1 机器学习中的统计学基础

1.1 机器学习与统计学习

统计强调推理，而机器学习则强调预测，他们一个很大的区别在于目的不同。

机器学习与统计学习都是支撑数据建模， 不同的是机器学习是数据建模的计算机视角，侧重技能； 统计学习是数据建模的数学视角，侧重推断。

1.2 部分机器学习算法中的统计学

算法	数理统计理论
贝叶斯分类器	随机变量，贝叶斯公式，随机变量独立性，正态分布，最大似然估计
贝叶斯网络	条件概率，贝叶斯公式
决策数	概率，熵
主成分分析	协方差矩阵，特征值与特征向量
logistic	概率，随机变量，最大似然估计
随机森林	抽样，方差
隐马尔可夫链	概率，离散型随机变量，条件概率，随机变量独立性，虽大似然估计
条件随机场	条件概率，数学期望，最大似然估计
高斯混合模型	正态分布，最大似然估计
生成对抗神经网络	条件分布

1.3 一些统计学概念

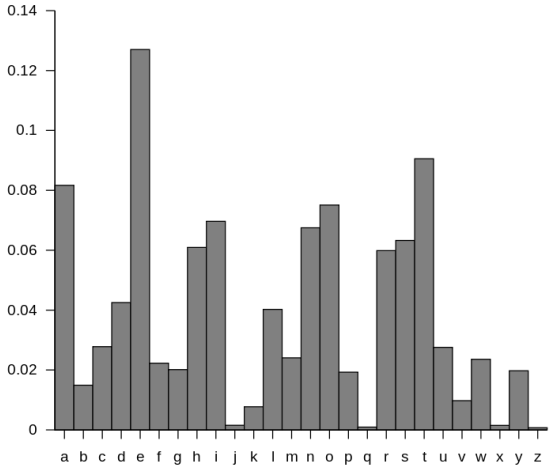
1.3.1 概率与条件概率

1.3.1.1 概率

现实世界中有一些概念是确定的，如石头下落，树木生长；同时一些是不确定的，如抛掷硬币，打靶射击。

不确定说明具有随机性，如果一个试验可以重复但每次结果不止一个，事先知道所有可能（样本空间）但不能确定每次出现什么结果（样本点），这就叫随机试验；随机试验中我们关心的子集就是随机事件。

抛掷一枚硬币100次，正面朝上与反面朝上分别60次（频数）和40次，频率分别是0.6和0.4，当重复试验次数增加，频率会变得稳定，下图表示英文单词中每个字母出现的频率。



但现实中我们不可能每个事件都做大量的试验，所以使用概率来表征时间可能性。后面会聊到大数定律，就是当数量足够大后，频率趋近与概率。

如果集合函数 $P(\cdot)$ 满足下列条件，P(A)就是事件A的概率：

1. 非负性：对于每一个事件 A，有 $P(A) \geq 0$;
2. 规范性：对于必然事件 S，有 $P(S) = 1$;
3. 可列可加性：设 A_1, A_{2}, ……是两两互不相容的事件，及对$A_iA_j = \Phi, i \neq j, i, j = 1,2,\ldots\ldots$，有

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

例：在1-2000的整数中随机取一个数，取到的整数既不能被6整除又不能被8整除的概率是多少：

$$P(\bar{A}\bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$$

$$= 1 - [P(A) + P(B) - P(AB)]$$

$$p = 1 - (\frac{333}{2000} + \frac{250}{2000} - \frac{83}{2000}) = \frac{3}{4}$$

1.3.1.2 条件概率

条件概率是在事件A发生了的基础上B发生的概率，且 $P(A) > 0$，则

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

为事件A发生的条件下事件B发生的条件概率。

同概率的定义，条件概率同样满足3个条件。

1. 非负性
2. 规范性
3. 可列可加性

所以对于任意事件 B_1，B_2，有

$$P(B_1 \cup B_2 \mid A) = P(B_1 \mid A) + P(B_2 \mid A) - P(B_1B_2 \mid A)$$

1.3.2 随机变量与分布

1.3.2.1 随机变量

随机试验的结果可以用数表示，每个样本空间中的元素都是数，但有时候样本空间S的元素不是一个数，此时难以研究，所以将S中的每个元素与实数空间联对应起来，于是有了随机变量。

同样抛掷硬币为例，一枚硬币投掷3次，样本空间为（假设正面为H，反面为T）：

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

若以X记三次投掷得倒正面H的总数，则S中的每一个样本点，在X中都有一个数与之对应，于是得到一个定义域为S，值域为$\{0, 1, 2, 3\}$的函数：

$$X = X(e) = \begin{cases} 3 & e = HHH, \\ 2 & e = HHT, HTH, THH, \\ 1 & e = HTT, THT, TTH, \\ 0 & e = TTT. \end{cases}$$

以上记$\{X = 2\}$，对应样本点集合 $A = \{HHT, HTH, THH\}$ ，这一事件当且仅当事件A发生时有$\{X = 2\}$。类似有：

$$P\{X \leq 1\} = P(HTT, THT, TTH, TTT) = \frac{1}{2}$$

1.3.2.2 离散型随机变量分布

离散型随机分布

离散型随机变量定义：全部可能去到的值是有限个或可列无限多个。 设全部可能取到的值为$x_k(k = 1, 2, 3, \ldots)$，则事件$\{X = x_k\}$的概率为：

$$P\{X = x_k\} = p_k, k = 1, 2, 3, \ldots$$

以上也叫随机变量X的分布律

(0-1)*分布*

设随机变量X只可能是取0与1两个值，它的分布律是

$$P\{X = k\} = p^k(1 - p)^{1-k}, k = 0, 1(0 < p < 1)$$

则称X服从p为参数的(0-1)分布或两点分布 其随机变量函数为：

$$X = X(e) = \begin{cases} 0 & e = e_1, \\ 1 & e = e_2. \end{cases}$$

实际中的0-1分布如：新生儿性别，产品质量是否合格，PM2.5是否超标等。

二项分布（伯努利试验）

设试验E只有两种可能结果：A 及 $\bar{A}$，则称E为伯努利试验。设$P(A) = p(0 < p < 1)$，此时$P(\bar{A}) = 1 - p$，将E独立重复（指每次$P(A) = p$保持不变）地进行n次，则这一串重复的独立试验为 n重伯努利试验。

n重伯努利试验某个A事件发生k次的概率可以记为：

$$p \cdot p \cdot \underbrace{\ldots}_{k\text{个}} \cdot p \cdot (1-p) \cdot (1-p) \cdot \underbrace{\ldots}_{n-k\text{个}} \cdot (1-p) = p^k(1-p)^{n-k}$$

这种指定的方式共有$\binom{n}{k}$种，它们两两互不容，故在n次实验中A发生k次的概率为:

$$P\{X=k\} = \binom{n}{k} p^k q^{n-k},\ k=0,1,2,\ldots,n.$$

$P\{X=k\} = \binom{n}{k} p^k q^{n-k}$正好是$(p+q)^n$的展开式中出现$p^k$的哪一项，故称变量X服从n, p的二项分布。

特别地：当n = 1时二项分布就是为

$$P\{X=k\} = p^k q^{1-k}$$

实际中二项分布的例子：单次射击命中率为0.02，独立射击400次

求至少命中两次的概率。

$$P\{X=k\} = \binom{400}{k} (0.02)^k (0.98)^{400-k},\ k=0,1,2,\ldots,400$$

于是所求概率为：

$$P\{X \geq 2\} = 1 - P\{X=0\} - P\{X=1\}$$

柏松分布

设随机变量X所有可能取的值为0,1,2,...，取各个值的概率为

$$PX=k = \frac{\lambda^k e^{-\lambda}}{k!},\ k=0,1,2,\ldots$$

其中λ是常数，则称X服从参数为λ的柏松分布。

实际中服从柏松分布的例子如：一本书一页中的印刷错误数，某地区一天内邮递遗失的信件数，某一医院一天内的急症病人数，某地区一个事件间隔内发生的交通事故次数。

柏松定理，设λ是一个常数，n是任意正整数，设$np_n = \lambda$，则对于任一个固定的非负整数k，有

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

直观理解一下 柏松分布就是当n很大，但p很小时候的二项分布逼近

1.3.2.3 连续型随机变量分布

连续型随机分布

区别与离散型随机变量，连续型随机变量的样本空间值是不能被一一列举的，比如误差，元件寿命等。对于这样的随机变量，我们不太关注具体数值出现的概率，而是落在某一区间内的概率。于是有：

$$P\{x_1 < X \leq x_2\} = P\{X \leq x_2\} - P\{X \leq x_1\}$$

记：

$$F(x) = P\{X \le x\},\ -\infty < x < \infty$$

为X的分布函数

对于分布函数F(x)，存在非负函数f(x)，使得任意实数x有：

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

则X为连续型随机变量，其中函数f(x)为X的概率密度函数（概率密度）。

均匀分布

均匀分布的概率密度函数：

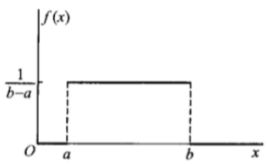
$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b, \\ 0 & \text{其 他}, \end{cases}$$

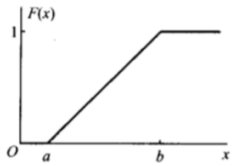
该密度函数表示随机变量X落在区间(a,b)中任意长度的子区间内的可能性是相同的，概率只依赖于子区间长度，而与子区间未知无关。

通过对均匀分布密度函数求积分可得分布函数为：

$$F(x) = \begin{cases} 0 & x < a, \\ \dfrac{x-a}{b-a} & a \le x < b, \\ 1 & x \ge b, \end{cases}$$

均匀分布密度函数与分布函数分别如下





指数分布

指数分布的密度函数为（$\theta > 0$ 为常数）：

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

通过对指数分布密度函数求积分可得分布函数为：

$$f(x) = \begin{cases} 1 - e^{-x/\theta} & x > 0, \\ 0 & \text{其 他}, \end{cases}$$

指数分布有一个重要的性质是无记忆性，比如一个电子元件已经使用了s小时，它总共能使用至少s+t小时的条件概率，与从开始使用时候算起至少使用t小时的概率相等，即对已经使用的s小时没有记忆。

指数分布无记忆性公式：

$$P\{X > s + t \mid X > s\} = P\{X > t\}$$

正态分布

正态分布的密度函数为（$\mu, \sigma(\sigma > 0)$为常数）：

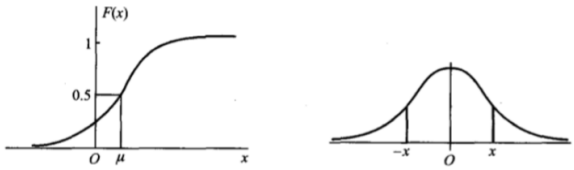
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

特别地，当$\mu = 0, \sigma = 1$时候称随机变量X服从标准正态分布。

对于任意一个正态分布，经过一个线性变换就能转化成标准正态分布。

若$X \sim N(\mu, \sigma^2)$，则$Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$

正态分布函数与密度函数分别如下



尽管正态变量的取之范围是$(-\infty, \infty)$，但它的值落在$(\mu - 3\sigma, \mu + 3\sigma)$内几乎是肯定的事，所以常被认为是3σ法则（1σ概率68.26%，2σ概率95.44%，3σ概率99.76%）。

1.3.3 随机变量的数字特征

1.3.3.1 数学期望（一阶原点矩）

若离散型随机变量X的分布律为：

$$P\{X = x_k\} = p_k, k = 1, 2, \dots$$

若级数

$$\sum_{k=1}^{\infty} x_k p_k$$

绝对收敛，则该级数的和为随机变量X的数学期望，记为E(X)。

同理：

设连续型随机变量X的概率密度函数为f(x)，若积分

$$\int_{-\infty}^{\infty} x f(x) dx$$

绝对收敛，则该积分的值为随机变量X的数学期望，记为E(X)。

数学期望E(X)由随机变量X的概率分布确定，若X服从某一分布，则E(X)为这一分布的数学期望。

1.3.3.2 方差（二阶中心矩）

数学期望可以评估平均水平，但实际中往往还需要评估偏离程度，于是用

$$E\{|X - E(X)|\}$$

来度量偏离程度，因绝对值运算不方便，故使用

$$E\{[X-E(X)]^2\}$$

来度量偏离程度，记为D(X)或Var(X)（在实际应用中引入$\sqrt{D(X)}$，记为σ(X)）表示标准差。

对于分布率为$P\{X=x_k\}=p_k, k=1,2,\ldots$的离散型随机变量有：

$$D(X)=\sum_{k=1}^{\infty}[X_k-E(X)]^2p_k$$

对于概率密度是f(x)的连续型随机变量有：

$$D(X)=\int_{-\infty}^{\infty}[x-E(X)]^2f(x)dx$$

1.3.3.3 协方差与相关系数

对于二维随机变量(X, Y)，除了X与Y的数学期望和方差之外，还需要度量X与Y之间的相互关系。

若两个随机变量X与Y相互独立，则容易得出：

$$E\{[X-E(X)][Y-E(Y)]\}=0$$

这说明当$E\{[X-E(X)][Y-E(Y)]\}\neq 0$时，X与Y存在一定关系（不相互独立）。

于是记$E\{[X-E(X)][Y-E(Y)]\}$为随机变量X与Y的协方差，记为Cov(X, Y)，即：

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

而

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

就是随机变量X与Y的相关系数。

于是对于任意两个随机变量X与Y，有

$$D(X+Y)=D(X)+D(Y)+2Cov(X,Y)$$

$$Cov(X,Y)=E(XY)-E(X)E(Y)$$

1.4 机器学习中的统计学基础

1.4.1 训练样本预测测试样本是可行的——样本与抽样分布

在概率论中，我们所研究的随机变量，它的分布都是假设已知的，在这一前提下去研究它性质，特点和规律；在数理统计中，我们研究的随机变量，总体的分布一般是未知的，或只知道它包含某种未知数的某种形式，于是通过对所研究对随机变量进行独立重复的观察，抽取一部分个体（样本）的观察值进行分析，根据获得的数据来对总体分布进行统计推断。对样本进行抽样时候一般有无放回抽样和有放回抽样，整理观察值一般有直方图与箱线图。

为了进行统计推断，一般不直接使用统计本身，而是针对不同问题构造适当的函数，常称其为统计量，以下为几个常用统计量。

样本平均值：

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

更广泛地有样本k阶（原点）矩：

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, \ldots$$

样本方差为二阶中心矩的（无偏估计）：

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

更广泛地有样本k阶中心矩：

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, k = 2, 3, \ldots$$

同理，可以作出与总体分布函数$F(x)$相应的统计量——经验分布函数。

设$X_1, X_2, \ldots, X_n$是总体F中的一个样本，用$S(x), -\infty < x < \infty$表示$X_1, X_2, \ldots, X_n$中不大雨x的随机变量的个数，定义经验分布函数$F_n(x)$为

$$F_n(x) = \frac{1}{n}S(x), -\infty < x < \infty$$

对于经验分布函数$F_n(x)$，格里汶科（Glivenko）在1993年证明了以下结果：对于任意实数x，当$n \to \infty$时，$F_n(x)$以概率1一致收敛与分布函数$F(x)$，即

$$P\{\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\} = 1$$

因此，对于任一实数x当n充分大时，经验分布函数的任意观察值$F_n(x)$与总体分布函数$F(x)$只有微小的差别，从而在实际中可以当作$F(x)$来使用。

1.4.2 机器学习中的预测——参数估计

1.4.2.1 点估计

设总体X的分布函数的形式已知，但它的一个或多个参数未知，借助于总体X的一个样本来估计总体未知参数的值的问题记为点估计。

点估计问题的一般提法如下：

设总体X的分布函数$F(x; \theta)$的形式为已知，θ是待估参数，$X_1, X_2, \ldots, X_n$是X的一个样本，$x_1, x_2, \ldots, x_n$是相应的一个样本值，点估计的问题就是构造一个适当的统计量$\hat{\theta}(X_1, X_2, \ldots, X_n)$，用它的观测值$\hat{\theta}(x_1, x_2, \ldots, x_n)$作为未知参数$\theta$的近似值，我们称$\hat{\theta}(X_1, X_2, \ldots, X_n)$为估计量，称$\hat{\theta}(x_1, x_2, \ldots, x_n)$为估计值。

常用构造估计量的方法为矩估计法和最大似然估计法

矩估计法

设X为连续型随机变量，其概率密度为$f(x;\theta_1,\theta_2,\ldots,\theta_k)$，或X为离散型随机变量，其分布律为
$P\{X=x\}=p(x;\theta_1,\theta_2,\ldots,\theta_k)$，其中$\theta_1,\theta_2,\ldots,\theta_k$为待估参数，$X_1,X_2,\ldots,X_n$是来自X的样本，假设总体X的前k阶矩$(l=1,2,\ldots,k)$

X连续型

$$\mu_l=E(X^l)=\int_{-\infty}^{\infty}x^l f(x;\theta_1,\theta_2,\ldots,\theta_k)dx$$

X离散型（R_X是X可能取值的范围）

$$\mu_l=E(X^l)=\sum_{x\in R_X}x^l p(x;\theta_1,\theta_2,\ldots,\theta_k)$$

存在，一般来说，它们是$\theta_1,\theta_2,\ldots,\theta_k$的函数，基于样本矩

$$A_i=\frac{1}{n}\sum_{i=1}^{n}X_i^l$$

依概率收敛于相应的总体矩$\mu_l(l=1,2,\ldots,k)$，样本矩的连续函数依概率收敛于相应的总体矩的连续函数，我们就用样本矩作为相应的总体矩的估计量，而以样本矩的连续函数作为相应总体矩的连续函数的估计量。这种估计方法就是矩估计法。

具体而言，设

$$\begin{cases}\mu_1=\mu_1(\theta_1,\theta_2,\ldots,\theta_k),\\ \mu_2=\mu_2(\theta_1,\theta_2,\ldots,\theta_k),\\ \cdots\cdots\\ \mu_k=\mu_k(\theta_1,\theta_2,\ldots,\theta_k)\end{cases}$$

一般而言对联立方程组求解可得

$$\begin{cases}\theta_1=\theta_1(\mu_1,\mu_2,\ldots,\mu_k),\\ \theta_2=\theta_2(\mu_1,\mu_2,\ldots,\mu_k),\\ \cdots\cdots\\ \theta_k=\theta_k(\mu_1,\mu_2,\ldots,\mu_k)\end{cases}$$

以A_i分别代替μ_i，就以

$$\hat{\theta}_i=\theta_i(A_1,A_2,\ldots,A_k),\ i=1,2,\ldots,k$$

分别作为θ_i的估计量，这种估计量称为矩估计量。

最大似然估计

若总体X属于离散型，其分布律$P\{X = x\} = p(x; \theta), \theta \in \Theta$的形式为已知，$\theta$为待估参数，$\Theta$是$\theta$可能取值的范围，设$X_1, X_2, \ldots, X_n$是来自X的样本，则$X_1, X_2, \ldots, X_n$的联合分布律为

$$\prod_{i=1}^{n} p(x_i; \theta)$$

又设$x_i, x_2, \ldots, x_n$是相应于样本$X_i, X_2, \ldots, X_n$的一个样本值，可得到$X_i, X_2, \ldots, X_n$取到观察值$x_i, x_2, \ldots, x_n$的概率值，即事件$\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$发生的概率为

$$L(\theta) = L(x_i, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta), \theta \in \Theta$$

这以概率随θ的取值而变化，它是θ的函数，$L(\theta)$称为样本的似然函数。

直观理解，现在已经取到样本值$x_i, x_2, \ldots, x_n$，这表明取到这一样本值的概率$L(\theta)$较大，所以我们不会考虑那些不能使样本$x_i, x_2, \ldots, x_n$出现的$\theta \in \Theta$作为θ的估计；反之，如果已知$\theta = \theta_0 \in \Theta$时有$L(\theta)$取得很大值，而$\Theta$中的其他$\theta$的值使$L(\theta)$取很小值，我们自然可以认为取$\theta_0$作为未知参数$\theta$的估计值较为合理。

由费希尔（R.A.Fisher）引进的最大似然估计法就是固定样本观察值$x_i, x_2, \ldots, x_n$，在θ取值的可能范围Θ内挑选使似然函数$L(x_i, x_2, \ldots, x_n; \theta)$达到最大的参数值$\hat{\theta}$，作为参数$\theta$的估计值，即取$\hat{\theta}$使

$$L(x_i, x_2, \ldots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_i, x_2, \ldots, x_n; \theta)$$

这样得到的$\hat{\theta}$与样本值$x_i, x_2, \ldots, x_n$有关，常记为$\hat{\theta}(x_i, x_2, \ldots, x_n)$，称为参数$\theta$的最大似然估计值，而相应的统计量$\hat{\theta}(X_i, X_2, \ldots, X_n)$称为参数$\theta$的最大似然估计量。

1.4.2.2 区间估计

对于一个未知量，在测量和计算时，有时近似值不足以满足要求，还需要估计误差，即要知道近似值的精确程度，类似地，对未知参数θ，除了对其进行点估计$\hat{\theta}$外，我们还需要估计出一个范围，并希望知道这个范围包含参数θ真值的可信程度，这样的范围通常以区间的形式给出，这样的区间就是置信区间。

设总体X的分布函数$F(x; \theta)$含有一个未知参数$\theta, \theta \in \Theta$（$\Theta$是$\theta$可能取值的范围），对于给定值$\alpha$（$0 < \alpha < 1$），若由来自样本$X_i, X_2, \ldots, X_n$确定的两个统计量$\underline{\theta}_1 = \underline{\theta}_1(X_i, X_2, \ldots, X_n)$和$\overline{\theta}_2 = \overline{\theta}_2(X_i, X_2, \ldots, X_n)$（$\underline{\theta}_1 < \overline{\theta}_2$），对于任意$\theta \in \Theta$满足

$$P\{\underline{\theta}_1(X_i, X_2, \ldots, X_n) < \theta < \overline{\theta}_2 = (X_i, X_2, \ldots, X_n)\} \geq 1 - \alpha$$

则称随机区间$(\underline{\theta}_1, \overline{\theta}_2)$是$\theta$的置信水平为$1 - \alpha$的置信区间，$\underline{\theta}_1$和$\overline{\theta}_2$分别为置信下限和置信上限。

直观理解，若反复抽样多次（各次得打的样本容量相等，都是n），每个样本值确定一个区间$(\underline{\theta}_1, \overline{\theta}_2)$，每个这样的区间要么包含$\theta$的真值，要不不包含$\theta$的真值，按伯努利大数定理，这么多的区间中，包含$\theta$真值的约占$100(1 - \alpha)\%$，不包含$\theta$真值的约占$100\alpha\%$。

1.4.3 判定模型是否合理——假设检验

1.4.3.1 假设检验与两类错误

在总体的分布函数完全未知或只知其形式，但不知具体参数的情况下，为了推断总体的某些未知特性，提出关于某些关于总体的假设，我们要根据样本对所提出的假设作出接受或拒绝的决策，假设检验就是这一决策的过程。

对于一个假设检验问题，在显著性水平α下，检验假设

$$H_0: \mu = \mu_0,$$

$$H_1: \mu \neq \mu_0.$$

其中H_0为原假设或零假设，H_1为备择假设。

由于检验是根据样本作出的，总有可能作出错误的决策，在假设H_0为真时，我们可能犯拒绝H_0的错误，称这一类“弃真”的错误为第I类错误；又当H_0实际上不真时，我们也有可能接受H_0，称这种“取伪”当错误为第II错误。

在确定检验法则时我们尽可能使犯两类错误的概率都小，但死当样本量固定的条件下，若减少犯一类错误的概率，往往另一类错误的概率会增加。一般来说，我们总是控制犯第I类错误的概率，使它不大于α（α一般取0.1, 0.05, 0.01, 0.005），这种只对犯第I类错误对概率进行控制，而不考虑犯第II类错误的概率的检验称为显著性检验。

1.4.3.2 一个正态分布的均值检验（例）

例：某机器生产的产品净重是一个随机变量，它服从均值0.5kg，标准差0.015kg，某日随机抽取9袋进行检查，称得净重为(kg)

0.497 | 0.506 | 0.518 | 0.524 | 0.498 | 0.511 | 0.520 | 0.515 | 0.512

问机器是否正常（置信度0.05）？

解：

$$H_0: \mu = \mu_0 = 0.5$$

$$H_1: \mu \neq \mu_0$$

当H_0为真时有

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

重而衡量$|\bar{x} - \mu_0|$当大小可以归结为衡量$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$的大小。我们可适当选择一正数k，使当观察值$\bar{x}$满足$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \geq k$时就拒绝假设$H_0$，反之$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < k$就接受$H_0$。

即当满足

$$|z| = |\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}| \geq k = z_{\alpha/2}$$

则接受H_0，而若

$$|z| = |\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}| < k = z_{\alpha/2}$$

则接受H_0。

取$\alpha = 0.05$计算得

$$|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}| = 2.2 > 1.96$$

故在0.05置信度下拒绝H_0，机器不正常。

1.4.3.3 假设检验解决的一些问题

- 单个总体均值检验
- 多个总体均值检验
- 单个总体方差检验
- 多个总体方差检验
- 样本容量选取
- 分布拟合检验

1.4.4 大数据让机器学习效果更好－大数定律与中心极限定理

1.4.4.1 大数定律

当随机事件A进行大量重复试验以后，其频率$f_n(x)$会随着n的增大而出现稳定性，稳定在一个常数附近，这种稳定性是概率定义的客观基础。

辛钦大数定理（弱大数定律）表示在数学期望存在的条件下其数学期望依概率收敛。设$X_1, X_2, \ldots$是相互独立同分布的随机变量序列，且具有数学期望$E(X_k) = \mu(k = 1, 2, \ldots)$作前n和变量的算术平均$\frac{1}{n}\sum_{k=1}^{n} X_k$，对任意的$\varepsilon > 0$有：

$$\lim_{n\to\infty} P\{|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu| < \varepsilon\} = 1$$

对于连续型随机变量，设f_A是n次独立重复试验中事件A发生的次数，p是事件A在每次试验中发生的概率，则对于任意正数$\varepsilon > 0$有：

$$\lim_{n\to\infty} P\{|\frac{f_A}{n} - p| < \varepsilon\} = 1$$

或

$$\lim_{n\to\infty} P\{|\frac{f_A}{n} - p| \geq \varepsilon\} = 0$$

这就是著名的伯努利大数定理。

弱大数定律是依概率收敛，强大数定律则是以概率1收敛（几乎处处收敛），只要n足够大，任意指定一个正整数ε，总能找到一个N，使当n > N时，前n个变量的算术平均与μ的差大于ε的次数有限的。表达式可定义为：

$$P\{\lim_{n\to\infty}\frac{X_1+X_2+\ldots+X_n}{n}=\mu\}=1$$

弱大数定律表明对于足够大的值n^*，随机变量$\frac{X_1+X_2+\ldots+X_{n^*}}{n^*}$的值靠近μ，但它不能保证所有的$n>n^*$，$\frac{X_1+X_2+\ldots+X_{n^*}}{n^*}$仍然停留在μ附近，因此$|\frac{X_1+X_2+\ldots+X_{n^*}}{n^*}-\mu|$可以无限多次离开0（尽管较大偏离的频率不会很高）。而强大数定律能保证这种情况不会发生，特别地，强大数定律表明这种表示以概率1成立。即对于任意的$\varepsilon>0$有：

$$|\sum_{i=i}^{n}\frac{X_i}{n}-\mu|>\varepsilon$$

只会出现有限次。

1.4.4.2 中心极限定理

在客观实际中有许多随机变量，他们由大量相互独立的随机因素综合影响所形成，其中每一个别因素影响在总的影响中所起的作用都是微小的，这种随机变量往往服从正态分布。这种现象构成了中心极限定理的客观背景。

独立同分布的中心极限定理（表达式构造比较复杂）表示均值为μ，方差为$\sigma^2>0$的独立同分布的随机变量$X_1, X_2, \ldots, X_n$之和$\sum_{k=1}^{n}X_k$的标准化变量，当n充分大时，近似成立

$$\frac{\sum_{k=1}^{n}X_A-n\mu}{\sqrt{n}\sigma}\sim N(0,1)$$

李雅普诺夫（Lyapunov定理）（表达式构造比较复杂）表示，在定理的条件下，随机变量

$$Z_n=\frac{\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}\mu_k}{B_n}$$

当n很大时，近似地服从正态分布N(0, 1)。

棣莫弗-拉普拉斯（De Moivre-Laplace）定理（表达式构造比较复杂）说明正态分布是二项分布的极限分布。

1.5 用数理统计解决实际问题

1.5.1 回归

这里主要讨论一元线性回归，一般表达式为$y=a+bx$。

假设对于x（在某一区间内）的每一个值有

$$Y\sim N(a+bx,\sigma^2)$$

其中a, b及σ^2都不是依赖于x的未知参数，及$\varepsilon=Y-(x+bx)$，相当于假设

$$Y=a+bx+\varepsilon,\varepsilon\sim N(0,\sigma^2)$$

下面对a, b进行参数估计（中间结果省略）得到一个方程组

$$\begin{cases} na + (\sum_{i=1}^{n} x_i)b = \sum_{i=1}^{n} y_i, \\ (\sum_{i=1}^{n} x_i)a + (\sum_{i=1}^{n} x_i^2)b = \sum_{i=1}^{n} x_i y_i \end{cases}$$

解得最大似然估计值为

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\hat{b}}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{b}\bar{x}$$

在得到a, b的估计$\hat{a}, \hat{b}$后，我们就取$\hat{a} + \hat{b}x$为回归函数$\mu(x) = a + bx$的估计。

$$\hat{y} = \hat{a} + \hat{b}x$$

就为一元回归方程。

1.5.2 朴素贝叶斯

贝叶斯的思想主要为通过先验概率加数据得后验概率。这里主要讨论朴素贝叶斯的算法过程。

设训练集为m和样本n和纬度

$$(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}, y_1), (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}, y_2), \ldots, (x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}, y_m)$$

共有k个特征输出类别，分别为$C_1, C_2, \ldots, C_k$，每个特征输出的类别的样本数为$m_1, m_2, \ldots, m_k$，在第k个类别中，如果是离散特征，则特征X_j各个类别取值为$m_{jl}, l = 1, 2, \ldots, S_j$，$S_j$为特征j的 不同取值数。

输出为X^t的分类

算法流程如下：

1.如果没有Y的先验概率，则计算Y的K个先验概率

$$P(Y = C_k) = (m_k + \lambda)/(m + K\lambda)$$

否则$P(Y = C_k)$为输入的先验概率

2.分别计算第k个类别的第j纬特征的第l个取值条件概率

$$P(X_j = x_{jl} | Y = C_k)$$

如果是离散值

$$P(X_j = x_{jl} | Y = C_k) = \frac{m_{kjl} + \lambda}{m_k + S_j\lambda}, \lambda > 0$$

如果是连续值直接求正态分布参数

$$P(X_j = x_j | Y = C_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x_j - \mu_k)^2}{2\sigma_k^2}}$$

μ_k为样本C_k中的所有X_j平均值，σ_k^2为方差。

对于需要预测的X^t，分别计算

$$P(Y = X_k)\prod_{j=1}^{n} P(X_j = x_j^t \mid Y = C_k)$$

X^t的分类C_{result}为

$$C_{result} = \max_{C_k} P(Y = C_k)\prod_{j=1}^{n} P(X_j = X_j^t \mid Y = C_k))$$

1.5.3 航空公司预售票策略

1.5.3.1 问题分析

对于一次航班，若航空公司限制订票对数量恰等于飞机法容量，那么由于总会有一些定了机票但不按时登机的乘客，致使飞机因不满员飞行而利润下降。而如果不限制预订票数量，那么当持票按时前来的乘客超过飞机容量时，必然引起乘客的抱怨，影响其社会声誉。所以航空公司需要综合考虑紧急利益与社会声誉，确定订票数量的最佳限额。

公司的经济利益可以用机票收入扣除飞行费用和赔偿金后的利润来衡量，社会声誉可以用持票前来登机，但因满员无法飞走的乘客限制在一定数量来衡量。该问题的关键因素——预定票乘客是否能按时来登机，这个因素是随机的。所以经济利益和社会声誉都需要在平均意义下衡量。这两个目标的优化问题，决策变量都是预定票数量的限额。

1.5.3.2 模型假设

1.飞机容量为n，假票价格为g，飞行费用为r，机票价格按照$g = r/\lambda n$来制定，其中λ为利润调节因子，如$\lambda = 0.6$表示飞机60%满员率就不亏本。

2.预定票数量的限额为m（$m > n$），每位乘客不按时登机的概率为p，各位乘客之间是否按时登机相互独立。

3.无法飞走的乘客获得赔偿金为b

1.5.3.3 模型建立

公司经济利益用平均理论S来衡量，每次航班的利润s为从机票收入中减飞行费用和可能发生的赔偿金。当m为乘客有k为不按时来登机时

$$s = \begin{cases} (m-k)g - r & m-k \le n, \\ ng - r - (m-k-n)b & m-k > n \end{cases}$$

由假设2，不按前来登机当乘客数k服从二项分布，于是概率

$$p_k = P(K = k) = \binom{m}{k} p^k q^{m-k}, q = 1 - p$$

平均利润S(s的期望)为

$$S(m) = \sum_{k=0}^{m-n-1} [ng - r - (m - k - n)b]p_k + \sum_{k=m-n}^{m} [(m - k)g - r]p_k$$

其中$\sum_{k=0}^{m} kp_k = mp$

$$S(m) = qmg - r - (g + b) \sum_{k=0}^{m-n-1} (m - k - n)p_k$$

当n, g, r, p给定后，可以求m使得S(m)最大。

从公司声誉和经济利益两方面考虑，应该让不能飞走当乘客不要太多，而由于这个数量是随机的，可以用不能飞走的乘客数超过若干人的概率作为度量指标，记不超过j个人的概率为$P_j(m)$，因为不能飞走的乘客超过j人，等价于m位预定票的乘客中不按时来登机的不超过$m - n - j - 1$人，所以

$$P_j(m) = \sum_{k=0}^{m-n-j-1} p_k$$

对于给定的n, j，显然当 $m = n + j$时不能飞走的乘客不超过 j 人，$P_j m = 0$，而当m变大时$P_j m$单调增加。

以上S(m)和$P_j(m)$无法解析求解，需要进行数值计算求解。

1.6 更多数学模型

- 优化模型
- 数学规划模型
- 微分方程模型
- 代数方程与差分方程模型
- 稳定性模型
- 离散模型
- 概率模型
- 统计回归模型
- 博弈模型
- 马氏链模型
- 动态优化模型