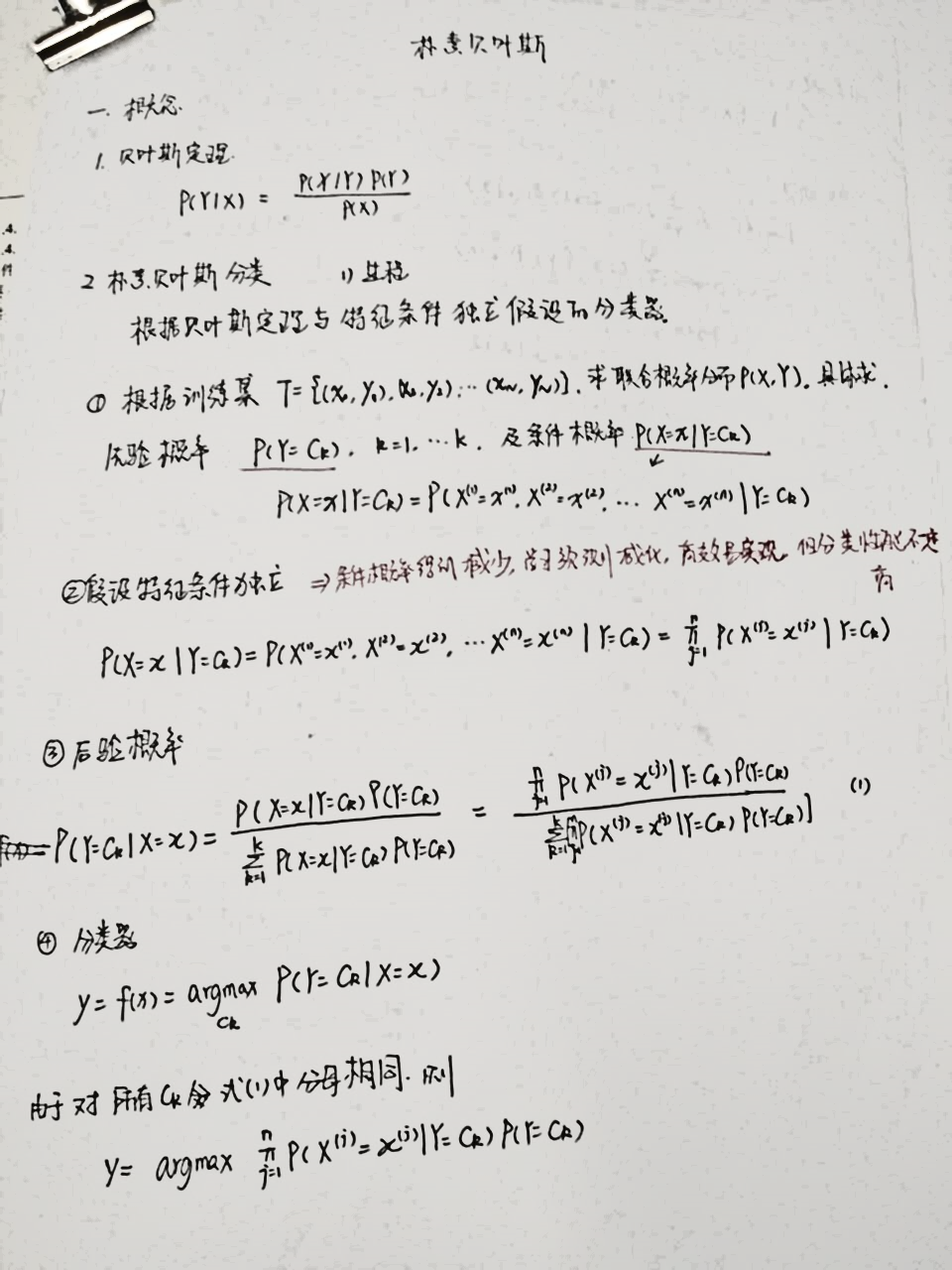
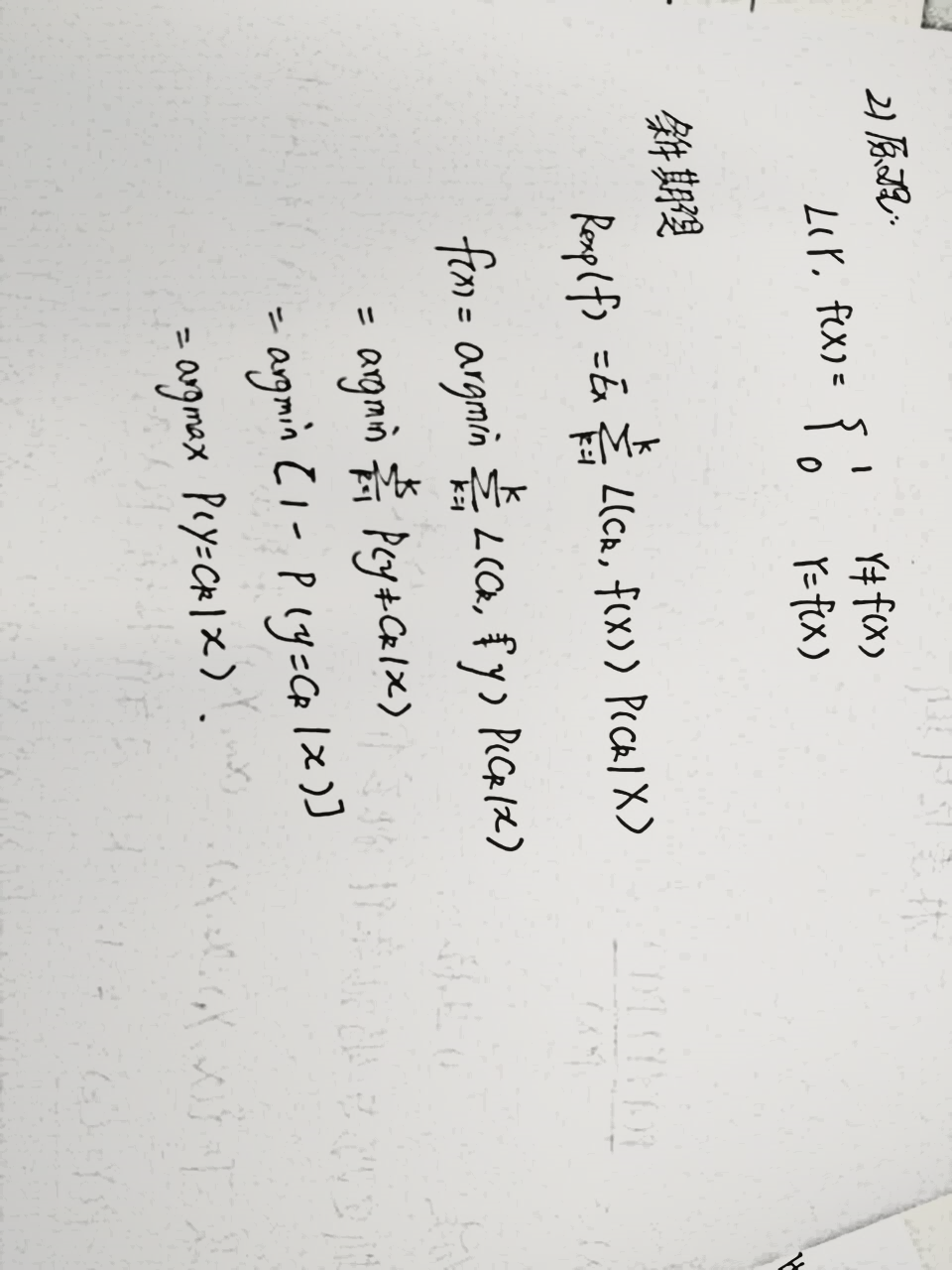
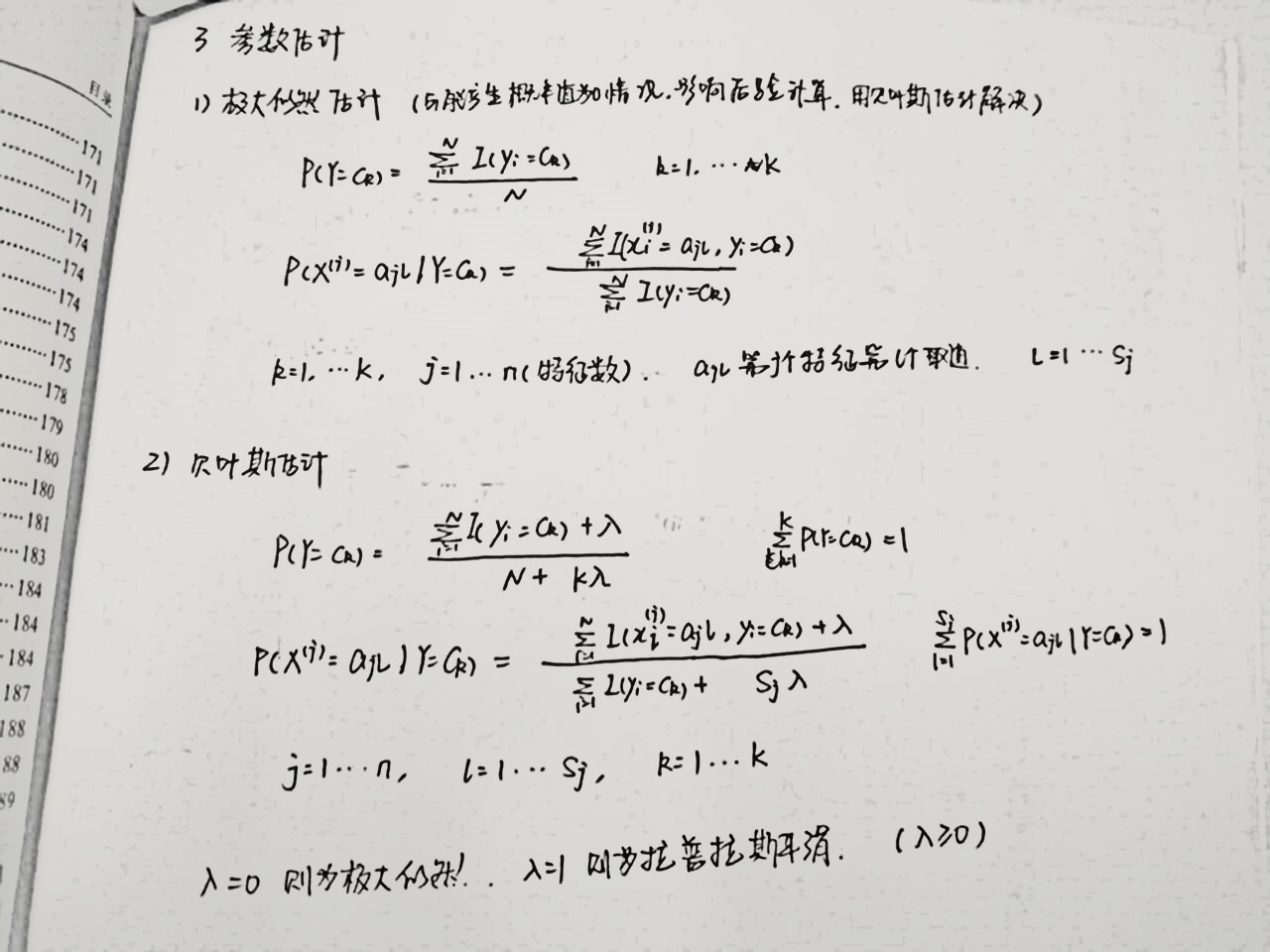
朴素贝叶斯







采用贝叶斯估计方法和极大似然估计方法得到的结果，概念在二.6中。、

《统计机器学习方法》李航

**二、问题与答案**

1. **说一下贝叶斯原理，贝叶斯分类过程和贝叶斯局限性。**

原理、分类过程见上。

**局限性：**

先验分布不同，使后验分布存在差异。（参考贝叶斯估计和朴素贝叶斯分类共同的局限性）

贝叶斯估计局限性：<https://wenku.baidu.com/view/cfcd21efa2161479161128f5.html>

1. **为什么朴素贝叶斯如此“朴素”？**

贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础。朴素贝叶斯分类是贝叶斯分类中最简单的一种分类方法。朴素贝叶斯分类器（Naive Bayes classifier）的朴素（Naive）之处在于，其假设了各个特征之间是独立的，使朴素贝叶斯变得简单，但会损失分类准确性。

1. **naive bayes和logistic regression的区别**

(1) Naive Bayes生成模型。先从训练数据中先计算P(x|y)和P(y)，再计算P(y|x)。

Logistic Regression判别模型，估计参数W，在数据集上最大化判别函数P(y|x)。

(2) Naive Bayes建立在条件独立假设基础之上。

LR限制宽松很多。数据满足条件独立假设，LR能取得好效果；不满足条件独立假设时，LR仍能通过调整参数让模型最大化符合数据的分布，从而得到最优模型。

(3) 数据集较小时，应选Naive Bayes。 Naive Bayes用了严格的条件独立假设，为了计算P(y|x)，需用统计的方法求得P(x|y)和P(y)。所需的数据量要小些，为O(log n)。

数据集较大时，应选Logistic Regression。计算时，在整个参数空间进行线性搜索，需要的数据集就更大，为O( n)。n为数据维度。

<http://blog.csdn.net/myue5/article/details/19409615>

<http://blog.csdn.net/cjneo/article/details/45167223>

**生成模型、判别模型：**

生成方法，产生样本模型（学习数据如何产生的，可以反映数据相关性，即联合概率分布），再产生预测模型（即条件概率分布）进行分类。主要模型有朴素贝叶斯和隐马尔科夫模型。

判别方法，直接产生预测模型（学习最优分类面，学习数据差异不学习数据间关系）。模型包括k近邻，感知级，决策树，支持向量机、逻辑回归等。

<http://blog.csdn.net/zouxy09/article/details/8195017>

1. **谈谈朴素贝叶斯的优缺点？**

**优点：**

1）朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

2）对小规模的数据表现很好，能处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。

3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。

**缺点：**

1） 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型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，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

2）需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

3）由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

4）对输入数据的表达形式很敏感。

<http://www.sohu.com/a/128627325_464088>

1. **解释贝叶斯公式和朴素贝叶斯分类。**

见上。

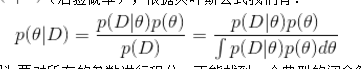
1. **什么是贝叶斯估计。**

参数估计方法。

最大似然估计(MLE)和最大后验估计(MAP)认为待估计参数是固定值的未知量。都是找使函数最大时的参数； MAP比MLE考虑先验分布，所以MAP会有超参数，其超参数代表一种信念(belief)，会影响推断(inference)结果。

贝叶斯派认为待估计的参数http://img.blog.csdn.net/20141110214603237是随机变量，服从一定的分布。样本X 是固定的，重点研究的是参数http://img.blog.csdn.net/20141110214603237的分布。让参数以某种概率密度函数分布（参数的先验），导致计算过程高复杂。

贝叶斯估计与MAP：

贝叶斯估计所有后验概率，积分难算。

MAP找使上述后验概率最大的参数。（参考网址2、3）

<http://blog.csdn.net/liu1194397014/article/details/52766760>

<http://blog.csdn.net/andyelvis/article/details/42423185>

<http://blog.163.com/silence_ellen/blog/static/1761042222014413112444364/>