数据挖掘 Homework 1

施俞晨 21210980116

1. 请列举实际中 1 个数据挖掘中有监督学习的应用，请包含以下部分:
(1) 背景介绍（请阅读材料：如何撰写背景介绍）
(2) 数据说明（说明因变量和自变量各是什么）
(3) 对以上特定问题如何进行模型评估
(4) 通过机器学习建模如何实现数据价值（无需列举具体的机器学习模型，但须说明模型如何落地，为谁服务产生价值）

1. 背景介绍

世界各地很多地方有盗猎现象，盗猎者通过一些陷阱、猎套等装置，等待动物上钩，然后杀死、卖掉动物的肉，或者更有价值的，例如象牙、犀牛角等动物制品来获利。这样的盗猎活动会影响到例如大象、老虎犀牛等特征动物，甚至可能进一步影响到整个生态系统，影响物种多样性。因此，反盗猎工作非常重要。

在保护区中，一般会有护林员巡逻来清理这些猎套。但是护林员时比较紧缺的，人力成本高，而猎套一般又都是比较便宜且容易制作的，可以到处放置，在盗猎者和护林员之间就存在着成本不平等的博弈问题。因此，希望能够通过机器学习的方式来帮助护林员制定一些更好的巡逻策略。

2. 数据说明

在本任务中，对于任务建模为：对于一定区域内的空间网格化划分，预测该区域内是否会发生盗猎现象，从而有针对性地进行巡逻检查，提升人工效率。因此，该任务可以建模为针对网络地点的回归问题，因变量为该地区发生盗猎行为的可能性，自变量为下述相关的特征。

研究者通过和乌干达当地政府的合作，可以提供当地 12 年的巡逻数据，从中可以提取一些特征。例如，在历史数据中可查找到在哪些区域曾经发生过到列现象。而相对动态的变量包括：往年的巡逻时间、在相邻区域的巡逻时间。还有地理空间相关的特征，包括：土地覆盖率、和河流的距离、和村庄的距离、和公路的距离、动物密度、海拔、坡度等。

例如，若一个地方离村庄太近，野生动物的量就可能比较少，因此对于盗猎者而言收益就比较低；而如果距离太远，则盗猎者跑到那个地方的时间和精力成本就会升高。因此，到村庄的距离是一个很容易会影响盗猎者决策的因素。

3. 模型与评估

在在历史数据中，一个难点在于，绝大多数的数据点上是没有找到猎套的，只有少量的数据点上被标记了一些套列行为。因此，在这一数据集中正样本是比较稀疏的，并且这一数据量无法支撑一个比较复杂的模型。

此外，这一数据也存在一定程度的不确定性：如果在一个地方发现了猎套，那么我们可以肯定该地存在盗猎现象；然而，在某些区域没有发生猎套，并不意味着该区域没有盗猎现象 —— 可能是护林员没有到该地进行巡逻或巡逻次数比较少，也有可能该地区的环境特殊性导致较难发现猎套。

通过对于区域进行网格化划分，将历史发生盗猎次数作为预测目标 Y，可以利用上面收集到的特征训练机器学习模型，从而进行回归预测。

机器学习模型可以学习到盗猎行为与相关特征之间的关系，从而预测可能发生盗猎行为的地点的分布模式。若模型学习的效果较好，可以对于护林员平常巡逻较少的区域进行精准预测。

为了验证模型，研究人员挑选出了模型预测出来盗猎风险较高的，但是在往常的巡逻极少设计的地点，请护林员去那些地点巡逻，去看看在那里是否可以发现猎套。

4. 效果与应用价值

在对于模型进行验证的一个月中，护林员发现了 19 次人类活动痕迹，还发现了已经被猎杀的大象，以及多个已经生效或部署好的羚羊猎套和大象猎套。将这一个月的数据跟以往的数据进行对比，发现其超过了 91%的历史月份，并且找到猎套的次数远高于历史平均数据。

此外，为了对比分析，护林人员还在更长的时间范围内（八个月）进行了验证，包括 5 个模型预测出来的高风险区域和 22 个低风险区域，结果发现在被预测为高风险的地区，每公里发现猎套的数量十倍于低风险区域。（高风险区域高于 0.1 个每公里，而低风险区域则约为 0.01 个每公里）

可见，通过训练好的机器学习模型，可以比较好的预测出盗猎的高风险和低风险区域，可以给当地的巡护人员作为重要参考。

机器学习模型可以学习到特征之间的一些关联，并且是比较通用的。除了在乌干达地区进行实践外，研究人员还通过世界自然基金会（WWF）和中国东北地区的护林团体取得联系，在当地开展了合作和研究。在我国的东北地区，历史数据和地理信息数据相较于乌干达林区更为缺少。为此，研究人员通过卫星图片提取出相关的地理信息，通过向护林员发放问卷的方式收集额外信息。然后再通过训练模型进行预测。根据实地的验证，发现模型给出的预测结果也能够帮助护林员更好地发现猎套和盗猎行为。

可见，通过历史和地理数据训练这一模型是很有价值的：它能够帮助预测盗猎行为发生的风险，从而帮助护林员规划巡逻路线，更为高效地利用人力资源。

5. 相关说明

上面介绍了该项目中的一个部分，利用机器学习解决一个建模为回归问题的实际问题。除此之外，后续还需要对于巡逻路线进行规划，这也可以利用机器学习的办法，在该项目中，研究人员利用了博弈论来完成这一规划，再进一步考虑当地的复杂地形进行具体的路线设计。

该项目为 PAWS (Protection Assistant for Wildlife Security)，参见 2021.06.19 来自 CMU 的方飞教授的一席演讲 博弈论+机器学习=？。

2. 证明《统计学习方法》习题 1.2：通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当损失函数是对数损失函数 时，经验风险最小化等价于极大似然估计。

经验风险最小化，即对于数据点的损失最小化 $\min \sum_i L(y|\hat{y})$。当损失函数为对数损失函数时，其形式为

$$\min \sum_i -\log P(y|x) = \min -\log \prod_i P(y|x)$$

而根据似然函数的定义 $L(\theta) = \prod_i P_\theta(y|x)$，可知最小化上式等价于最大化似然函数。