Chapter 8: Sampling

采样的作用

从特定的概率分布中抽取对应的样本点

信息降维
- 采样得到的样本集可以看做一个非参数模型，即用少量的样本点（经验分布）来近似总体分布
- 由于总体分布可能存在无穷样本点，全部采集训练不现实，因此一般采用总体分布的一个样本集作为训练集来作为总体分布的近似，其之为训练集。
- 训练模型的时候是最小化模型在训练集上的损失函数(经验风险)，在评估模型时，使用另一个总体分布的样本集作为测试集
- 用于信息可视化，帮助人们快速、直观地了解总体分布中数据的结构和特性

对当前数据集重采样，充分利用已有数据集，挖掘更多信息
- 自助法
- 刀切法
- 利用重采样，有意识地改变样本的分布，以更适应后续的模型训练和学习
- 如利用重采样来处理分类模型的训练样本不均衡的问题

随机模拟
- 含有隐变量的复杂模型，可以利用古布斯采样对隐变量的分布进行采样
- 最数值型的近似求解方法，不同于一些确定性的近似求解方法，如变分贝叶斯方法、期望传播等

均匀分布随机数

计算机是确定性的，存储和计算单元只能处理离散状态值

计算机不能产生完全的均匀分布随机数，只能产生离散分布的伪随机数来逼近连续分布，使用很大的离散空间来提供足够的精度

生成方法

线性同余法 Linear Congruential Generator

$$x_{k+1} \equiv a \cdot x_k + c (mod \quad m)$$

根据当前生成的随机数x_t来进行适当变换，进而产生下一次的随机数x_(t+1)，初始值x_0称为随机种子

得到区间[0,m-1]上的随机整数，除以m得到区间[0,1]

随机数并不是相互独立的，最多产生m个不同的随机数

如果进行多次操作，得到的随机数序列会进入循环周期，这就需要精心选择合适的乘法因子a和模数m

$$\begin{cases} m & = 2^{31} - 1 \\ a & = 1103515245 \\ c & = 12345 \end{cases}$$

gcc采用版本

主要产生方式

几乎所有的采样方法都是以均匀分布随机数作为基本操作

常见采样方法

简单分布如有限离散分布可以用轮盘赌算法来采样

函数变换法
- 如果变换关系φ(·) 是x的累积分布函数的话，则得到所谓的逆变换采样(Inverse Transform Sampling)
- 假设待采样的目标分布的概率密度函数为p(x)，它的累积分布函数为
- 则逆变换采样法按照如下过程进行采样：
  1. 从均匀分布U(0,1)产生一个随机数u_i
  2. 计算x_i = Φ^-1(u_i)，其中Φ^-1(·) 便累积分布函数的逆函数

则通过该过程采样的x_i服从p(x)分布

如果带采样的目标分布的来及分布函数的逆函数无法求解或者不容易计算，则不适用逆变换采样法

改进

可以构造一个容易采样的参考分布，先对参考分布进行采样，然后对得到的样本进行一定的拒绝/重塑操作，使得最终的样本服从目标分布

拒绝采样(Accept-Regect Sampling)
- 自适应拒绝采样

在拒绝采样中，如果在某一步中采样被拒绝，则该步不会产生新样本，需要重新进行采样

重要性采样(Importance Sampling)

对于高维空间的随机向量，通常难以寻找合适的参考分布，采样效率低下，样本的接受概率小或重要性权重低

马尔科夫蒙特卡洛采样法MCMC：Markov Chain Monto Carlo

基于采样的数值型近似求解方法
- 蒙特卡洛法

用于进行采样
- 马尔科夫链

MCMC基本思想

针对特采样的目标分布，构造一个马尔科夫链，使得该马尔科夫链的平稳分布就是目标分布

然后，从任何一个初始状态出发，沿着马尔科夫链进行状态转移，最终得到的状态转移序列会收敛到目标分布，由此可以得到目标分布的一系列样本

核心点是构造合适的马尔科夫链，不同马尔科夫链对应不同的MCMC采样法
- Metropolis-Hastings采样法
- 吉布斯采样法

MCMC采样法每一步都会产生一个样本，只是有时候这个样本与之前的样本一样而已

MCMC采样法最在不断迭代过程中逐渐收敛到平稳分布的，因此实际应用中一般会对得到的样本序列进行burn-in处理，即截除掉序列中最开始的一本样本，只保留后面的样本

产生独立样本

MCMC产生的样本不是独立的，因为后一个样本是由前一个样本根据特定的转移概率得到的，或者有一定概率就是前一个样本，如果仅仅是采样，并不需要样本之间相互独立

如果确实需要产生独立同分布的样本，可以同时运行多条马尔科夫链，这样不同链上的样本是独立的；或者时一条马尔科夫链上每隔若干个样本选取一个，这样选取出来的样本也是近似独立的。

不均衡样本集的采样

基于数据的方法

对数据进行重采样，使原本不均衡的样本变得均衡

对样本集进行随机采样

过采样Over Sampling
- 随机从少数类样本集中重复抽取样本(有放回)以得到更多的样本
- 对少数类样本进行了多次复制，扩大了数据规模，增加了模型的复杂度，导致过拟合
- 不是简单复制样本，而是采用一些方法生成新的样本，可以降低过拟合的风险
  - SMOTE
    - 为每个少数类的样本合成相同数量的新样本，这可能会增加类间重叠度，并且会生成一些不能提供有益信息的样本
  - Borderline-SMOTE
    - 只给那些分类边界上的少数类样本合成新样本
  - ADASYN
    - ADASYN则给不同的少数类样本合成不同个数的新样本

欠采样Under Sampling
- 随机从多数类样本集中选取较少的样本，有放回或无放回
- 会丢弃一些样本，可能会损失部分有用信息，造成模型只学到了整体模式的一部分，导致欠拟合

Informed Undersampling
- NearMiss
- Easy Ensemble
- Balance Cascade
- One-sided Selection

基于算法的方法
- 将问题转化为单类学习One-class learning、异常检测Anomaly detection
- 改变模型训练时的目标函数，如代价敏感学习中不同类别有不同的权重来矫正这种不平衡性

贝叶斯网络的采样

概率图模型描述多个随机变量的联合概率分布，又称信念网络或有向无环图模型

贝叶斯网络是利用有向无环图来刻画一组随机变量之间的条件概率分布关系

- 祖先采样
- 逻辑采样
- 似然加权采样
- MCMC采样法

高斯分布的采样

逆变换法
- Box-Muller
  - 需要计算三角函数
- Marsaglia Polar Method

拒绝采样法

2018.12.20 Thursday. By KuKuXia@github.com

ML/DL/RL深度交流讨论群: 622545311，加群时请备注: 姓名-方向-公司or高校