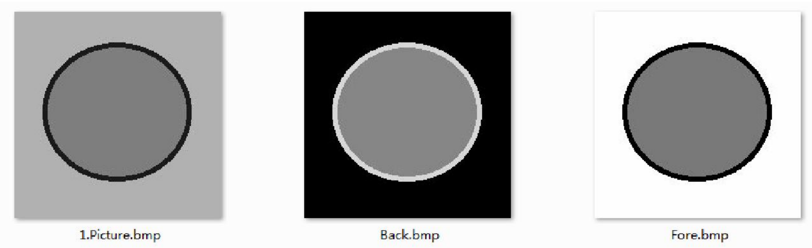


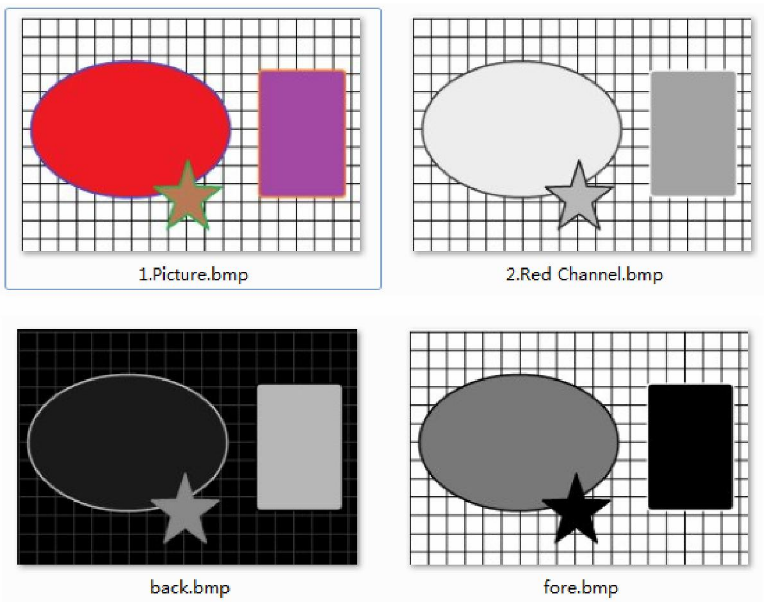


em3.py

```python
def calcEM(height):
    N = len(height)
    gp = 0.5    #girl probability
    bp = 0.5    #boy probability
    gmu,gsigma = min(height),1  #先验：直接取最大和最小值
    bmu,bsigma = max(height),1
    ggamma = range(N)
    bgamma = range(N)
    cur = [gp, bp, gmu, gsigma, bmu, bsigma]
    now = []

    times = 0
    while times < 100:
        i = 0
        for x in height:
            ggamma[i] = gp * gauss(x, gmu, gsigma)
            bgamma[i] = bp * gauss(x, bmu, bsigma)
            s = ggamma[i] + bgamma[i]
            ggamma[i] /= s
            bgamma[i] /= s
            i += 1

        gn = sum(ggamma)
        gp = float(gn) / float(N)
        bn = sum(bgamma)
        bp = float(bn) / float(N)
        gmu = averageWeight(height, ggamma, gn)
        gsigma = varianceWeight(height, ggamma, gmu, gn)
        bmu = averageWeight(height, bgamma, bn)
        bsigma = varianceWeight(height, bgamma, bmu, bn)

        now = [gp, bp,gmu,gsigma,bmu,bsigma]
        if isSame(cur, now):
            break
        cur = now
        print "Times:\t", times
        print "Girl mean/gsigma:\t", gmu,gsigma
        print "Boy mean/bsigma:\t", bmu,bsigma
        print "Boy/Girl:\t", bn, gn, bn+gn
        print "\n\n"
        times += 1
    return now
```

GMM与图像



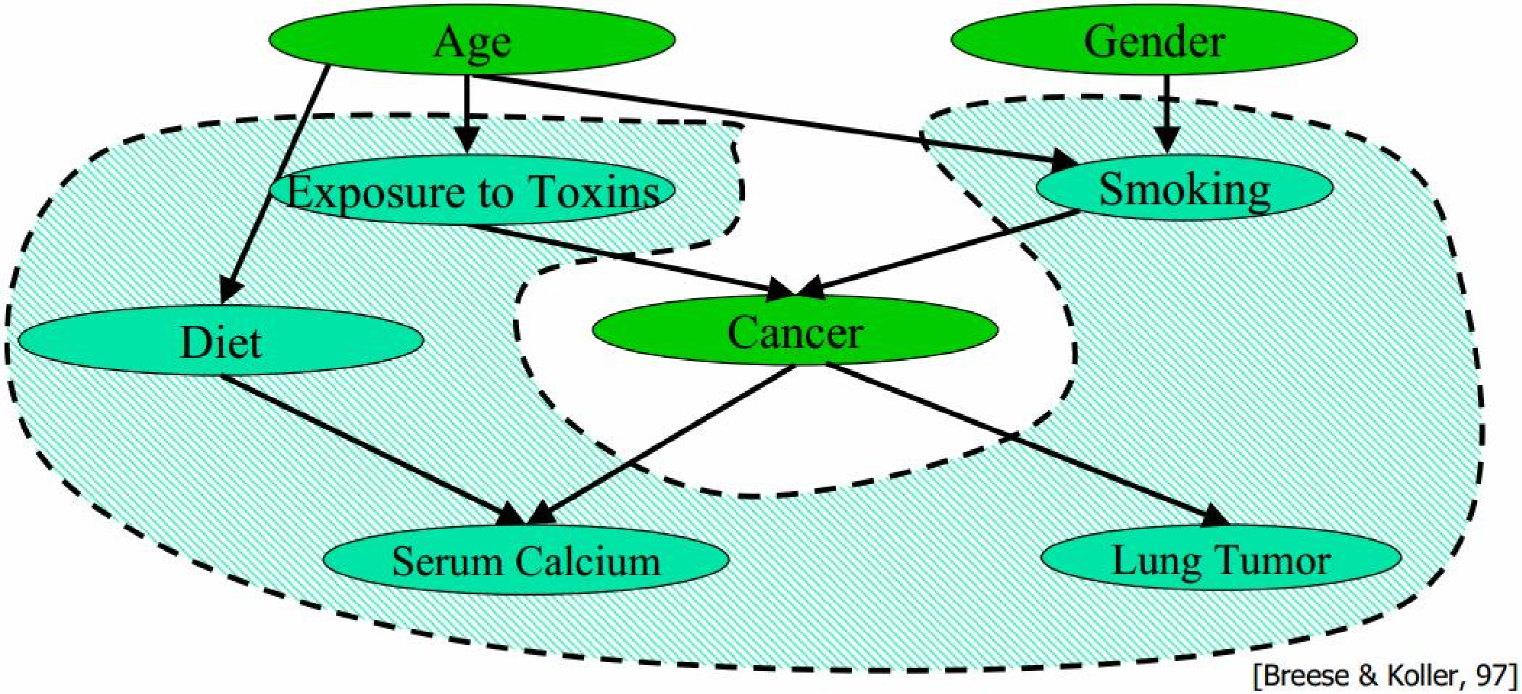



```python
def composite(band, parameter):
    c1 = parameter[0]
    mu1 = parameter[2]
    sigma1 = parameter[3]
    c2 = parameter[1]
    mu2 = parameter[4]
    sigma2 = parameter[5]

    p1 = []
    p2 = []
    for pixel in band:
        p1.append(c1 * gauss(pixel, mu1, sigma1))
        p2.append(c2 * gauss(pixel, mu2, sigma2))

    scale(p1)    #灰度均衡
    scale(p2)
    return [p1, p2]

if __name__ == "__main__":
    im = Image.open('.\\Pic\\test.bmp')
    print im.format, im.size, im.mode

    im = im.split()[0]    #只处理第一个通道
    nb = []               #处理后的新通道
    data = list(im.getdata())
    parameter = GMM(data)
    t = composite(data, parameter)

    im1 = Image.new('L', im.size)
    im1.putdata(t[0])
```

贝叶斯网络

背景知识：Serum Calcium(血清钙浓度)高于2.75mmol/L即为高钙血症。许多恶性肿瘤可并发高钙血症。恶性肿瘤病人离子钙增高的百分比大于总钙，也许可用于肿瘤的过筛试验。当高钙血症的原因难于确定时，必须考虑到恶性肿瘤的存在。

http://www.wiki8.com/xueqinggai_131584/

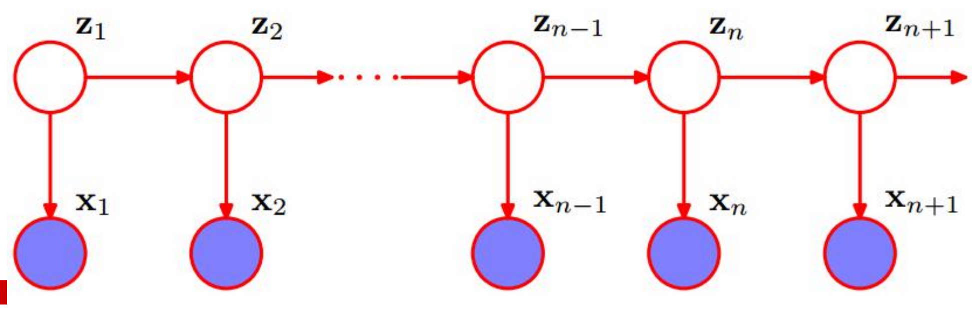


[Breese & Koller, 97]

阴影部分的结点集合，是Cancer的"马尔科夫毯" (Markov Blanket)

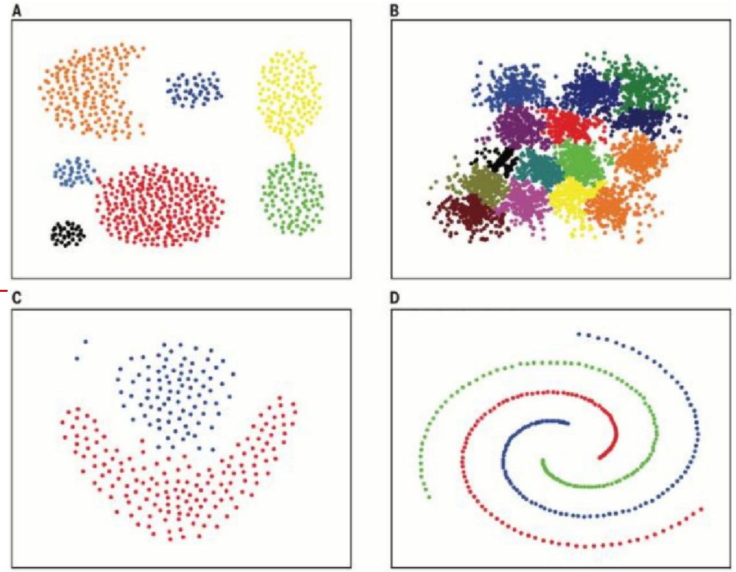
条件独立：$P(S,L|C) = P(S|C) * P(L|C)$

理解HMM框架

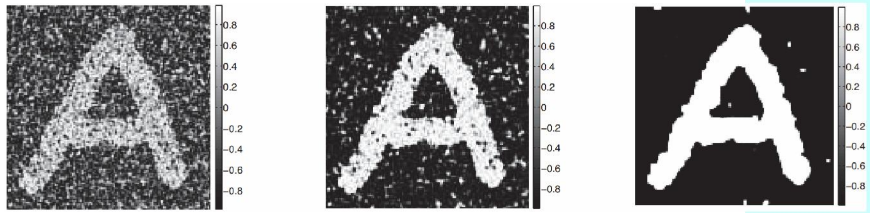


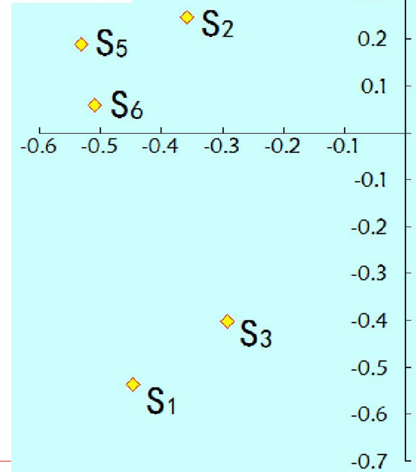
- 概率计算问题
  - 给定模型$\lambda=(A,B,\pi)$和观测序列$O=\{o_1,o_2,\ldots o_T\}$，计算模型λ下观测序列O出现的概率P(O| λ)
  - 前向-后向算法——动态规划
- 学习问题
  - 已知观测序列 $O=\{o_1,o_2,\ldots o_T\}$，估计模型$\lambda=(A,B,\pi)$的参数，使得在该模型下观测序列P(O| λ)最大
  - 极大似然估计(给定状态序列)，Baum-Welch算法(状态序列未知)——EM算法
- 预测问题
  - 即解码问题：已知模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=\{o_1,o_2,\ldots o_T\}$ 求对给定观测序列条件概率P(I|O)最大的状态序列I
  - Viterbi算法——动态规划

其他内容



- 最大熵模型
  - 自然语言处理解决标记问题
- 聚类
  - K-means/K-Mediods/密度聚类/谱聚类
- 降维
  - PCA/SVD/ICA





- SVM
  - 与核技术相结合
- 主题模型pLSA/LDA
  - 与聚类、标签传递算法相结合
- 条件随机场
  - 无向图模型，链式条件随机场解决标记问题
- 变分推导Variation Inference
  - 与EM、贝叶斯相结合，参数、隐变量的学习
- 深度学习
  - 大规模人工神经网络

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

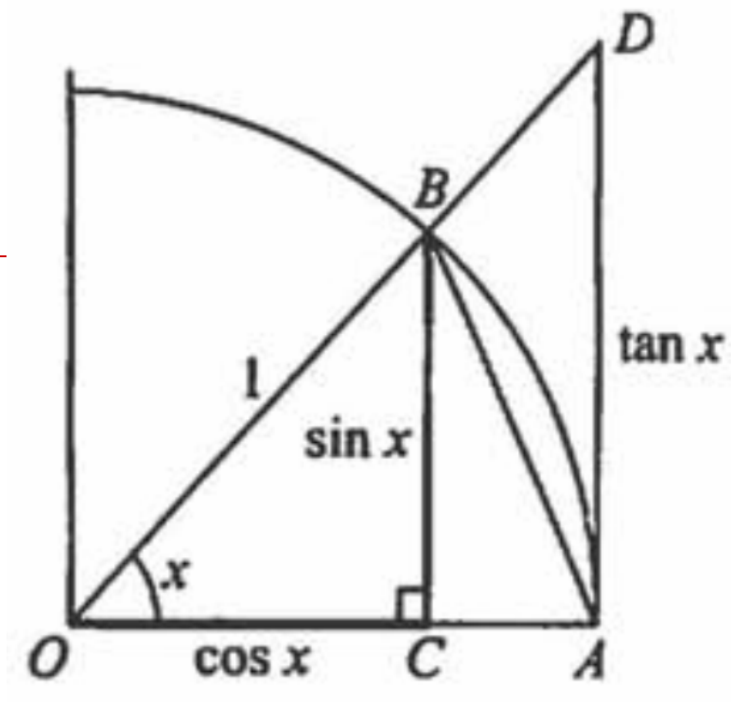
- 当 $x \in U(x_0,r)$ 时，有 $g(x) \leqslant f(x) \leqslant h(x)$ 成立，并且 $\lim\limits_{x \to x_0} g(x) = A$，$\lim\limits_{x \to x_0} h(x) = A$，那么

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

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： $\lim\limits_{x\to 0}\cos x = \cos 0 = 1$
- 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

- 根据前文中$a_n=\left(1+\frac{1}{n}\right)^n$的二项展开式，已经证明数组$\{a_n\}$单增有上界，因此，必有极限。
- 同时：

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

$$\lim_{n\to\infty}\left(1+\frac{1}{n+1}\right)^n=\lim_{n\to\infty}\frac{\left(1+\frac{1}{n+1}\right)^{n+1}}{1+\frac{1}{n+1}}=\frac{\lim\limits_{n\to\infty}\left(1+\frac{1}{n+1}\right)^{n+1}}{\lim\limits_{n\to\infty}\left(1+\frac{1}{n+1}\right)}=\frac{e}{1+0}=e$$

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n+1}=\lim_{n\to\infty}\left(\left(1+\frac{1}{n}\right)^n\left(1+\frac{1}{n}\right)\right)=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n\cdot\lim_{n\to\infty}\left(1+\frac{1}{n}\right)=e\cdot(1+0)=e$$

- 根据两边夹定理，函数$f(x)=\left(1+\frac{1}{x}\right)^x$的极限存在，为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
  - 在信息熵章节中将再次遇到它
- 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1

- 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：

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

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

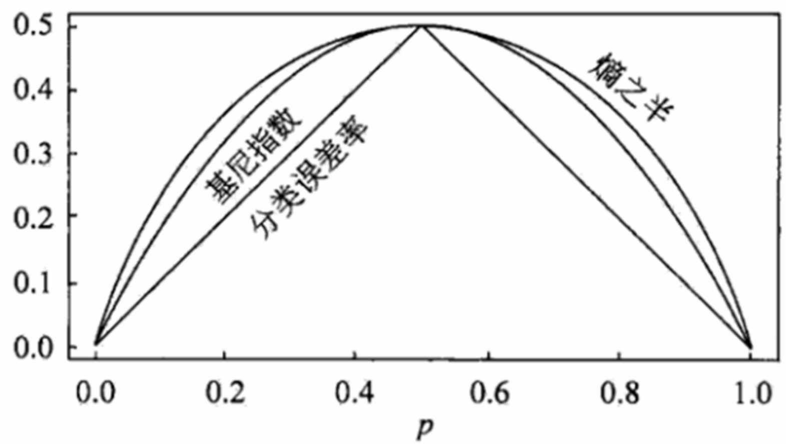
$$= 2^k\cdot e^r$$

Taylor公式的应用2

- 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x=1处一阶展开，忽略高阶无穷小，得到f(x)≈1-x

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

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



  - 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

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

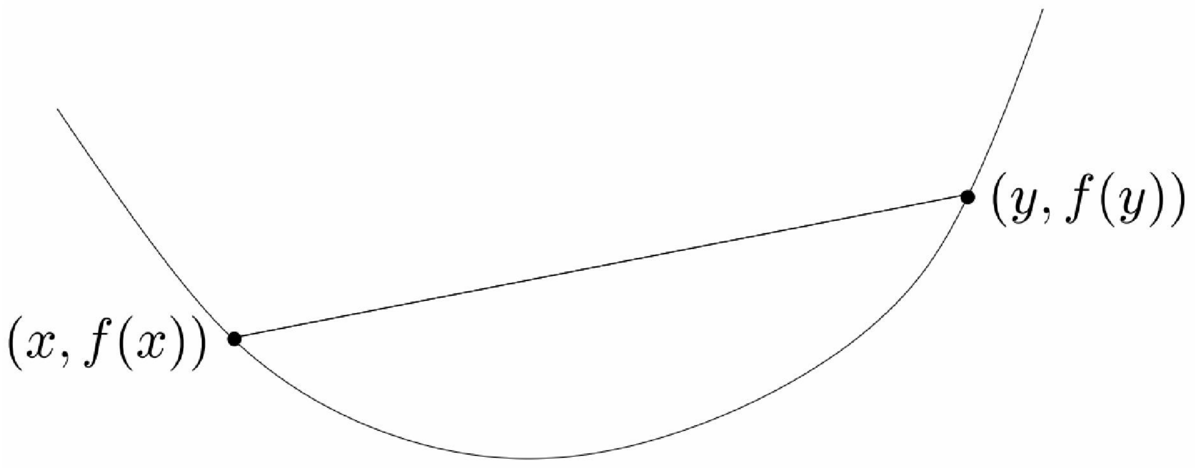
  为函数z=f(x,y)在点P的梯度，记做**grad**f(x,y)
- 梯度的方向是函数在该点变化最快的方向
  - 考虑一座解析式为z=H(x,y)的山，在(x_0,y_0)的梯度是在该点坡度变化最快的方向。
- 梯度下降法
  - 思考：若下山方向和梯度呈θ夹角，下降速度是多少？

凸函数

- 若函数f的定义域domf为凸集，且满足

$$\forall x, y \in \operatorname{dom} f, 0 \le \theta \le 1,\ \text{有}$$

$$f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y)$$



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

- 凸函数的表述

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

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：可以在确定函数的凹凸性之后，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：

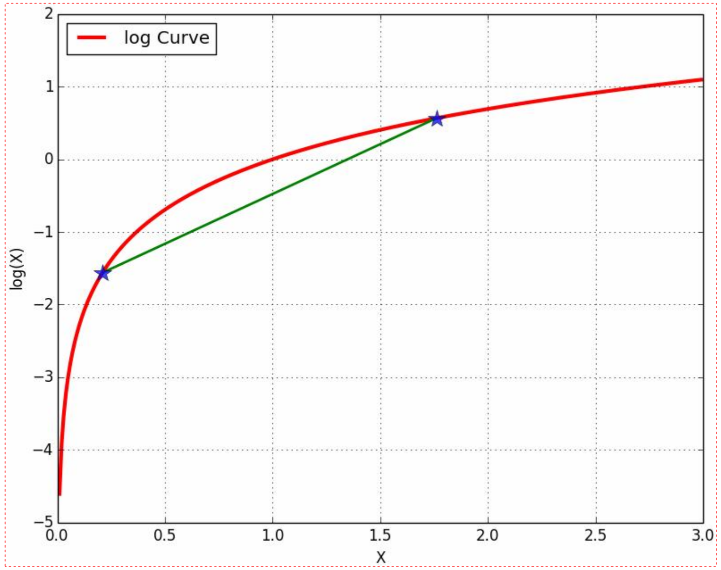
$$D(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\log\frac{p(x)}{q(x)}$$

- 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在定义域上是凸函数

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



概率论

- 对概率的认识：$P(x) \in [0,1]$
  - P=0：事件出现的概率为0→事件不会发生？
  - 若x为离散/连续变量，则P(x=x0)表示x0发生的概率/概率密度
- 累计分布函数：$\Phi(x)=P(x \leqslant x0)$
  - $\Phi(x)$一定为单增函数
  - $\min(\Phi(x))=0$， $\max(\Phi(x))=1$
  - 将值域为[0,1]的某函数y=f(x)看成y事件的累积概率
  - 若y=f(x)可导，则p(x)=f'(x)为某概率密度函数
- P.S.
  - cumulative distribution function,CDF
  - Probability Density Function,pdf

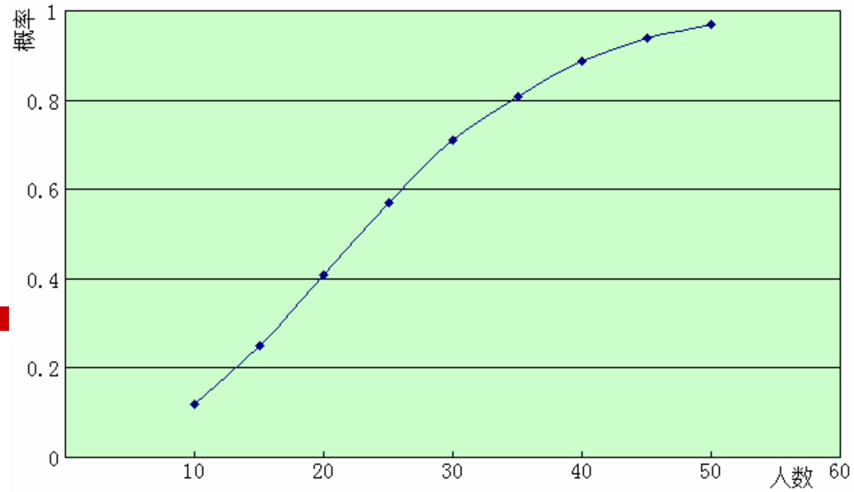
古典概型

- 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 $P(A)=\frac{P_N^n}{N^n}$

- 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：$N(N-1)(N-2)\cdots(N-n+1)=P_N^n$

生日悖论



- 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

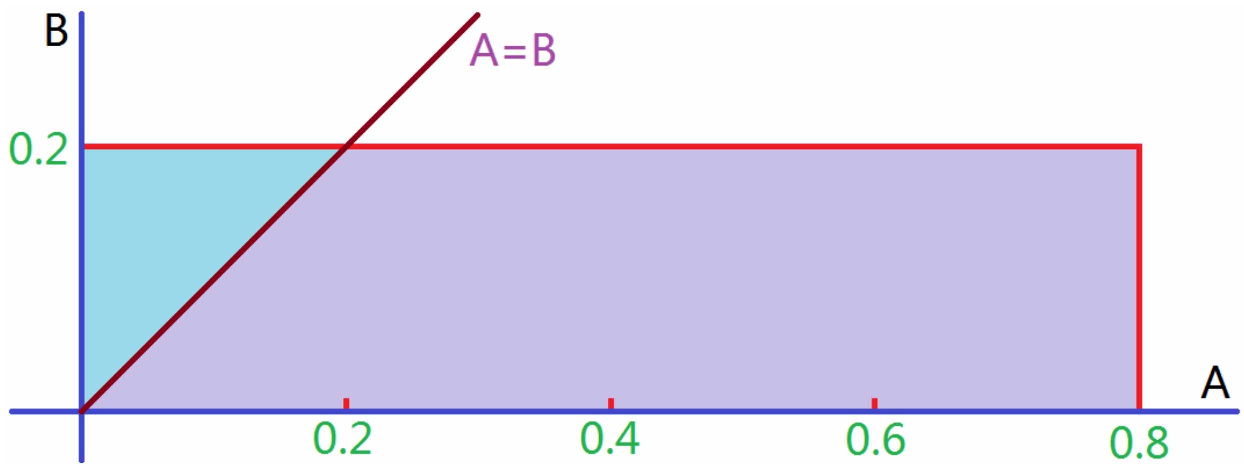
- 把n个物品分成k组，使得每组物品的个数分别为n1,n2…nk，(n=n1+n2+…+nk)，则不同的分组方法有$\frac{n!}{n_1!n_2!n_3!\cdots n_k!}$种。
- 上述问题的简化版本，即n个物品分成2组，第一组m个，第二组n-m个，则分组方法有$\frac{n!}{m!(n-m)!}$，即：C_n^m。

商品推荐

- 商品推荐场景中过于聚焦的商品推荐往往会损害用户的购物体验，在有些场景中，系统会通过一定程度的随机性给用户带来发现的惊喜感。
- 假设在某推荐场景中，经计算A和B两个商品与当前访问用户的匹配度分别为0.8分和0.2分，系统将随机为A生成一个均匀分布于0到0.8的最终得分，为B生成一个均匀分布于0到0.2的最终得分，试计算最终B的分数大于A的分数的概率。

商品推荐

- A=B的直线上方区域，即为B>A的情况。
- $S_{蓝色}=0.02$　　$S_{矩形}=0.16$
- $p=0.02/0.16=0.125$



概率公式

- 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。

贝叶斯公式的应用

- 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。
- 解：

$$P(G=1)=\frac{5}{8} \qquad P(G=0)=\frac{3}{8}$$

$$P(A=1|G=1)=0.8 \qquad P(A=0|G=1)=0.2$$

$$P(A=1|G=0)=0.3 \qquad P(A=0|G=0)=0.7$$

$$P(G=1|A=1)=?$$

$$P(G=1|A=1)=\frac{P(A=1|G=1)P(G=1)}{\sum_{i\in G}P(A=1|G=i)P(G=i)}=\frac{0.8\times\frac{5}{8}}{0.8\times\frac{5}{8}+0.3\times\frac{3}{8}}=0.8163$$

两种认识

- 给定某系统的若干样本，求该系统的参数。
- 矩估计/MLE/MaxEnt/EM等：
  - 假定参数是某个/某些未知的定值，求这些参数如何取值，能够使得某目标函数取极大/极小。
  - 频率学派
- 贝叶斯模型：
  - 假定参数本身是变化的，服从某个分布。求在这个分布约束下使得某目标函数极大/极小。
  - 贝叶斯学派
- 无高低好坏之分，只是认识自然的手段。只是在当前人们掌握的数学工具和需解决的实践问题中，贝叶斯学派的理论体系往往能够比较好的解释目标函数、分析相互关系等。
  - 前面章节的内容，大多是频率学派的思想；下面的推理，使用贝叶斯学派的观点。

贝叶斯公式 $P(A|B)=\frac{P(B|A)P(A)}{P(B)}$

- 给定某系统的若干样本x，计算该系统的参数，即

$$P(\theta|x)=\frac{P(x|\theta)P(\theta)}{P(x)}$$

  - P(θ)：没有数据支持下，θ发生的概率：先验概率。
  - P(θ|x)：在数据x的支持下，θ发生的概率：后验概率。
  - P(x|θ)：给定某参数θ的概率分布：似然函数。
- 例如：
  - 在没有任何信息的前提下，猜测某人姓氏：先猜李王张刘……猜对的概率相对较大：先验概率。
  - 若知道某人来自"牛家村"，则他姓牛的概率很大：后验概率——但不排除他姓郭、杨等情况。

贝叶斯公式带来的思考 $P(A|D)=\frac{P(D|A)P(A)}{P(D)}$

- 给定某些样本D，在这些样本中计算某结论A_1、A_2……A_n出现的概率，即$P(A_i|D)$

$$\max P(A_i|D)=\max\frac{P(D|A_i)P(A_i)}{P(D)}=\max\left(P(D|A_i)P(A_i)\right)\rightarrow\max P(D|A_i)$$

$$\Rightarrow\max P(A_i|D)\rightarrow\max P(D|A_i)$$

- 第一个等式：贝叶斯公式；
- 第二个等式：样本给定，则对任何A_i,$P(D)$是常数；
- 第三个箭头：若这些结论A_1、A_2……A_n的先验概率相等(或近似)，则得到最后一个等式：即第二行的公式。

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，

所以 $$E(X)=\sum_{i=1}^{n} E(X_i)=np.$$

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

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

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

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

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

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

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

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

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

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

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

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

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

指数分布

设随机变量 X 服从指数分布，其概率密度为

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

$$P(x > s + t | x > s) = P(x > t)$$

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

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

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

正态分布

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

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

$$=\mu\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$

$$=\mu.$$

正态分布

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

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

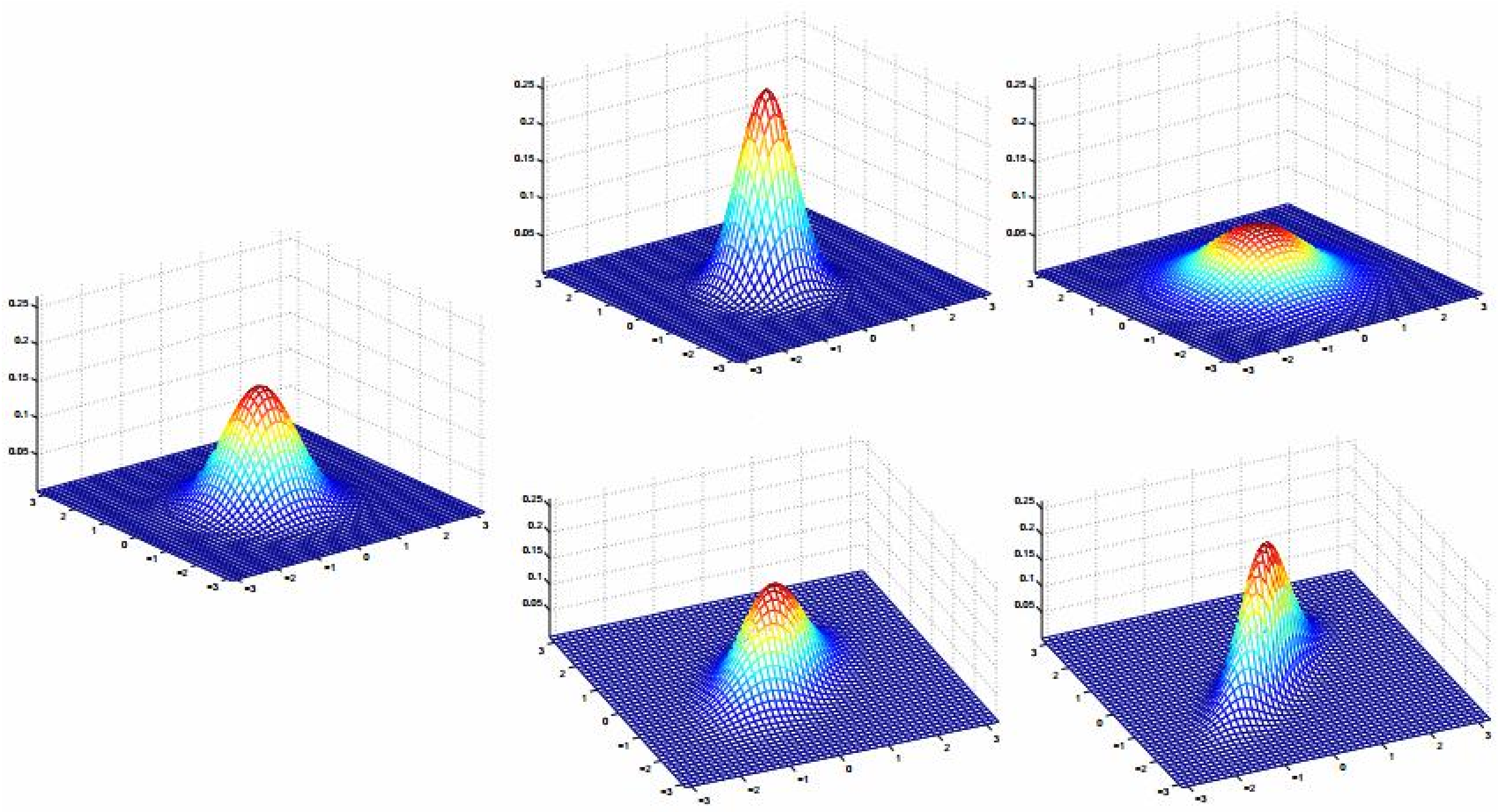
令 $\dfrac{x-\mu}{\sigma}=t$，得

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

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

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

二元正态分布



集合Hash问题

- 某Hash函数将任一字符串非均匀映射到正整数k，概率为2^{-k}，如下所示。现有字符串集合S，其元素经映射后，得到的最大整数为10。试估计S的元素个数。

$$P\{Hash(<string>)=k\}=2^{-k},\quad k\in Z^{+}$$

问题分析 $P\{Hash(<string>)=k\}=2^{-k}, \quad k\in Z^+$

- 由于Hash映射成整数是指数级衰减的，“最大整数为10”这一条件可近似考虑成“整数10曾经出现”，继续近似成“整数10出现过一次”。
- 字符串被映射成10的概率为$p=2^{-10}=1/1024$，从而，一次映射即两点分布：

$$\begin{cases} P(X=1)=\dfrac{1}{1024} \\ P(X=0)=\dfrac{1023}{1024} \end{cases}$$

问题分析

- 从而n个字符串的映射，即二项分布：

$$P\{X=k\}=C_n^k p^k(1-p)^{n-k}，\text{其中}p=\frac{1}{1024}$$

- 二项分布的期望为：$E(P\{X=k\})=np，\text{其中}p=\frac{1}{1024}$

- 而期望表示n次事件发生的次数，当前问题中发生了1次，从而：

$$np=1\Rightarrow n=\frac{1}{p}\Rightarrow n=1024$$

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	$n\ge 1,$ $0<p<1$	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.

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

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

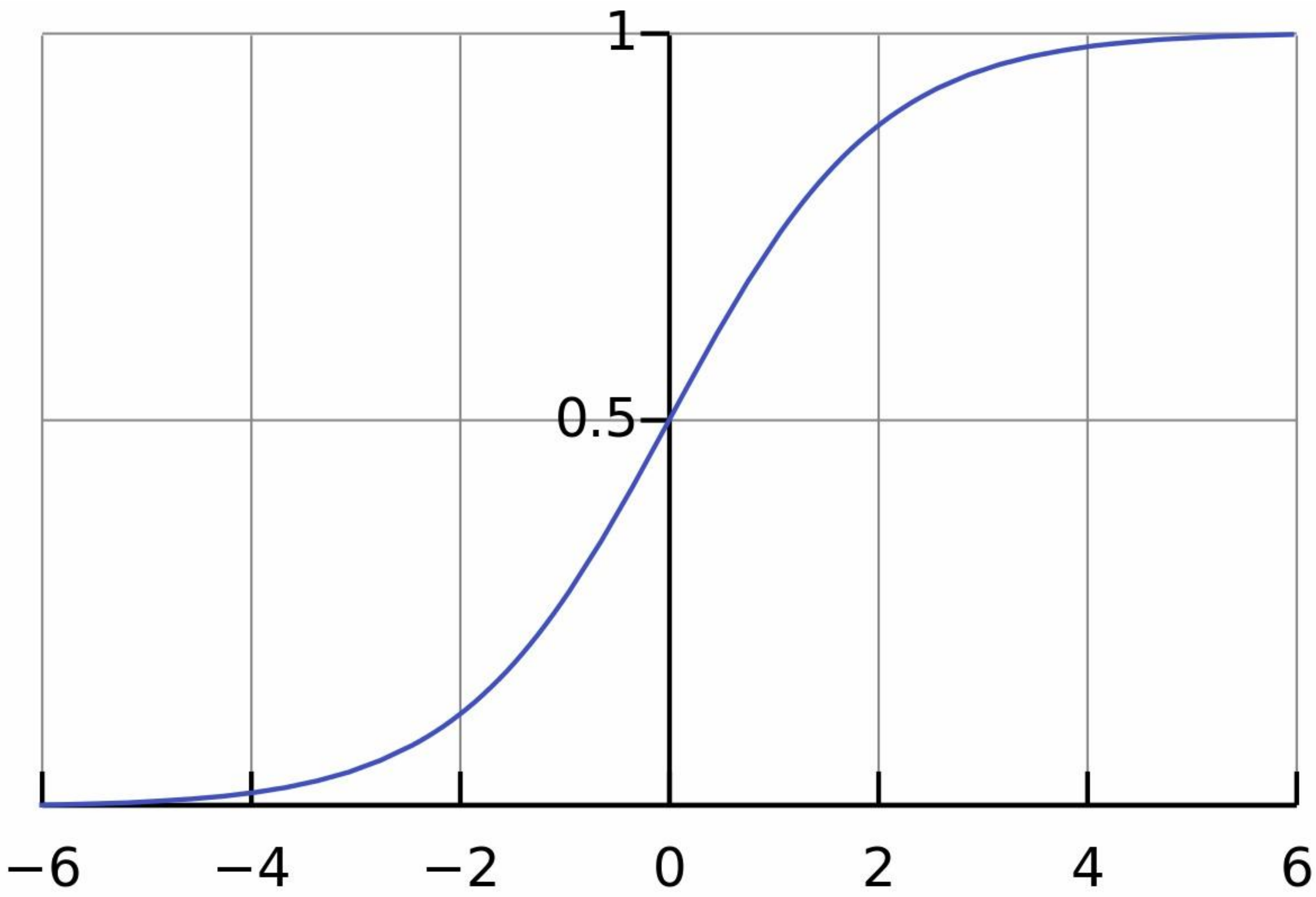
考察参数Φ

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

$$
\begin{aligned}
f'(x) &= \left(\frac{1}{1+e^{-x}}\right)' \\
&= \frac{e^{-x}}{\left(1+e^{-x}\right)^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \left(1-\frac{1}{1+e^{-x}}\right) \\
&= f(x)\cdot\left(1-f(x)\right)
\end{aligned}
$$

- 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:

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

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

课下作业1

- 查阅有关Gamma分布的相关内容：

$$p(x;\alpha,\beta)=\frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}\ ,\quad x\geq 0(\text{常系数}\alpha,\beta>0)$$

- Gamma函数：

$$\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}e^{-t}dt \qquad \Gamma(n)=(n-1)!$$

- Gamma分布的期望为：

$$E(X)=\frac{\alpha}{\beta}$$

- 主题模型章节中将有所涉及

课下作业2

- A、B两国元首相约在首都机场晚20点至24点交换一份重要文件。如果A国的飞机先到，A会等待1个小时；如果B国的飞机先到了，B会等待2个小时。假设两架飞机在20点至24点降落机场的概率是均匀分布，试计算能够在20点至24点完成交换的概率。
  - 假设交换文件本身不需要时间。

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

我们在这里

- 7 七月算法 http://www.julyedu.com/
  - 精品视频/直播课程/问答社区
- 七月题库APP
  - http://www.julyapp.com/
- 微博
  - @研究者July
  - @七月题库
  - @邹博_机器学习
- 微信公众号
  - julyedu

感谢大家！

恳请大家批评指正！