第一周

Day4-Day5

1. 回顾：请参考课本 3.3 节，对决策树的几个算法（ID3，C4.5，CART）进行总结。其中总结需要包括：算法的整体流程是什么？什么是熵？什么是信息增益？什么是基尼指数？

参考答案：参考课本 3.3 节进行总结。

2. 回顾：请参考课本 3.3 节，为什么决策树需要剪枝？如何进行剪枝？

参考答案：参考课本 3.3 节进行总结。

剪枝处理是决策树学习算法用来解决过拟合问题的一种办法。在决策树算法中，为了尽可能正确分类训练样本， 节点划分过程不断重复， 有时候会造成决策树分支过多，以至于将训练样本集自身特点当作泛化特点， 而导致过拟合。 因此可以采用剪枝处理来去掉一些分支来降低过拟合的风险。

剪枝的基本策略有预剪枝（pre-pruning）和后剪枝（post-pruning）。

预剪枝：在决策树生成过程中，在每个节点划分前先估计其划分后的泛化性能， 如果不能提升，则停止划分，将当前节点标记为叶结点。

后剪枝：生成决策树以后，再自下而上对非叶结点进行考察， 若将此节点标记为叶结点可以带来泛化性能提升，则修改之，而后剪枝中的代价复杂度剪枝（CCP）需要重点掌握。

3. 补充：决策树的优缺点是什么？

参考答案：

优点：

1、决策树算法易理解，机理解释起来简单。

2、决策树算法可以用于小数据集。

3、决策树算法的时间复杂度较小，为用于训练决策树的数据点的对数。

4、相比于其他算法只能分析一种类型变量，决策树算法可处理多种类别。

5、能够处理多输出的问题。

6、对缺失值不敏感。

7、可以处理不相关特征数据。

8、效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

9. 是其他更复杂的算法的基础。

缺点：

1、对连续性的字段比较难预测。

2、容易出现过拟合。

3、当类别太多时，错误可能就会增加的比较快。

4、在处理特征关联性比较强的数据时表现得不是太好。

5、对于各类别样本数量不一致的数据，在决策树的一些算法中，信息增益的结果偏向于那些具有更多数值的特征。

4. 回顾：逻辑回归能够手推一把吗？请拍照上传（其中包括：伯努利过程，极大似然，损

失函数，梯度下降）

参考答案：参考本章中逻辑回归视频课程，手推逻辑回归。

参考链接：https://www.cnblogs.com/ModifyRong/p/7739955.html

5. 补充：逻辑回归优缺点是什么？

参考答案：

优点：

(1) 形式简单，模型的可解释性非常好。从特征的权重可以看到不同的特征对最后结果的影响，某个特征的权重值比较高，那么这个特征最后对结果的影响会比较大。

(2) 模型效果不错。在工程上是可以接受的（作为 baseline），如果特征工程做的好，效果不会太差，并且特征工程可以大家并行开发，大大加快开发的速度。

(3) 训练速度较快。分类的时候，计算量仅仅只和特征的数目相关。并且逻辑回归的分布式优化 sgd 发展比较成熟，训练的速度可以通过堆机器进一步提高，这样我们可以在短时间内迭代好几个版本的模型。

(4) 资源占用小,尤其是内存。因为只需要存储各个维度的特征值，。

(5) 方便输出结果调整。逻辑回归可以很方便的得到最后的分类结果，因为输出的是每个样本的概率分数，我们可以很容易的对这些概率分数进行 cutoff，也就是划分阈值(大于某个阈值的是一类，小于某个阈值的是一类)。

缺点：

(1) 准确率并不是很高。因为形式非常的简单(非常类似线性模型)，很难去拟合数据的真实分布。

(2) 很难处理数据不平衡的问题。举个例子：如果我们对于一个正负样本非常不平衡的问题比如正负样本比 10000:1.我们把所有样本都预测为正也能使损失函数的值比较小。但是作为一个分类器，它对正负样本的区分能力不会很好。

(3) 处理非线性数据较麻烦。逻辑回归在不引入其他方法的情况下，只能处理线性可分的数据，或者进一步说，处理二分类的问题 。

(4) 逻辑回归本身无法筛选特征。有时候，我们会用 gbdt 来筛选特征，然后再上逻辑回归。

6. 补充：为什么逻辑回归需要归一化？

参考答案：逻辑回归使用梯度下降方法进行优化，归一化可以提高收敛速度，增加收敛精度。

7. 补充：关于逻辑回归，连续特征离散化的好处？

参考答案：在工业界，很少直接将连续值作为逻辑回归模型的特征输入，而是将连续特征离散化为一系列 0、1 特征交给逻辑回归模型，这样做的优势有以下几点：

(1) 离散特征的增加和减少都很容易，易于模型的快速迭代；

(2) 稀疏向量内积乘法运算速度快，计算结果方便存储，容易扩展；

(3) 离散化后的特征对异常数据有很强的鲁棒性：比如一个特征是年龄>30 是 1，否则 0。如果特征没有离散化，一个异常数据“年龄 300 岁”会给模型造成很大的干扰；

(4) 逻辑回归属于广义线性模型，表达能力受限；单变量离散化为 N 个后，每个变量有单独的权重，相当于为模型引入了非线性，能够提升模型表达能力，加大拟合；

(5) 离散化后可以进行特征交叉，由 M+N 个变量变为 M*N 个变量，进一步引入非线性，提升表达能力；

(6) 特征离散化后，模型会更稳定，比如如果对用户年龄离散化，20-30 作为一个区间，不会因为一个用户年龄长了一岁就变成一个完全不同的人。当然处于区间相邻处的样本会刚好相反，所以怎么划分区间是门学问；

(7) 特征离散化以后，起到了简化了逻辑回归模型的作用，降低了模型过拟合的风险。
参考链接：https://www.zhihu.com/question/31989952/answer/54184582

8. 补充：逻辑回归能否解决非线性的分类问题?
参考答案：
可以，只要使用核技巧(kernel trick，在下周的 SVM 中会讲到核技巧)。

不过，通常使用的 kernel 都是隐式的，也就是找不到显式地把数据从低维映射到高维的函数，而只能计算高维空间中数据点的内积。在这种情况下，logistic regression 模型就不能再表示成$w^T x + b$的形式（primal form），而只能表示成$\sum_i a_i \langle x_i, x \rangle + b$的形式（dual form）。忽略那个$b$的话，primal form 的模型的参数只有$w$，只需要一个数据点那么多的存储量；而 dual form 的模型不仅要存储各个a_i，还要存储训练数据x_i本身，这个存储量就大了。

SVM 也是具有上面两种形式的。不过，与 logistic regression 相比，它的 dual form 是稀疏的——只有支持向量的a_i才非零，才需要存储相应的x_i。所以，在非线性可分的情况下，SVM 用得更多。
参考链接：https://www.zhihu.com/question/29385169/answer/45023550