《统计学习方法》第四章读书笔记

一、自己提出的问题的理解（罗列全部）：

1. （讨论后补充）属性独立性假设是一个较强的假设，在实际情况中很难成立，但为什么朴素贝叶斯仍能取得较好的效果？

**讨论后的理解**：在使用分类器之前，首先做的第一步（也是最重要的一步）往往是特征选择，这个过程的目的就是为了排除特征之间的共线性、选择相对较为独立的特征。

对于分类任务来说，只要各类别的条件概率排序正确，无需精准概率值就可以导致正确分类；如果属性间依赖对所有类别影响相同，或依赖关系的影响能相互抵消，则属性条件独立性假设在降低计算复杂度的同时不会对性能产生负面影响。

二、别人提出的问题的理解（选择几个问题罗列，并给出理解）：

1. 极大似然估计和贝叶斯估计的区别是什么？

**个人的理解：贝叶斯估计是极大似然估计“升级版”**

在实际的模型训练过程中，采用极大似然估计可能会出现零概率问题（因为先验概率和反条件概率是根据训练样本算的，但训练样本数量不是无限的，所以可能出现有的情况在实际中存在，但在训练样本中没有，导致为0的概率值，影响后面后验概率的计算）。

即便可以继续增加训练数据量，但对于有些问题来说，数据怎么增多也是不够的。此时的模型是不平滑的，要使之平滑，一种方法就是将训练（学习）的方法换成贝叶斯估计。对分子加，分母加上类别数，这样在解决零概率问题的同时，也保证了概率和依然为1。

1. 在属性之间相关性较大的情况下，条件独立性假设可能会大大增加误差率，我们该如何解决？

**个人的理解：**首先，朴素贝叶斯模型名称中“朴素”的含义正是假设属性之间相互独立，这个假设在实际应用中往往是不成立的，个人认为这种较强的假设可以算是这个模型的先天缺陷。

但正如问题1所述，使用分类器之前要做的第一步就是特征选择，这个过程的目的就是为了排除特征之间的共线性、选择相对较为独立的特征。而在属性相关性较小时，朴素贝叶斯性能最为良好。

若应用场景属性个数较多或者属性之间相关性较大时，朴素贝叶斯分类效果不佳。对于这一点，可采用半朴素贝叶斯之类的算法，通过考虑部分关联性适度改进。

三、读书计划

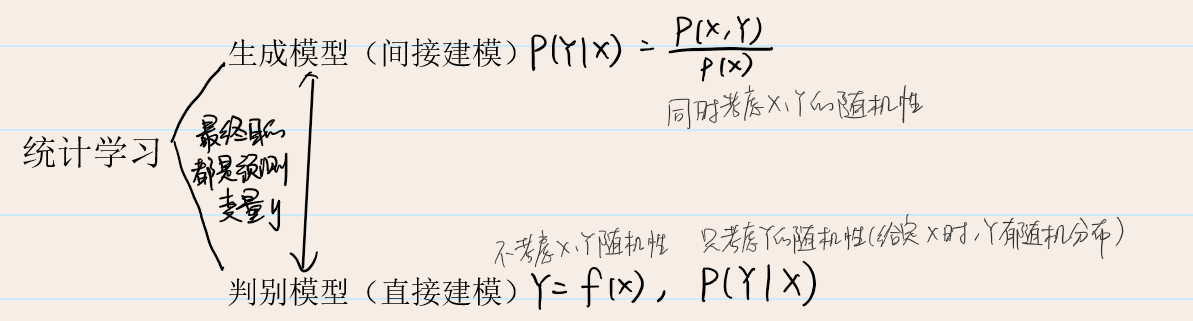
1. 本周完成的内容章节：第四章
2. 下周计划：第五章

四、读书摘要及理解或代码的具体实现

1. 读书摘要及理解

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入输出的联合概率分布；然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。

朴素贝叶斯法仍属于监督学习范畴，但与之前不同的是，它基于一种全新的建模理论：搭建模型的出发点不是为了对未知数据做预测，而是为了弄清楚并模拟数据产生的原理与机制。



**先验概率：**

**条件概率：**

**朴素贝叶斯算法的工作流程**

**输入：**训练数据集

其中,表示第i个样本的第j个特征，,表示第j个特征可能取的第l个值，j=1,2,…,n，l=1,2,…,，

**输出：**实例x的分类

1、计算先验概率及其条件概率

2、对于给定的实例计算

3、确定实例x的类

1. 代码的具体实现

调用Weka api实现了朴素贝叶斯分类，数据集为鸢尾花iris.arff，测试得到分类的正确率0.96。