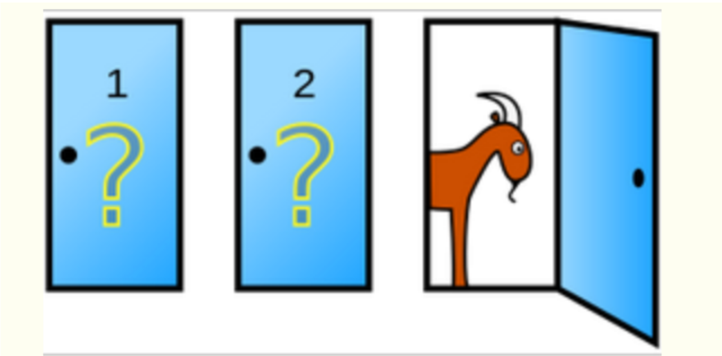
开课吧
kaikeba.com
慧科集团旗下企业

概率论

机器学习中数学基础第二课

- 蒙特霍尔问题：
    - 参赛者面前有三扇关闭着的门，其中一扇的后面是一辆汽车，选中后面有车的那扇门就可以赢得该汽车，而另外两扇门后面则各藏有一只山羊。当参赛者选定了一扇门，但未去开启它的时候，主持人会开启剩下两扇门中的一扇，露出其中一只山羊。主持人其后会问参赛者要不要更换选择，选另一扇仍然关着的门。



- 统计:
  - 统而计之

  - 统：观察数据
  - 计：分析数据

  - 是搜集和分析数据的科学与艺术

大纲

- 1. 几个概念
- 2. 概率
- 3. 分布
- 4. 估计

几个概念

- 均值: 用于度量样本平均水平的变量。

- 给出一组薪资数据：
  - 8K,10K,15K,20K,25K,30K,32K

  - 数据挖掘人员的薪资水平怎么样?

  - 均值:$X_{mean} = \sum_{i=1}^{n} X_i / n$

几个概念

- 如果马云来了，一个月收入是10000K

- 数据变为：
  - 8K,10K,15K,20K,25K,30K,32K,10000K

  - 平均值为：1267.5K

  - 中位数：一组数按照升序排列，排序位于中间的数，为中位数，如果中间数为偶数，则为中间两个数的平均

几个概念

- 马云、马化腾，雷军，扎克伯格，王健林，都来了，他们按照一个月收入分别是9000K,10000K,10000K,10000K,11000K

- 数据变为：
  - 8K,10K,20K,20K,20K,30K,32K,9000K,10000K,10000K,10000K,11000K

  - 均值和中位数都有误导

  - 众数：数据中出现频数最大的数值。上述双峰数据给出两个众数，20K，10000K。
  - 众数可以用于分类。

几个概念

- 均值: 数据对称时或者趋势单一
- 中位数：由于存在异常值发生偏移时
- 众数：数据可以分为多个组时

几个概念

- 两个球员得分情况：
- A: 7， 9， 9， 10， 10， 10， 10， 11， 11， 13
- B: 2， 4， 6， 7， 7， 10， 10， 10， 11， 13， 30

- 均值、众数、中位数都一样。
- ???

几个概念

- 方差，标准偏差

- 方差：用于度量数据分散性的一种方法：
- 计算公式为：
  - 方差：$\delta^2 = \sum(x-\mu)^2/(n\text{-}1)$
  - 标准差：$\delta = \sqrt{\sum(x-\mu)^2/(n-1)}$

  - A,B 标准偏差为：1.48和7.02 （稳定球员和神经刀球员）

几个概念

- 假设有两位球员：
  - 第一位均值得分为70分，标准偏差为20分
  - 第二位均值得分为40分，标准偏差为10分
  - 现在球员1得分为75分，球员2得分为55分，就球员本身相对于其历史记录来说，哪位表现更优异呢?
  - 由此我们引出了机器学习中常用的特征处理的技巧: 标准化和归一化

几个概念

- 协方差：

- $\delta^2 = \sum(x-\mu_x)\ (y-\mu_y)\ /(n\text{-}1)$

- 协方差用于衡量两个变量的总体误差。
- 通俗的理解：两个变量变化一致性程度的度量。

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

几个概念

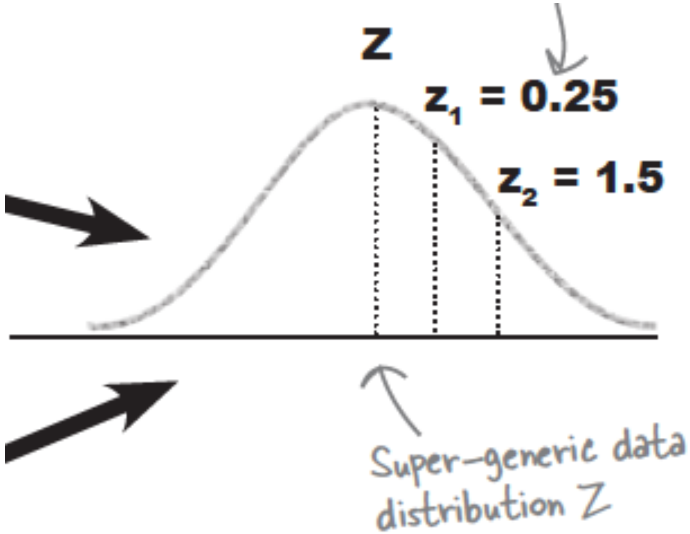
- 相关系数

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

- 相关系数的取值范围在＋1到－1之间

几个概念

- 标准化和归一化:
  - 最大最小值归一化：
    - （x-最小值）/（最大值-最小值）
  - z-score标准化
    - （x-均值）/标准差



概率

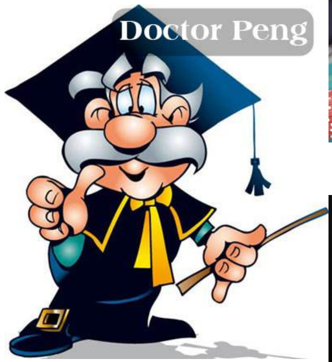
- 概率：表示某件事发生的可能性大小的一个量。

- 黑盒子里面有20个黑球和15个白球，随机从盒子里面取出一个球来，让你猜球是什么颜色，猜对了得10元钱，请问你会怎么猜？

- Probability 概率 用P(x)表示x发生的概率

有两个赌徒，他们拥有相同的获胜概率，他俩下赌金之后，约定谁先赢满5局，谁就获得全部赌金。赌了半天，A赢了4局，B赢了3局，时间很晚了，他们都不想再赌下去了。那么，这个钱应该怎么分？是不是把钱分成7份，赢了4局的就拿4份，赢了3局的就拿3份呢？或者，因为最早说的是满5局，而谁也没达到，所以就一人分一半呢？

这两种分法都不对。正确的答案是：赢了4局的拿这个钱的3／4，赢了3局的拿这个钱的1／4。



$$E(X)=\sum_{i=1}^{n}x_i\cdot p(x_i)$$

- 一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。随机选取一个学生，他（她）穿长裤的概率和穿裙子的概率是多大。

- P1=0.6+0.4*0.5=0.8

- P2=1-P1=0.2

- 你在校园里面随机遇到一个穿长裤的人，他/她是男生的概率是多少？

概率

- 贝叶斯定理
  - 是概率论中比较难掌握的一部分之一，也是机器学习中运用最广泛的一种，贝叶斯决策理论的思想是很多分类算法的核心。

概率

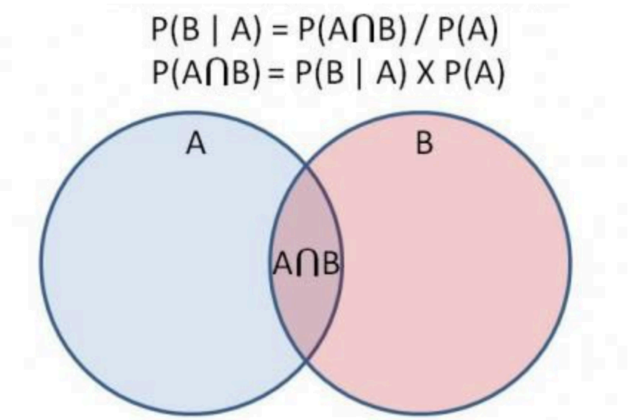
贝叶斯(约1701-1761) Thomas Bayes，英国数学家。约1701年出生于伦敦，做过神甫。1742年成为英国皇家学会会员。1761年4月7日逝世。贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。他死后，理查德·普莱斯(Richard Price)于1763年将他的著作《机会问题的解法》(An essay towards solving a problem in the doctrine of chances)寄给了英国皇家学会，对于现代概率论和数理统计产生了重要的影响。1774年，法国数学家皮埃尔-西蒙·拉普拉斯才给出了我们现在所用的贝叶斯公式的表达。

贝叶斯公式

设 $A_1, A_2, \cdots, A_n$ 为一个完备事件组，

$P(A_i) > 0$, $i = 1, \cdots, n$, 对任一事件 B, 若 $P(B) > 0$, 有

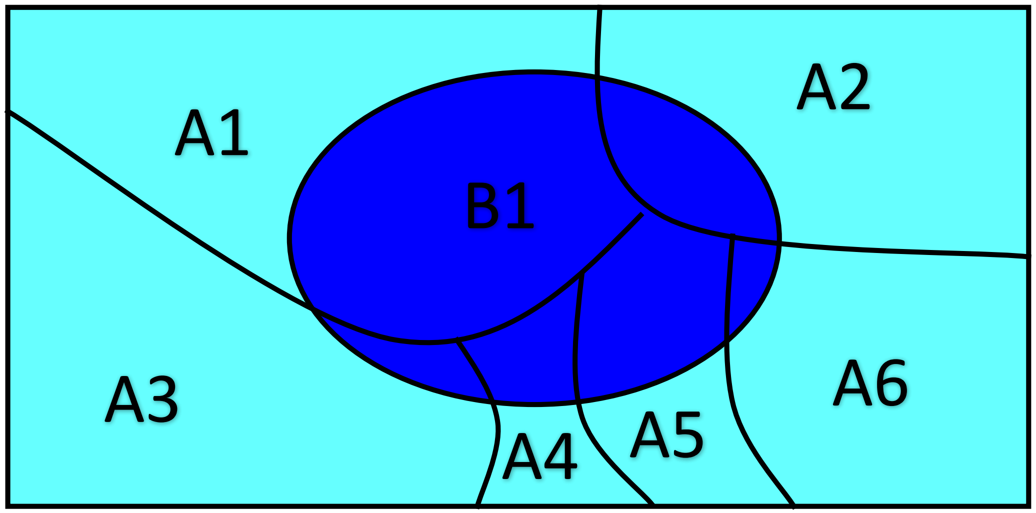
$$P(A_k \mid B) = \frac{P(A_k B)}{P(B)}$$



$$= \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}, \quad (k = 1, 2, \cdots, n)$$

全概率公式

$$P(B)=\sum_{i=1}^{n}P(A_i)\,P(B\mid A_i)$$



贝叶斯概率

贝叶斯公式

$$P(A_k \mid B) = \frac{P(A_k B)}{P(B)} = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)},$$

$\mathbf{P}(\boldsymbol{A_k}|\mathbf{B})$ $\boldsymbol{A_k}$的后验概率

$\mathbf{P}(\boldsymbol{A_k})$先验概率

$\mathbf{P}(\boldsymbol{B}|\boldsymbol{A_k})$似然函数

某游戏公司测试新游戏，调查了玩家对游戏的满意程度，结果如下：80%玩家选择游戏1，20%玩家选择游戏2；游戏1玩家中，60%人觉得好玩；游戏2玩家中，70%觉得好玩。随机挑选了一位玩家，他表示玩的这个游戏很好玩，问题来了：请问：他玩这个游戏是游戏2的概率多大。

P(游戏2| 满意)=P(游戏2)*P(满意|游戏2)/P(满意)

P(满意)=P(游戏1)*P(满意|游戏1)+P(游戏2)*P(满意|游戏2)

P(游戏2| 满意)=0.2*0.7/(0.2*0.7+0.6*0.8)=0.23

概率

全概率公式

贝叶斯定理

概率

独立性假设

贝叶斯公式的重要性假设：朴素的意思

事件A发生的概率不受事件B的影响，成为事件A和事件B是独立的。

独立事件的概率计算为：P(AB)=P(A)*P(B)

分布

分布其实一种统计出来的频数图（直方图）。
比如下面一组数据：

1，2，2，3，3，3，4，4，5

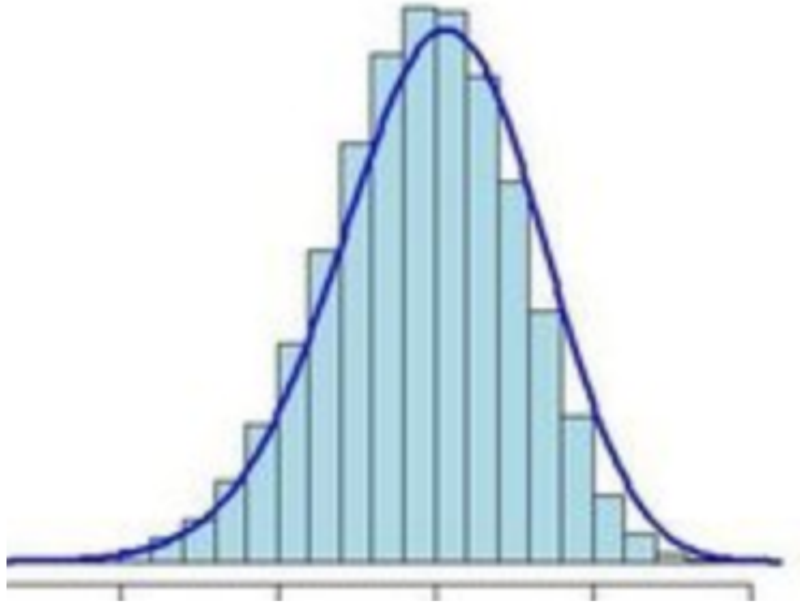
如何看这组数据的分布呢？

分布

离散分布和连续分布：

有限个-离散

无限个-连续

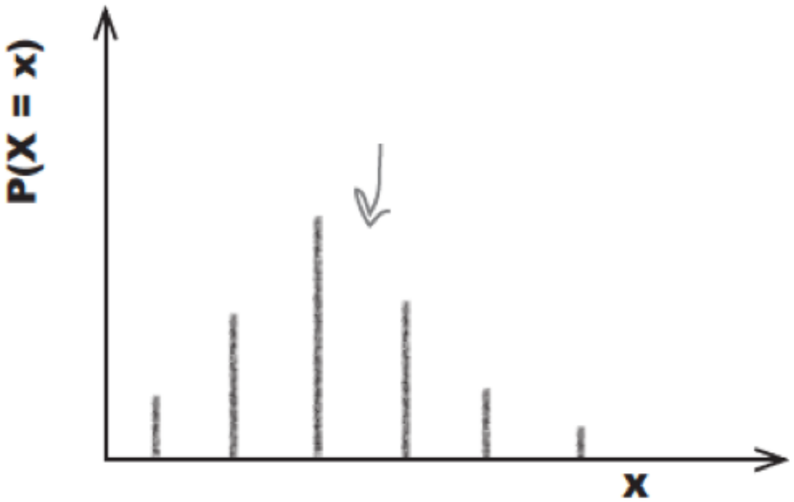


分布

- 二项分布

- 问题: 抛硬币，抛5次硬币，有2次正面朝上的概率是多少

$$p(x) = C_n^x p^x (1-p)^{n-x}$$



分布

- 二项分布应用场景

- 1. 多次试验是独立的
- 2. 每次试验概率相同
- 3. 试验结果为二分类

分布

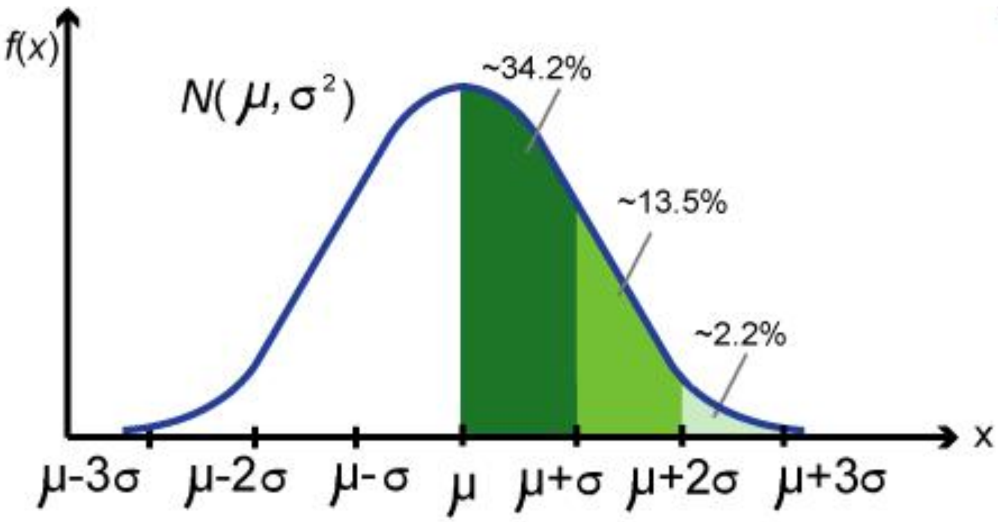
- 二项分布
- E(x)=np
- Var(x)=npq

- 退化为单次的话(1重伯努利试验)，则为伯努利分布-（很多机器学习的基础分布）：

$$p(x) = p^x * (1-p)^{1-x}$$

分布

- 正态分布



$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

均值为μ:

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$$

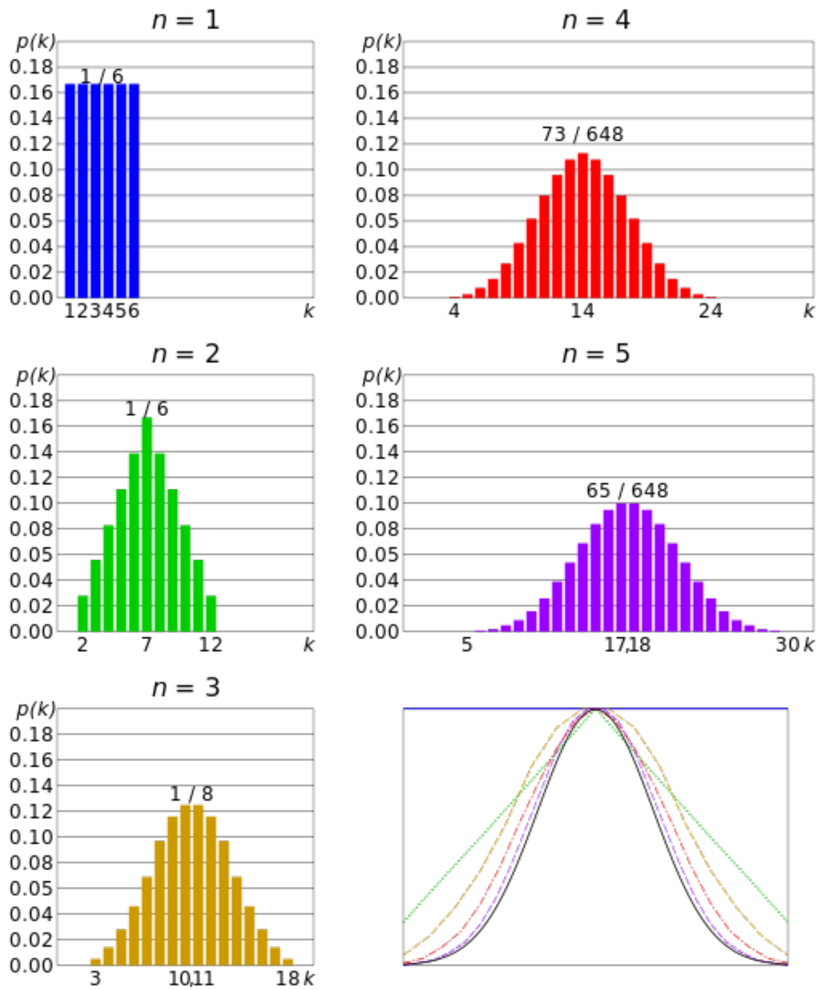
方差为σ:

$$s^2=\frac{\sum_{i=1}^{n}(x-u)^2}{n-1}$$

正态分布也叫高斯分布，其曲线两头低，中间高，也称作钟形曲线

- 中心极限定理

- 从总体中抽取样本容量为n的简单随机样本,当样本容量很大时,样本均值的抽样分布与正态概率分布近似。



分布

- 正态分布

- E(X)=u

- Var(X)= σ^2　　　#σ为标准差

当$\mu = 0, \sigma = 1$时的正态分布是标准正态分布

似然

- 似然估计
  -
  - 现在有一个正反面不是很匀称的硬币，如果正面朝上记为H，反面朝上记为T，抛10次的结果如下:
  - T,T,T,H,T,T,H,T,T,T
  - 问正面朝上的概率是多少？

似然

- 似然估计
  -
  - 设反面朝上的概率为u, 则正面朝上的概率为1-u；
  - T,T,T,H,T,T,H,T,T,T
  - 那么出现上述可能性的概率是多大？
  - 为：u*u*u*(1-u)*u*u*(1-u)*u*u*u

似然

- 似然估计
  - 为：u*u*u*(1-u)*u*u*(1-u)*u*u*u
  - 更一般的形式，我们假设正面朝上的x=1,反面朝上x=0
  - 一次的概率为: $u^{x}(1-u)^{(1-x)}$

$$p(\mathbf{X};\mu)=\prod_{i=1}^{n}p(x_i;\mu)=\prod_{i=1}^{n}\mu^{x_i}(1-\mu)^{1-x_i}$$

似然

- 似然估计

$$
\begin{aligned}
\log p(\mathbf{X};\mu) &= \log \prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i} \\
&= \sum_{i=1}^{n} \log\left\{\mu^{x_i}(1-\mu)^{1-x_i}\right\} \\
&= \sum_{i=1}^{n} \left[\log \mu^{x_i} + \log(1-\mu)^{1-x_i}\right] \\
&= \sum_{i=1}^{n} \left[x_i \log \mu + (1-x_i)\log(1-\mu)\right]
\end{aligned}
$$

似然

- 似然估计

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \log p(\mathbf{X}; \mu) &= \sum_{i=1}^{n} \frac{\partial}{\partial \mu} [x_i \log \mu + (1 - x_i) \log(1 - \mu)] \\
&= \sum_{i=1}^{n} x_i \frac{\partial}{\partial \mu} \log \mu + \sum_{i=1}^{n} (1 - x_i) \frac{\partial}{\partial \mu} \log(1 - \mu) \\
&= \frac{1}{\mu} \sum_{i=1}^{n} x_i - \frac{1}{1 - \mu} \sum_{i=1}^{n} (1 - x_i)
\end{aligned}
$$

似然

- 似然估计

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 发现是正面朝上的概率是0.2， 我们实验了10次，有两次是正面。

- 步骤

- 1. 写出似然函数；

- 2.对似然函数取对数并求导数；

- 3. 求解模型中参数的最优值。

- 蒙特霍尔问题：
  - 参赛者面前有三扇关闭着的门，其中一扇的后面是一辆汽车，选中后面有车的那扇门就可以赢得该汽车，而另外两扇门后面则各藏有一只山羊。当参赛者选定了一扇门，但未去开启它的时候，主持人会开启剩下两扇门中的一扇，露出其中一只山羊。主持人其后会问参赛者要不要更换选择，选另一扇仍然关着的门。