# 采样的作用

**场景描述**

采样是从特定的概率分布中抽取对应的样本点。

## 问题： 举例说明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。采样可以让人对随机事件及其产生过程有更直观的认识。例如，通过对二项分布的采样，可以模拟“抛硬币出现正面还是反面”这个随机事件，从而模拟产生一个多次抛硬币出现的结果序列；

采样得到的样本也可以看做是一个非参数模型，即用少量的样本点（经验分布）来**近似总体分布**，并刻画总体分布中的不确定性。从这个角度来说，采样其实是一种对信息的降维，可以引起简化问题的作用。比如机器学习中的训练集和测试集，这种**信息降维的特性**，使得采样在数据可视化方面也有很多应用，可以帮助人们快速、直观了解总体分布中数据的结构和特性。

# 均匀分布随机数

## 场景描述

均匀分布是指整体样本空间中的每一个样本点对应的概率密度都是相等的。根据空间是否连续，又分为离散均匀分布和连续均匀分布。

均匀分布可以算作是最简单的概率分布。从均匀分布中进行采样，即生成均匀分布随机数，几乎是所有采样算法都需要用到的基本操作.

## 问题：如何编程实现均匀分布随机数生成器

计算机程序都是确定性的，因此并不能产生真正意义上的完全均匀分布随机数，只能产生**伪随机数**（伪随机数是指这样数字虽然是通过确定性程序产生的，但是他们能通过近似的随机性测试）。

一般可以采用线性同余法：；

# 常见的采样方法

## 场景描述

* 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，给定随机变量一个取值，可以根据概率密度函数来计算该值对应的概率。
* 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。

## 问题：抛开那些针对特定分布而精心设计的采样方法，说一些你所知道的通用采样方法或者采样策略，简单描述他们主要思想以及具体操作步骤。

* 几乎所有的采样都是以均匀分布随机数作为基本操作。均匀分布随机数一般用线性同于法来产生。
* 首先假设已经有[0,1]均值分布随机数。对于一些简单分布，可以直接用均匀采样的一些扩展方法来产生样本点，比如有限离散分布可以用轮盘赌算法来采样。然而，很多分布一般不好直接进行采样，可以考虑函数变换法。一般x和u存在变换，则他们的概率密度函数如下关系：



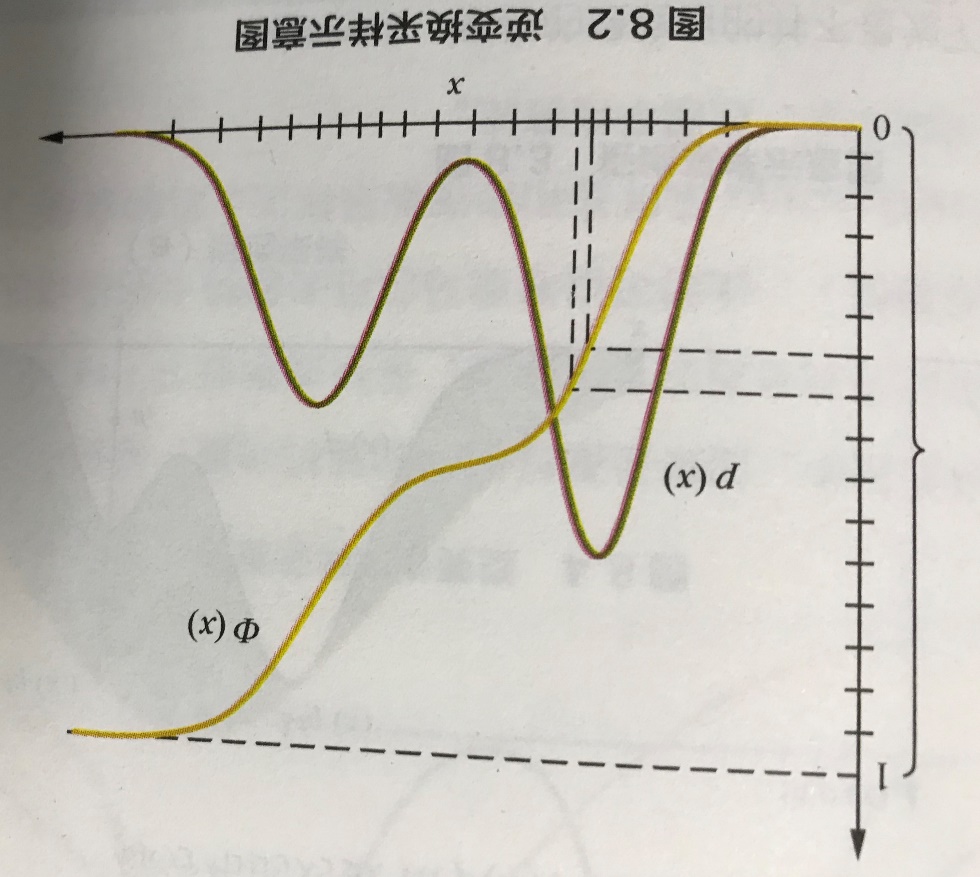


F(x)为分布，f(x)为概率目的,p为概率

* 如果从目标分布p(x)中不好采样x，可以构造一个变换，使得从变换后的分布p(u)中采样u比较容易，这样可以通过中间接得到x。
* 特别地，在函数变换中，如果变换关系是x的累积分布函数的化，则得到所谓的逆变换采样，假设待采样的目标分布的概率密度函数为p(x),则有：

，则逆变换按如下过程采样：

1. 从均匀分布U(0,1)产生一个随机数;
2. 计算 ，p(x)概率密度，表示分布，纵轴是u(均匀分布）



* 如果待采样的目标分布的累积分布函数的逆函数无法求解或者不容故意计算，则不适用于逆变换采样法，此时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后处理操作，使得最终的样本服从目标分布。
  + 拒绝采样：对目标分布p(x)，选择一个容易的参考分布q(x)，使得对于任意x都有

,

1. 参考分布q(x)中采样一个样本；
2. 从均匀分布U(0,1)产生一个随机数;
3. 如果，则接受；否则拒绝，重新（1）（2）（3）

# 高斯分布的采样

## 问题：如何对高斯分布进行采样

首先，假设随机变量z服从标准正态分布，N(0,1)，令，

则x服从均值为方差为的高斯分布；因此，任意高斯分布都可以从标准正态分布通过拉伸和平移得到；

### 步骤：

### 产生（0,1）上的均匀分布随机数u；

### 令，则z服从标准正态分布，其中是高斯误差函数，它是标准正态分布的累积分布函数经过简单平移和拉伸变换后的形式，

### 上述逆变换法

# 贝叶斯网络的采样

## 场景描述

概率图模型经常被用来描述多个随机变量的联合概率分布。贝叶斯网络，又称为信念网络或者有向无环图模型。它是一种概率图模型，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。

## 问题：如何对贝叶斯网络进行采样？如果只需要考虑一部分变量的边缘分布，如何采样？如果网络中含有观测变量，又该如何采样？

* 对一个没有观测变量的贝叶斯网络进行采样，最简单的方法是祖先采样，它的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完成采样，才对该节点进行采样。
* 含有观测变量的贝叶斯网络的采样，最直接的方法是逻辑采样，还是利用祖先采样得到所有变量的取值，如果这个样本在观测变量上的采样值与实际观测值相同，则结构，否则拒绝。

# 不均衡样本集的重采样

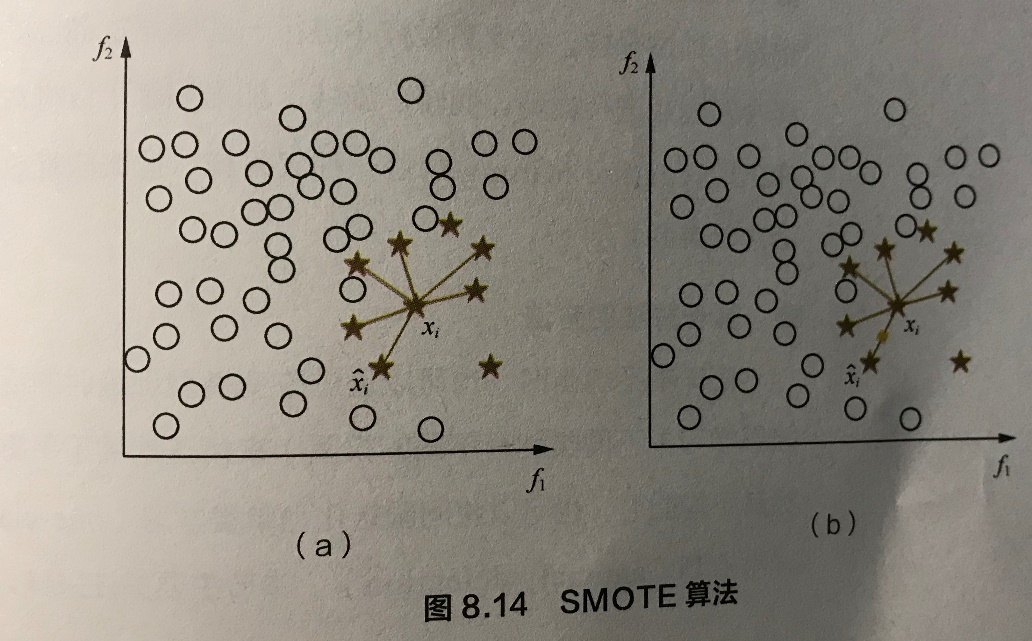
## 问题：对于二分类问题，当训练集中正负样本非常不均衡时，如何处理数据以更好地训练分类模型

* 模型在训练时优化目标函数和人们在测试时使用的评价标准不一致的化就会发生问题；这种不一致可能是由于数据分布不一致，

### 基于数据的方法

对于数据进行重采样，使原来不均衡的样本变得均衡。首先，记样本数大的类别，类别数小的，样本集。

* 最简单的处理不均衡样本集的方法是**随机采样**。采样一般分为过采样和欠采样。随机过采样是从少数类样本集中随机重复抽取样本（有放回）以得到更多样本。
  + 过采样对于少数类样本进行多次复制，扩大了数据规模，增加了模型训练的复杂度，同时也容易造成过拟合。
  + 欠采样会丢弃一些样本，可能会损失部分有用的信息，造成模型只学到了整体模式的一部分。
* **SMOTE方法：** 对少数类样本集中每个样本x，从它在的K近邻中随机选择一个样本y，然后在x,y的连线上选取一个点作为新合成的样本，

****

* + 为每个少数类样本合成相同数量的新样本，这可能会增大类间重叠度，并且会生成一些不能提供有益信息的样本。
* 对于**欠采样**：

1. **easy ensemble**算法：每次从多数类中随机抽取一个子集E（），然后用训练一个分类器，重复上述过程若干次，得到多个分类器，最终的分类器是多个分类器结果的融合。
2. Balance Cascade:级联结构，在每一级从多数类中随机抽取一个子集E，然后用训练一个分类器；然后将中能够被当做分类器正确判断的样本剔除，继续下一个级的操作，重复若干次得到级联结构，最终的输出结果也是个级联分类器结果的融合。

### 2． 基于算法的方法

### 在样本不均衡时，也可以通过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，当样本数目及其不均衡时，也可以将问题转化为单类学习，异常检测。