第二章 概率论

1. **基本概念**
    - 概率
        - 频率学派：从事物本身出发，完全从客观数据的角度来理解概率，使用随机事件的发生的频率描述概率
        - 贝叶斯学派: 从人的角度出发，表示为人对事物的感受。可通俗理解如下：在接触某个事物前，人对事物有个预想的认识(**先验概率**)，在接触这个事物后，人会适当改变自己对这个事物的认识，得到新的认识(**后验概率**)
        - 如上可见，频率派关注事物本身，贝叶斯派关注我们对事物的认识。从点估计和贝叶斯估计角度可以这样理解
            - 我们对事物的数学认识就是一个模型，具体的模型有具体的参数，当确定模型空间后，我们认识事物就是确定模型参数。事物本身是确定的，则模型的参数也应该是确定，频率学派就是要学习这个确定参数。这就是点估计，选择最优的参数值(最大后验，最大似然)可见这个学派是追求真理的，希望能找到事物的真面目。而贝叶斯派不理会事物本身，在乎我们对事物的认识，事物本身会有一个确定的参数值，但我们不管，不选择参数为某个确定的值，而是对所有可能参数值，我们都有个相信的程度。即参数是一个随机变量，满足我们认识的一个分布，在看到数据前，这个分布是先验概率分布，在看到数据后，分布变成后验概率分布
    - 随机变量
        - 数学建模的一个基本单位，描述某个事物的状态，它具有多个状态
        - 区别于函数里的自变量，自变量是纯数学的，没有实际含义的。而随机变量是对应着实际事物，是事物的一个数学表示。
        - 随机变量的取值可以是离散的，也可以是连续的
    - 分布
        - 用来描述随机变量的取值情况。
        - 区分实际分布和数学分布
            - 实际分布指的是事物的真实情况。大多情况下，我们不能知道事物的真实情况或者不能简单地表达真实情况。因此我们需要数学分布，我们定义一些分布函数，比如高斯分布，来极大可能地描述事物的真实情况。可以看做对事物的真实情况的数学建模
    - 在贝叶斯体系里，先验概率是边缘概率，后验概率是条件概率(看到数据后)
2. **一些公式**
    - 乘法法则(product rule)

    $$p(A,B)=p(A \Lambda B)=p(A|B)p(B)$$

    - 加法法则(sum rule)

    $$p(A)=\sum_b p(A,B)=\sum_b p(A|B=b)p(B=b)$$

    - 链式法则(chain rule)

    $$p(X_{1:D})=p(X_1)p(X_2|X_1)p(X_3|X_2,X_1)\ldots p(X_D|X_{1:D-1})$$

    - 条件概率

    $$p(A|B)=\frac{p(A,B)}{p(B)} \quad if \quad p(B)>0$$

    - 贝叶斯法则

    $$p(X=x|Y=y)=\frac{p(X=x,Y=y)}{p(Y=y)}=\frac{p(X=x)p(Y=y|X=x)}{\sum_{x'}p(X=x')p(Y=y|X=x')}$$

        - 贝叶斯公式最大作用是当$p(X|Y)$难求时，可以转化成求$p(X)$和$p(Y|X)$
        - 关于生成模型和判别模型(个人想法)
            - 在分类问题里，我们最后需要的是$p(Y|X)$
            - 在判别模型中，我们直接学习得到$p(Y|X)$，直接判断
            - 在生成模型中，我们学习的是$p(Y)$和$p(X|Y)$，然后通过贝叶斯法则得到$p(Y|X)$。注意其中$p(X|Y)$是数据的分布，可以生成数据的，可能因此称为生成模型吧。
3. **独立性**
    - 事物X和事物Y独立，记为$X \perp Y$

    $$X \perp Y \Leftrightarrow p(X,Y)=p(X)p(Y)$$

    - 在实际中很少存在事物A和事物B是真的独立的，很多东西表面上看是没有关系的，但它们之间通过一些其他事物关联起来，一个能通过内在关联事物影响另一个，所以只有当内在的关联事物确定时，两者才独立。于是有条件独立如下：$X \perp Y|Z \Leftrightarrow p(X,Y|Z)=p(X|Z)p(Y|Z)$
    - 在给实际问题数学建模时要注意使用独立性，因为通过独立性可以将联合分布变成边缘分布，而边缘分布往往比联合分布更容易处理。当然独立性是少有的，所以我们要利用好条件独立
4. **关于分布的其他一些概念**
    - 分位数quantile
        - 通过对累计分布函数$F(X)$求逆$F^{-1}(X)$,利用$F^{-1}(X)$可以方便地计算分布的分位点。比如$F^{-1}(0.5)$表示分布

的中点，$F^{-1}(0.25)$表示分布的四分位点

- 均值

$$E[X]=\sum_{x\in\chi}xp(x)$$

- 方差

$$var[X]=E[X-\mu]^2$$

5. **一些重要的离散分布**
   - 二项分布
     - 一个试验，有两种结果(成功或失败)，成功的概率为θ，独立进行n次，其中成功的次数K是一个随机变量，满足二项分布，公式如下

$$Bin(k|n,\theta)=\frac{n!}{(n-k)!k!}\theta^k(1-\theta)^{n-k}$$

期望$n\theta$，方差$n\theta(1-\theta)$

     - 当n为1时，二项分布称为伯努利分布
   - 多项分布
     - 一个试验，有多种结果，设有D种结果，第i种结果发生的概率为θi。独立进行n次，则全部试验结果可以用一个D维向量$\vec{X}$来表示，每个维度对应一种结果，其数字表示该种结果在试验中发生的次数，向量$\vec{X}$满足多项分布，公式如下

$$Mu(\vec{x}|n,\theta)=\frac{n!}{x1!x2!\ldots xd!}\prod_{j=1}^{D}\theta_j^{xj}$$

   - 泊松分布
     - 多用来描述一段时间内，发生某事件的次数X
     - 实际上来源于二项分布的极限形式，可以如下理解：把一段时间等分成n份，当n无限大时，每份时间无限小，可以认为每份时间里只有两种情况发生了某事件一次或者没发生，不会发生事件多次。那么一份时间就是一个伯努利试验，因为时间无限小，所以发生事件的概率θ无限小。对应到二项分布，则是取n无限大，θ无限小。最后的结果即是泊松分布，公式如下

$$Poi(x|\lambda)=e^{-\lambda}\frac{\lambda^x}{x!}$$

     - 泊松分布的期望和方差均为λ
   - 经验分布(the empirical distribution)
     - 利用看到的数据(样本数据)来估计近似总体分布。比如数据集中a类占10%，则认为总体分布中p(a)=0.1
6. **一些重要的连续分布**
   - 高斯(正态)分布

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

     - 最大熵意义：在统计中，我们在满足已知限制后，希望分布的熵越大越好。在没有任何限制情况下，均匀分布熵最大。在已知均值和方差情况下，高斯分布熵最大。
     - 中心极限定理:多个独立的随机变量，它们的分布的均值相同且方差相同。则在这些随机变量组成的总体中，多次采样，随着采样次数增多，采样分布趋向于高斯分布
   - 拉普拉斯分布

$$Lap(x|\mu,b)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$$

     - 其中$mean=\mu, mode=\mu, var=2b^2$
   - 学生t分布(Student t distribution)

$$T(x|\mu,\sigma^2,v)\propto[1+\frac{1}{v}(\frac{x-\mu}{\sigma})^2]^{-(\frac{v+1}{2})}$$

     - 其中$mean=\mu, mode=\mu, var=\frac{v\sigma^2}{v-2}$
   - 三个分布的分析与比较
     - 高斯分布，拉普拉斯分布，学生t分布都是表示整个实数域上的随机变量，它们的形状都是一个单峰，中间高，两边低。由公式可知，高斯分布是x和均值的差的平方作为指数的指数下降，下降速度很快。而拉普拉斯分布则是以x和均值的差的绝对值作为指数的指数下降，中间时(差小于1时)下降得比高斯分布快，两边时(差大于1时)下降得比高斯分布慢。学生t分布则以负幂函数的速度下降，远小于指数下降速度。
     - 长尾问题:如上三个分布都是单峰形状，数据集中在中间的波峰，两边称为尾巴。尾巴包含数据多则称该分布具有长

尾。而长尾直接依赖于下降速度，尾巴下降地慢，则显得平坦一点。从如上分析可以知道学生t分布具有最好的长尾性质，拉普拉斯分布次之，高斯分布最差。长尾性质影响着分布的对数据的鲁棒性，受outlier的影响程度。越大的尾巴，鲁棒性越好。如果像高斯分布那种两边急速下降，当数据中出现outlier，因为两边的分布量太少满足不了outlier，所以高斯分布不得不偏移向outlier，效果不好。

- 伽马分布
  - 用于表示范围在正实数的连续随机变量，即$x > 0$
  - 伽马函数

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

  - 伽马分布

$$Ga(T|shape = a, rate = b) = \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb}$$

  - 其中mean=$\frac{a}{b}$, mode=$\frac{a-1}{b}$, var=$\frac{a}{b^2}$
- 贝塔分布
  - 用于表示范围在[0,1]之间的连续随机变量。
  - 贝塔函数

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

  - 贝塔分布

$$Beta(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

  - 其中mean=$\frac{a}{a+b}$, mode=$\frac{a-1}{a+b-2}$, var=$\frac{ab}{(a+b)^2(a+b+1)}$
- Pareto distribution
  - 用来对具有长尾现象的数据进行建模，比如单词的使用的频率，把全世界人民对单词的使用看做一个量的话，可以发现占总单词数一小部分的常用词占使用量的大部分，而其他那些词则占使用量的小部分。这就是长尾，这个长尾虽然小，但稳定，持久，不为零。这是引申到经济统计学中“二八法则”。80%的资源掌握在20%的人手里，这就是一种长尾现象。在生活中有很多长尾现象，比如淘宝大部分销量在于一小部分商品上。

$$Pareto(x|k,m) = km^k x^{-k+1} I(x \geq m)$$

  - 本质上是一个负幂函数，k控制下降速度，m控制截断阈值
  - 其中mean=$\frac{km}{k-1}$ if k > 1, mode=m, var=$\frac{m^2k}{(k-1)^2(k-2)}$ if k > 2

7. **联合分布**
   - 之前的分布是对于单个随机变量来说的，对离散变量来说，其pmf是一维列表，对连续变量来说，其pdf是一元函数。现在联合分布是对多个随机变量的分布，对离散变量来说，其pmf是多维列表，对连续变量变量来说，其pdf是多元函数。
   - 随机变量之间的线性相关性
     - 协方差(covariance)
       - 用来度量两个随机变量之间的线性关系

$$cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X][Y]$$

       - 在多个随机变量情况中，算两两之间的协方差，组成协方差矩阵
     - 协相关性系数(correlation coefficient)
       - 用协方差度量线性相关性有个问题就是它的大小依赖与数据本身，而不仅仅依赖两个变量之间的线性相关性。而协相关性系数就是对协方差进行归一化，限制到[-1,1]之间。大于零代表正相关，小于零代表负相关。

$$corr[X,Y] = \frac{cov[X,Y]}{\sqrt{var[X]var[Y]}}$$

       - 同样地算两两之间的协相关性可以组成协相关性系数矩阵
       - 两个随机变量独立则肯定不线性相关，但是不线性相关不一定独立。可以这样理解独立是一点关系都没有，而线性相关只是关系的一种而已。
8. **随机变量之间的变换**
   - 已经知道随机变量X的分布，那么X的变换随机变量Y=f(X)的分布是什么呢？
   - 线性变换
     - 在线性变换$\vec{y} = f(\vec{x}) = AX + \vec{b}$中，我们可以通过如下公式从X的均值和方差来算出Y分布的均值和方差

$$E[\vec{y}] = E[A\vec{x} + \vec{b}] = A\vec{\mu} + \vec{b}$$

$$cov[\vec{y}] = cov[A\vec{x} + \vec{b}] = Acov[\vec{x}]A^T$$

        - 因为高斯分布只要确定均值和方差即完全确定分布本身了，则如果Y的分布是高斯分布且是线性变换，通过如上公式就可以确定Y的均值和方差，从而确定Y的分布
    - 一般变换
        - 对离散变量，我们只需要计算y=f(x)，再把对应的值累加起来，具体见如下公式:

$$p_y(y) = \sum_{x:f(x)=y} p_x(x)$$

        - 对连续变量，为了代入y=f(x)，我们先计算y的cdf，公式如下:

$$P_y(y) = P(Y \le y) = P(f(X) \le y) = P(X \le f^{-1}(y)) = P_x(f^{-1}(y))$$

            - 对y的cdf求导就得到y的pdf

$$p_y(y) = \frac{d}{dy}P_y(y) = \frac{d}{dy}P_x(f^{-1}(y)) = \frac{dx}{dy}\frac{d}{dx}P_x(x) = \frac{dx}{dy}p_x(x)$$

    - 蒙特卡洛近似
        - 通常来说，通过如上变换公式来计算变换后变量的分布是很困难的(求导一般较难)。因此我们提出一种方法来近似计算变换后的分布
        - 核心思想是用样本分布近似总体分布
        - 具体方法：在原X分布中多次采样得到样本集{x1,x2,x3...}，对样本应用变换得到变换后的样本集{f(x1),f(x2),f(x3)...},可以认为这个样本集是变换后的Y分布的一个样本集，可以用其样本分布来近似Y的总体分布

9. **信息论**
    - 熵(entropy)
        - 度量一个随机变量的不确定性

$$H(X) = -\sum_{k=1}^{K} p(X=k)log_2 p(X=k)$$

        - 可以从编码角度来了解，$log_2 p(X=k)$表示在分布p情况下编码X=k信息所需要的编码数，$p(X=k)$表示这种信息的比例，作为权值，加权相加。因此熵表示编码该随机变量的最小平均编码数。里面的编码思想是最小化编码数，比例越大的信息，用越少的编码数(参考哈夫曼编码)
    - KL离散度(KL divergence)
        - 交叉熵

$$H(p,q) = -\sum_k p_k log q_k$$

            - 对比熵的公式，交叉熵表示在分布q的情况下编码信息，但是实际分布是p，因为p和q不同，肯定不是最小编码。而且当p和q差异越大时，编码数越大。因此可以用交叉熵来度量两个分布之间的差异性。注意交叉熵不是对称的，即H(p,q)不一定等于H(q,p)
        - 因为交叉熵的大小不是纯粹地依赖两个分布的差异性，还依赖与分布本身的熵值。为了更好地度量两个分布之间的差异性，我们多使用KL离散度

$$KL(p||q) = \sum_{k=1}^{K} p_k log\frac{p_k}{q_k} = \sum_k p_k log p_k - \sum p_k log q_k = -H(p) + H(p,q)$$

        - 从公式可知，KL离散度表示在错误的分布q的情况下编码信息比正常在分布p情况下编码信息所额外耗费编码数。这样就排除分布本身熵大小的影响，更好地度量两个分布之间的差异性。
        - 同样地，KL离散度也不是对称的，即KL(p||q)不一定等于KL(q||p)
        - 在机器学习中，经常使用KL离散度来作为损失函数
    - 互信息(mutual information)
        - 对两个随机变量X和Y，两者包含的共同信息称为互信息。可以通过比较分布p(X,Y)和p(X)p(Y)的差异性来度量互信息。如果随机变量X和Y独立，则p(X,Y)等于p(X)p(Y)，两者差异性为零，互信息为零。互信息公式如下

$$I(X;Y) = KL(p(X,Y)||p(X)p(Y)) = \sum_x \sum_y p(x,y) log\frac{p(x,y)}{p(x)p(y)}$$

        - 条件熵
            - 熵表示随机变量的不确定性，而条件熵表示在知道另一个随机变量情况，该随机变量的情况。

$$H(Y|X) = \sum_x p(x)H(Y|X=x)$$

            - 其中H(Y|X=x)表示条件概率分布P(Y|X=x)的熵
            - 通过条件熵来计算互信息

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$