白话大数据与机器学习

# 大数据产业

简单可以分为：数据收集，数据存储，数据建模，数据分析，数据变现。

# 步入数据之门

信息是用来消除随机不确定性的东西。

# 排列组合与古典概率

# 统计与分布

## 4.1 加和值、平均值和标准差

### 4.1.1 加和值

### 4.1.2 平均值

### 4.1.3标准差

标准差公式为：

## 4.2 加权均值

## 4.3 众数，中位数

## 4.4 欧式距离



表现为：两点之间，直线最短

## 4.5 曼哈顿距离

曼哈顿距离（Manhattan Distance），又叫出租车距离。是一种典型的对欧氏距离的简化。



## 4.6 同比和环比

同比：与相邻时段的同一时期相比。例如：每月一次的报告，今年7月与去年7月相比。

环比：直接和上一个报告时期相比。例如：今年7月的报告与6月的相比。

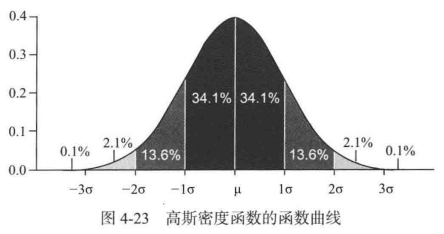
## 4.7 抽样

抽样（sampling）。

## 4.8 高斯分布

正态分布（Normal Distribution）又名高斯分布（Gaussian Distribution）。高斯函数的概率密度函数如下：





密度函数：，是样本特征自变量，是在这个样本特性上的数量比例。这个函数的峰值在的位置，对应的函数值为。样本数量在此用的是定积分的定义，即整个函数曲线在其下方围住的面积占比。在左右两侧的函数对称，在和之间的样本数量占整个样本数量的68.2%，和占有95.4%，和之间占有99.6%。

根据统计的情况在不同条件下和值可能不同。

1. 比较大，则整个函数图像的中轴向右挪动比较多；
2. 比较小，则函数图像的中轴向左挪动比较多；
3. 比较大，则整个曲线绵延比比较长，整个坡度显得平缓；
4. 比较小，整个曲线窄而立陡；

**4.9 泊松分布**

泊松（Poisson）分布是一种统计和概率中常见的离散概率分布。其概率函数如下：



泊松分布的参数是单位时间（或单位面积）内随机时间的平均发生率。泊松分布适合于描述单位时间内随机事件发生的次数。

