COMP9417 - 机器学习教程。非参数建模

问题1.树木的表现力

给出决策树来表示下列布尔函数，其中变量A、B、C和D的值为t或f，而类值为真或假。你能观察到函数的复杂性增加对它们作为决策树的表达形式有什么影响吗？

(a) $A \wedge \neg B$

(b) $A \vee [B \wedge C]$

(c) A *XOR* B

(d) $[a \wedge b] \vee [c \wedge d]$。

解决方案。

```
A=t。
|    B = f: 真
|    B = t。假的
A=f。
|    B = f: 假的
|    B = t。
|    |    C = t。真
|    |    C = f: 假的
```

解决方案。

```
A=t。
|    B = t。假的
|    B = f: 真
A=f。
|    B = t。真
|    B = f: 假的
```

解决方案

```
。
A=t。B = t:True B
|    = f:
|    |    C = t。
|    |    |    D = t。真
|    |    |    D = f: 错
|    |    C = f    假的
A=f。。
|    C=t。
|    |    D = t。真
|    |    D = f: 错
|    C = f: 假的
```

请注意，随着目标表达式变得更加复杂，重复子树的*复制*效应，例如，在*d*的树中。这是一种假设类模型（这里是决策树）可以适合任何布尔函数的情况，但为了表示函数，树可能需要非常复杂。这使得它很难学习，并且需要大量的数据！

问题2.决策树学习

(a) 假设我们从下面的训练集中学习一棵决策树来预测给定属性A、B和C的Y类，并且不做任何修剪。

A	B	C	Y
0	0	0	0
0	0	1	0
0	0	1	0
0	1	0	0
0	1	1	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	0	1
1	1	1	0
1	1	1	1

这个数据集的训练集误差是多少？将你的答案表示为12个例子中被错误分类的数量。

解决方案。

2. 有两对例子，属性A、B和C的值相同，但Y类的值不同（矛盾）。这些对中的每一个例子总是会有一个被错误分类（噪音）。

(b) 决策树分类器的一个很好的特点是，它们可以学习树来进行*多类*分类，即问题是学习将每个实例准确地分类到$k>2$类中的一个。

假设要在一个任意的数据集上学习决策树，其中每个实例都有一个离散的类值，属于k>2类中的一个。任何数据集的最大训练集误差，以分数表示，是多少？

解决方案。
首先考虑所有k类值都是均匀分布的情况，所以有$\frac{1}{k}$ 个例子
训练集中的每一个类别。在最坏的情况下，可以学习一棵决策树，预测
所有例子的k类中的一个。这将使$\frac{1}{k}$ 训练集的正确性，并使
$1 - \frac{1}{k} = \frac{k-1}{k}$ 错误占训练集的比例。
如果任何一个类别有超过$\frac{1}{k}$ 的例子，那么最坏的情况是保证决策树的
预测该类，这将r 由于该类别现在代表了超过$\frac{1}{k}$ ，因此减少了错误。
训练集。

问题3.线性平滑

在本周的实验中，我们介绍了线性平滑法，也称为核平滑法，我们从头开始实施，并将其应用于模拟数据集。下图取自Hastie、Tibshirani和Friedman的《统计学习要素》，是线性平滑器工作时的最佳写照。数据是从以下数据生成过程中模拟出来的。

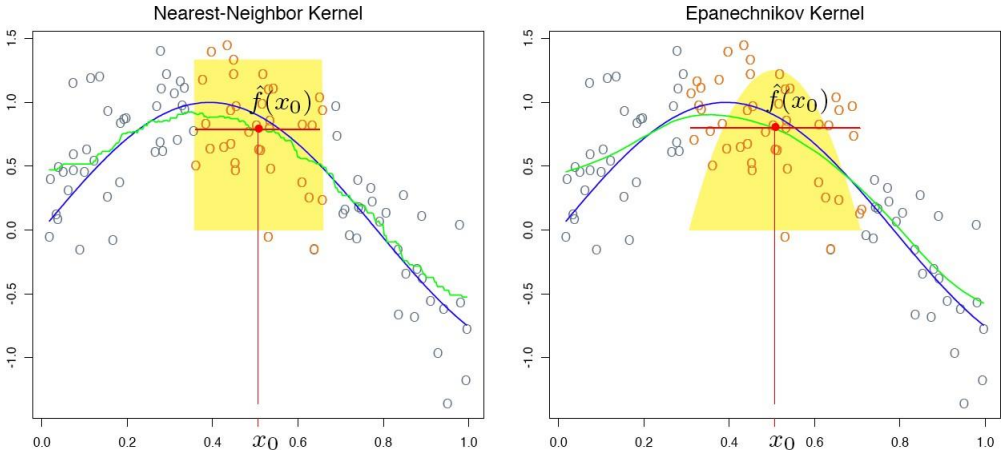
$$y = \sin(4x) + E。 \qquad e \sim n\ (0,\ 1/3)\ 。$$

在左边，我们看到了工作中的 "最近邻 "核，我们在实验室中称之为盒式核，在右边我们看到了埃帕尼科夫核，也是在实验室中介绍的。我们在这里包括他们的定义。

$$K(u) = \mathbf{1}\{|u| \leq 1/2\}。 \qquad \text{箱型车内核}$$
$$K(u) = \mathbf{1}\{|u| \leq 1\}\frac{3}{4}(1 - u^2) \qquad \text{埃帕尼科夫内核}$$
。

还记得在实验室里，线性平滑器预测的形式。

$$\hat{f}(x_0) = \sum_{i=1}^{n} \frac{K\left(\frac{|x_i - x_0|}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{|x - x_0|}{h}\right)} Y_i。$$



复习一下实验室的线性平滑部分，然后写下几句话，描述图中发生的事情。一定要详细描述以下每一项所代表的内容。

(i) 蓝色曲线

(ii) 黑色的散落物

(iii) 红色散点

(iv) 黄色区域

(v) 水平红线

(vi) 水平红线上的红点

(vii) 绿色曲线

解决方案。

(i) 蓝色曲线：这是数据生成过程中使用的真实函数$f(x) = \sin(4x)$。

(ii) 黑色散点：这些是真实函数的取样点，加入了噪声。所选择的内核只看满足特定条件的点（它们需要足够接近查询点x_0）。具体来说，在盒式车内核中，如果$|X_i - x_0|_2 / h \leq 0.5$，分子为1，否则为零。同样，在Epanechnikov的情况下，分子为1。
只有当$|X_i - x_0|_2 / h \leq 1$时，分子才是非零。这只是kNN的一个平滑版本
退步。

(iii) 红色散点 这些是满足内核条件的点，因此被用于估计输入点x_0 的预测。换句话说，黑色散点对预测$\hat{f}(x_0)$没有贡献。

(iv) 的黄色区域 这显示了内核之间的重要区别，因为它代表了分配给红色点的权重。在左图中，分配给所有红色点的权重是统一的，所以附近的点和远处的点（从x_0）对$\hat{f}(x_0)$的贡献都是一样的。在右图中，附近的点比远处的点有更高的权重（更多的贡献），这似乎更合理。

(v) 水平红线 红色散点是x的邻域$_0$ ， 基于某种意义上的内核K，红色水平线表示包含这个邻域的最大矩形的边长。

(vi) 水平红线上的红色点 这就是预测值$f\hat{}(x_0)$。

(vii) 绿色曲线 这是运行线性平滑器对域中所有点的拟合结果。