第八章 采样

1.采样的作用

采样在机器学习中的应用

采样本质上是对随机现象的模拟，根据给定的概率分布，来模拟产生一个对应的随机事件。另一方面，采样得到的样本集也可以看作是一种非参数模型，即用较少量的样本点（经验分布）来近似总体分布，并刻画总体分布中的不确定性。从这个角度说，采样其实也是一种信息降维，可以起到简化问题的作用。对当前的数据集进行重采样，可以充分利用已有数据集。通过对样本多次重采样来估计统计量的偏差、方差等。另外，利用重采样技术，可以在保持特定的信息下（目标信息不丢失），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，例如利用重采样来处理分类模型的训练样本不均衡问题。

2.均匀分布随机数

一般可采用线性同余法来生成离散均匀分布伪随机数，但线性同余法生成的随机数并不是互相独立的，gcc中采用的glibc可以产生更好地均匀分布随机数。

3.常见的采样方法

对于一个随机变量，通常用概率密度函数来刻画该变量的概率分布特性。具体来说，给定随机变量的一个取值，可以根据概率密度函数来计算该值对应的概率（密度）。反过来，也可以根据概率密度函数提供的概率分布信息来生成随机变量的一个取值，这就是采样。从某种意义上来讲，采样是概率密度函数的逆向应用。

常见的采样方法：

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。

（1）均匀分布随机数

（2）拒绝采样

（3）重要性采样

（4）马尔科夫蒙特马洛采样法

（5）吉布斯采样法

4.高斯分布的采样

Box-Muller算法，Marsaglia polarmethod

5.马尔科夫蒙特卡洛采样阀

使用MCMC采样思想，包含两个MC（Monte Carlo, Markov Chain）。蒙特卡洛法是指基于采样的数值型近似求解方法，而马尔科夫链则用于进行采样。
MCMC采样基本思想是：针对带采样的目标分布，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。
常见的MCMC采样法：
（1）Metropolis-Hastings采样法
（2）吉布斯采样法

6.贝叶斯网络的采样

祖先采样法：核心思想是根据有向图的顺序，先对祖先节点进行采样，只有当某个节点的所有父节点都已完成采样，才对该节点进行采样。

7.不均衡样本集的重采样

最简单的处理方法是随机采样，一般又分为过采样（Over-sampling）和欠采样（Under-Sampling）。随机过采样是从少数类样本集中随机重复抽取（有放回）以得到更多样本；随机欠采样则相反，从多数类样本集中随机选取较少的样本（有放回或无放回）。
直接的随机采样可以使样本集变得均衡，但会带来一些问题：
（1）过采样对少数类样本进行多次采样，扩大了数据规模，增加了模型训练的复杂度，也容易造成过拟合。
（2）欠采样会丢弃一些样本，可能会损失部分有用信息，造成模型只学到了整体模型的一部分。
为了解决上述问题，通常在采样时不是简单地复制样本，而是采用一些方法生成新的样本。

在过采样方法方面，比如SMOTE算法对少数类样本集中每个样本x，从它在少数样本集中的K近邻中随机选取一个样本y，然后在x, y连线上随机选取一点作为新合成的样本。这种合成新样本的过采样方法可以降低过拟合的风险。也有一些改进的算法如Borderline-SMOTE,ADASYN等。Borderline-SMOTE只给那些处在分类边界上的少数样本合成新样本。ADASYN则给不同的少数类样本合成不同个数的新样本。其次还有数据清理方法（如基于Tomek Links）来进一步降低合成样本带来的类间重叠，以得到更加良定义的类簇，从而更好地训练分类器。

在欠采样方法方面，可以采用Informed Undersampling来解决由于随机欠采样造成的数据丢失问题。常见的Informed Undersampling算法有：

（1）Easy Ensemble算法（每次从多数类上随机选取一个子集E，E的大小与少数类样本集相似，然后用E+少数类样本集训练一个分类器，重复上述过程若干次，得到多个分类器，最终的分类结果是这么多个分类器结果的融合）

（2）Balance Cascade算法

基于算法的方法：可以通过改变模型训练时的目标函数（如代价函数敏感学习中不同类别有不同的权重）来矫正这种不平衡性；当样本数目极其不均衡时，也可以将问题转化为单类学习、异常检测。