1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：朴素贝叶斯中每个属性对于类别决策都有相同的影响，不符合实际i情况，怎么解决这个问题？

讨论后的理解：可以对各特征属性赋予一个权值，通过属性的重要性来确定权重；

1. 提出的问题2：对于朴素贝叶斯有哪些可以改进的方面？

讨论后的理解：

1. 结构扩展（主要方向为探索属性间的关联关系）：

朴素贝叶斯是星型结构， Friedman在1997年提出的TAN算法，将原算法的星型结构演变成更高的树型依赖结构，目的是解决属性变量之间存在的依赖关系。2004年石洪波提出的利用条件互信息选择若干TAN进行组合，提高TAN算法；Cooper在1992年提出了K2算法，在已获得先验信息下，利用爬山法来搜索最优的结构；

2）局部学习（将原数据集切割成若干子集，使用其中某些来进行分类算法）

局部学习的原理是将原本不符合朴素贝叶斯假设的训练集中的部分数据能满足属性间没有依赖性的要求。

3）属性选择（将冗余属性剔除删除的方法）

Langley在1994年提出的贪心策略对于属性进行选择，以此提高分类算法的性能。

也有很多学者使用粗糙集理论，寻找能够度量单个属性特征对类别的决策贡献度指标，即关联规则，并且基于此来剔除贡献度第的属性特征。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：Navie Bayes和Logistic回归区别是什么？

讨论后的理解：前者是生成式模型，后者是判别式模型，二者的区别就是生成式模型与判别式模型的区别。

1）首先，Navie Bayes通过已知样本求得先验概率P(Y), 及条件概率P(X|Y), 对于给定的实例，计算联合概率，进而求出后验概率。也就是说，它尝试去找到底这个数据是怎么生成的（产生的），然后再进行分类。哪个类别最有可能产生这个信号，就属于那个类别。

优点：样本容量增加时，收敛更快；隐变量存在时也可适用。

缺点：时间长；需要样本多；浪费计算资源

2）相比之下，Logistic回归不关心样本中类别的比例及类别下出现特征的概率，它直接给出预测模型的式子。设每个特征都有一个权重，训练样本数据更新权重w，得出最终表达式。梯度法。

优点：直接预测往往准确率更高；简化问题；可以反应数据的分布情况，类别的差异特征；适用于较多类别的识别。

缺点：收敛慢；不适用于有隐变量的情况。

1. 问题4：为什么说朴素贝叶斯是稳定的算法？

讨论后的理解：因为朴素贝叶斯是基于古典数学理论，使用已经获得的数据的先验概率，通过先验概率来计算出后验概率，以此对一条数据进行预测，得到需要的与该数据所对应的类别，因为先验概率是固定的，计算的方法和路径的都是固定的，因此每次计算所得的后验概率都是相同的，所以这是一种稳定的方法‘

1. 问题5：多项式模型和伯努利模型在实现上有什么区别，如何评价优劣?

讨论后的理解：伯努利模型并不考虑词在文档中出现的次数，只考虑出不出现，在这个意义上相当于假设词是等权重的。多项式模型考虑此在文档中的出现次数。二者的计算粒度不一样，多项式模型以单词为粒度，伯努利模型以文件为粒度。当训练集文本较短时，我们更倾向于使用贝努利模型。而文本较长时，我们更倾向于多项式模型，因为，在一篇文档中的高频词，会使该词的似然概率值相对较大。这两种方法是用来处理离散数据的，对于连续型数据采用高斯模型。

1. （必填）读书计划

1、本周完成的内容章节：《统计学习方法》第三章

2、下周计划：《统计学习方法》第四章