1. （必填）自己提出的问题的理解（罗列全部）：
2. **提出的问题1：**

P128提到特征选择算法用来去除冗余的特征，其中的逻辑是什么？

讨论后的理解：

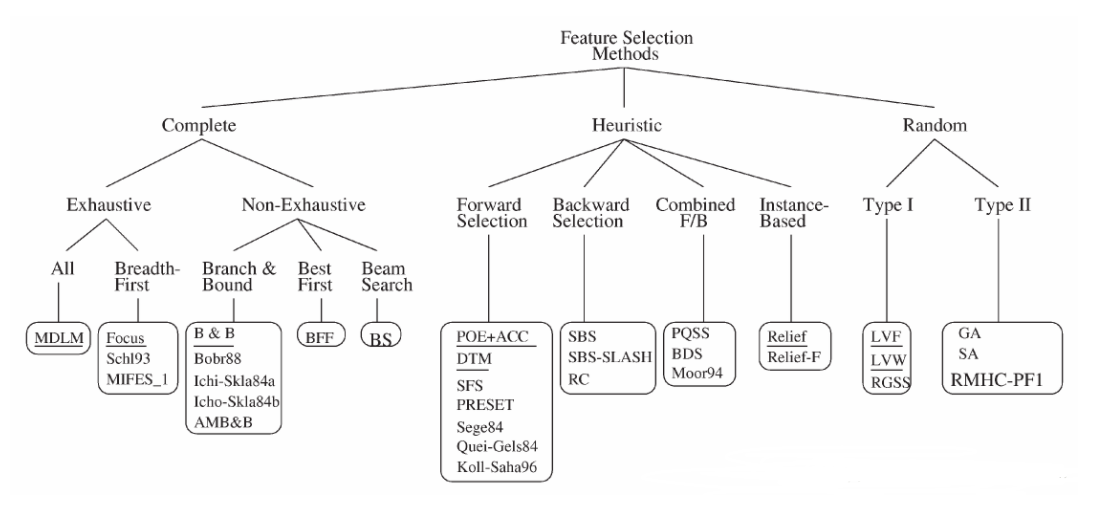
一个对象一般具有多种特征，特征一般可以分为三种类型：

相关特征：对于学习任务（例如分类问题）有帮助，可以提升学习算法的效果；

无关特征：对于我们的算法没有任何帮助，不会给算法的效果带来任何提升；

冗余特征：不会对我们的算法带来新的信息，或者这种特征的信息可以由其他的特征推断出；

对于一个特定的学习算法来说，哪一个特征是有效的是未知的。因此，需要从所有特征中选择出对于学习算法有益的相关特征。而且在实际应用中，经常会出现维度灾难问题，尤其是在文本处理中。例如，可以把一篇文档表示成一个词向量，但是往往会使用所有的单词作为字典，因此对于一篇可能仅仅包含100或者200个单词的文档，可能需要上万的维度（也就是特征）。如果可以从中选择一部分相关特征构建模型，这个问题就可以得到一定程度的解决。所以，特征选择和降维有一定的相似之处。另外，从上面的例子中可以发现，如果只选择所有特征中的部分特征构建模型，那么可以大大减少学习算法的运行时间，也可以增加模型的可解释性。所以特征选择算法的主要目的包括降维、降低学习任务的难度和提升模型的效率。特征选择算法主要分为如下三类：



1. **提出的问题2：**

P130倒数第二段，无标注数据的惩罚值从一个较小的值迭代增加至合适值，其增大了找到一个好的局部最优解的可能性并使得在惩罚值很小的情况下能更加容易的改进代价函数。为什么这样会是改进代价函数变得更加容易呢？

讨论后的理解：

当标记点不均匀时，相对较多的那一类目标函数中乘上的常数较小，每个点越界的惩罚就较小，场平面就会相对往点多的那一类的方向移动。对正例和反例数据的软件间隙误差采用不同的惩罚值，可以使得代价函数向着局部最优方向移动，即可以使得在惩罚值相对较小的情况下更容易改进代价函数。

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**：

co-Training算法同时使用两个分类器的意义是什么？

自己的理解：

首先分别在每个视图上利用有标记样本训练一个分类器；然后,每个分类器从未标记样本中挑选若干标记置信度(即对样本赋予正确标记的置信度)高的样本进行标记,并把这些“伪标记”样本(即其标记是由学习器给出的)加入另一个分类器的训练集中,以便对方利用这些新增的有标记样本进行更新。这个“互相学习、共同进步”的过程不断迭代进行下去,直到两个分类器都不再发生变化,或达到预先设定的学习轮数为止。分类器间的分歧程度是协同训练法泛化错误率的上界。

1. **问题4：**

自学习算法因为自己训练自己，会不会出现严重的过拟合结果？

自己的理解：

会。过拟合问题是指在模型参数你和过程中，由于训练数据包含抽样误差，训练时复杂的模型将抽样误差也考虑在内，将抽样误差进行了很好地拟合，使得训练集上呈现很好的效果，但是在测试集上效果很差。一般可用通过增加数据量、调整模型来防止过拟合。

1. **问题5：**

EM算法能够保证必定收敛嘛？

要证明算法的收敛性，可以证明对数似然函数的值在迭代过程中一直在增大。即：

由条件概率性质可知：

由于：

令：

上两式相减得：

将分别取和代入上式，并相减得：

要证明EM算法具有收敛性，只需要证明上式右边是非负即可。

由于使得极大，所以恒成立

而：

因此：

恒成立。所以，EM算法具有收敛性，但不能保证收敛到全局的极大值点。

1. （必填）读书计划

1、本周完成的内容章节：第五章（5.1）

2、下周计划：统计学习方法（第一章）

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

摘要：本周完成了半监督学习的从已标注数据和无标注数据中学习的部分，包括EM算法、Co-Training和自学习算法。

**EM算法**（全称为最大化期望算法）是一种用于在含有不完全数据的概率模型中，通过不断迭代求解参数的最大似然估计（Maximum Likelihood Estimate, MLE）或最大后验估计（Maximum A Posteriori, MAP）的一种算法。不完全数据在算法实现中一般可以解释为三情况。第一，存在真实的观测数据缺失；第二，在求解最大似然估计很难的情况下，如果人为的添加变量可以使求解最大似然估计变得更容易，则原始数据也可以看成是人为的不完全数据；第三，含有隐变量（latent variable）的概率模型求解最大似然估计或最大后验估计问题。

EM 算法基于 Bayes 算法的一种计算方法——数据添加算法，即通过在观测数据（observed data）的基础上添加一些“潜在数据”，使极大化或模拟操作变得更加简单，“潜在数据”可以是缺失数据或未知参数。其原理如下：设可观测数据为 X，而𝜃关于 X 的验后分布𝑝(𝜃|𝑋)难以进行各种统计计算，于是我们假定一些观测值 Z 已知，并且𝜃关于 X 和 Z 的验后分布𝑝(𝜃|𝑋, 𝑍)相对更易于做极大化、抽样等，然后再对 Z 做检查和改进。

**Co-Training**也是一种从无标注数据和已标注数据中学习的方法。算法假设在不同的目标函数下数据的类别分布式想容的，且在给定数据标识的情况下，一个集合中的特征与另一个集合的特征是条件独立的。算法利用属性的分割来从已知的已标注数据和无标注数据进行迭代学习。其主要思想是，首先从已标注数据集L的每个特征集中学习得到一个分类器，然后利用这个分类器去分类U中的无标注数据。

**自学习**是利用自己的预测结果来训练自己。它是一种不需要特征分割的增量式算法。算法开始时利用少量已标注数据的全部特征训练得到一个分类器，然后利用此分类器去分类无标注数据，再将其中可信度最高的数据及其预测结果与已标注数据一起作为下一轮分类器训练的训练集。

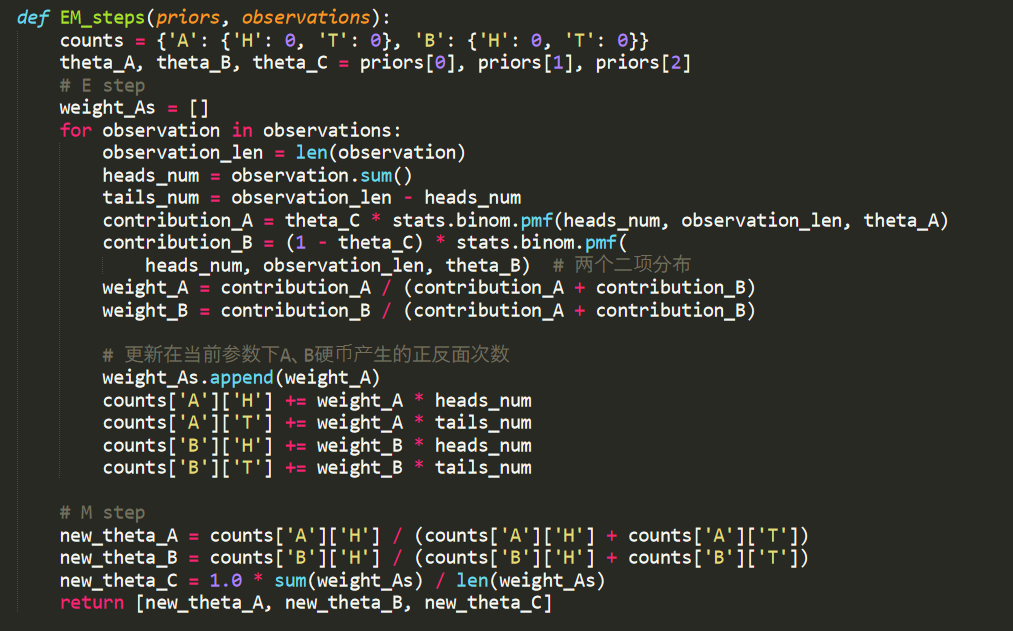
**直推式向量机**是利用无标注数据，为无标注数据选择类别标识使得得到的分类器具有最大的间隙，其中以标注那些已知的无标注数据为目的的训练过程称为直推式。直推式向量机的假设为，数据集满足最大间隙假设。当数据集满足最大间隙假设时，它可能会改进结果，但是当它不满足这个假设时，效果则会下降到和直接使用监督学习方法一样。

**基于图的方法**：可以看成是最近邻有监督学习算法的同时使用有标注和无标注数据的一种拓展。算法的基本思想是把已标注数据和无标注数据都看成是一张图中的顶点，并且使用某种相似度函数来定义各个顶点之间的边的权重。图中边的权重体现了两个顶点之间相似度的大小，权重越大，对应的相似度高。在书中，提到了三种基于图的三种LU学习方法：最小割、高斯厂和谱图直推器。

2、伪代码的具体实现(选做)

实现了基于二项分布的EM算法：

算法准备：



算法实现：

