第一次作业

作业在qq群中的作业中提交电子版，可以是word、pdf、或图像文件（手写后拍照），注意在作业上写出姓名和学号。

姓名：乔翱

学号：202222080625

1. 什么是机器学习？请查找资料给出一个你觉得满意的定义或解释。

以性能度量P进行衡量，如果一个计算机程序在某类任务T上的性能，随着经验E而提升，那么我们称这个计算机程序从经验E中学习。

1. 机器学习按任务可以分为哪些主要类型？并分别描述其定义。

监督学习：通过让机器学习大量带有标签的样本数据，训练出一个模型，并使该模型可以根据输入预测相应输出的过程。

无监督学习：直接对输入数据进行建模，提取数据有效信息，探索数据的整体结构。

强化学习：通过一个智能体在与复杂而不确定的环境交互中最大化总回报来学习的一种计算方法。

1. 请说明在机器学习任务中，训练集、验证集和测试集的作用。

训练集：用于训练模型，找出模型参数。

验证集：用于确定模型超参数，选出最优模型。

测试集：用于评估最终模型的泛化能力。

1. 什么是过学习？有哪些避免过学习的方法？

过拟合：模型在训练集上误差很低，但是在测试数据上误差很高。

过拟合解决方法：扩大训练集，正则化，通过验证集来评估选择模型。

1. 什么是交叉验证？它有哪些具体做法？

K折交叉验证：将数据集分层采样划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为验证集，最终返回k个测试结构的均值，k最常用的取值是10。

留一法：考虑K=N的情况很合适，其中N是数据点的总数，这种技术叫做留一法，不受随机样本划分方式的影响，结果往往比较准确，当数据集比较大时，计算开销难以忍受。

留一交叉验证：哪个模型E小选哪个。

1. 请查找资料给出一个正则化方法的文字定义。

正则化：通过修改目标函数来惩罚模型的复杂度。通过系数λ控制正则化项相对于平方和误差项的重要性。

1. 请给出三种决策方法。

生成式模型、判别式模型、判别函数

1. 简述极大似然估计和最大后验参数估计的区别和联系

极大似然估计：极大似然估计是在模型已知，参数未知的情况下，利用采样得到的数据，类似对现实中的一些数据进行人工记录下来，得到一组已知的观测数据，在假设的模型中训练，将使似然函数取得最大值时的参数θ，即为所取得的的参数。

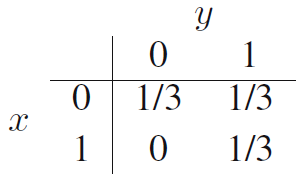
最大后验参数估计：最大后验概率则是在规则化极大似然估计，由于贝叶斯学派认为参数不能直接从观测数据中得到，因为参数具有许多的不确定性，不能因为仅有的数据就认为参数是由这些观测数据得到的，因此在这个前提下要加上一个先验概率。

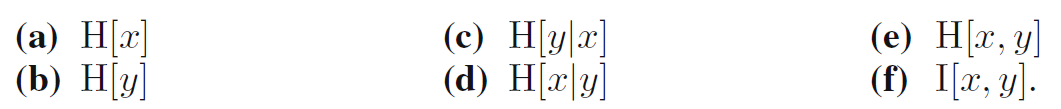
当先验概率是分布均匀的情况下，则相当于没有给参数提供任何有用的信息，例如每种情况都是等概率的事件，那次此时的极大似然估计就等于最大后验估计。因此，可以把极大似然估计看做一种特殊的先验概率为均匀分布的最大后验估计，也可以把最大后验估计看出是必须考虑先验概率的极大似然估计，最大后验估计是规则化的极大似然估计。

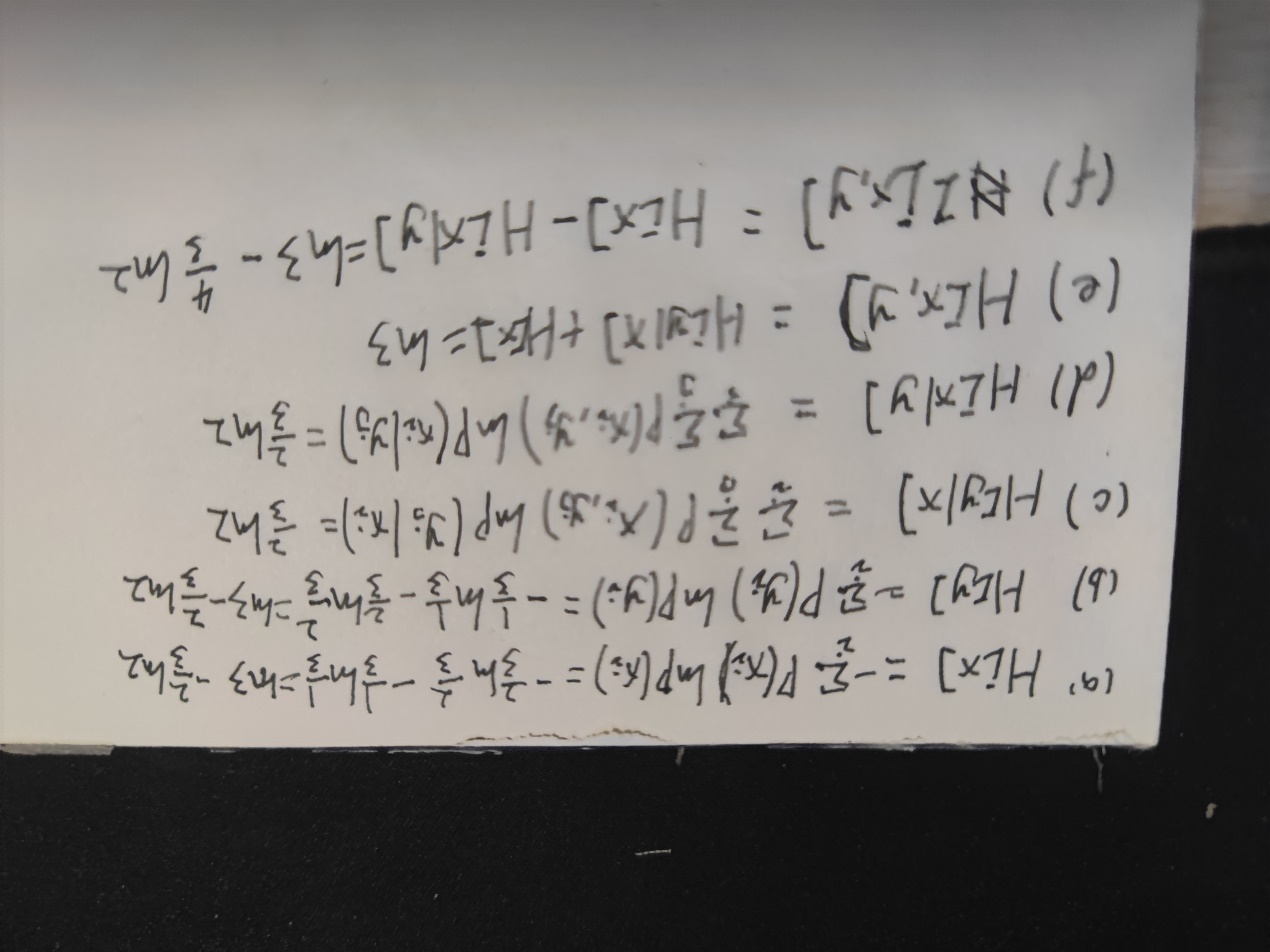
1. 一种癌症，得了这个癌症的人被检测出为阳性的几率为90%，未得这种癌症的人被检测出阴性的几率为90%，而人群中得这种癌症的几率为1%。一个人被检测出阳性，问这个人得癌症的几率为多少？

事件A表示测出为阳性，用事件B\_1表示得癌症,B\_2表示未得癌症。根据题目，P(A|B\_1)=0.9,P(A|B\_2)=0.1,P(B\_1)=0.01, P(B\_2)=0.99，则已知为阳性的情况下，得癌症的几率 P(B\_1|A) = 0.01\*0.9 = 0.009。

1. Bishop书第一章习题1.39（信息论）考虑两个变量x和y，每个变量只有两个可能的取值。它们的联合概率分布在表中给出。计算下⾯各式的值。







1. 什么是共轭先验，它有什么好处，试举出几种分布参数的共轭先验。

在贝叶斯统计理论中，如果某个随机变量θ的后验概率p(θ|x)与先验概率p(θ)属于同一个分布簇（同类分布），那么称p(θ|x)和p(θ)为共轭分布，这种性质叫做共轭性。同时，也称p(θ)为似然函数p(θ|x)的共轭先验。

共轭先验，对于某些似然函数，如果选择了某个先验，使得后验和先验具有相同的分布，则将该先验称为共轭先验。共轭先验的好处主要是在于代数上的方便性，可以直接给出后验分布的封闭形式，否则的话只能数值计算。共轭先验也有助于获得关于似然函数如何更新先验分布的直观印象。

正态分布-正态分布，二项分布-贝塔分布，多项式分布-狄利克雷分布，泊松分布-伽玛分布。

1. Bishop书第一章习题1.11（极大似然）



