Data Mining, Spring 2018

Problem Set #2: Supervised Learning II

(Due on May 4 Friday, 2018 at 11:59pm)

**Submission Instructions**

These questions require thought but do not require long answers. Please be as concise as possible. You should submit your answers as a write-up in PDF format to [DataMining\_2018@126.com](mailto:DataMining_2018@126.com). The email title is formatted as “hwk2\_学号\_姓名”.

**Questions**

1. **模型的性能度量**

我们需要比较两个分类模型和。他们在10个二类（+或-）样本所组成的测试集上的分类结果如下表格中所示。假设我们更关心正样本是否能被正确检测。

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | True Class | Scores from | Scores from |
| 1 | + | 0.73 | 0.61 |
| 2 | + | 0.69 | 0.03 |
| 3 | - | 0.44 | 0.68 |
| 4 | - | 0.55 | 0.31 |
| 5 | - | 0.67 | 0.45 |
| 6 | + | 0.47 | 0.09 |
| 7 | - | 0.08 | 0.38 |
| 8 | - | 0.15 | 0.05 |
| 9 | + | 0.45 | 0.01 |
| 10 | - | 0.35 | 0.04 |

1. 对于分类模型，取阈值为0.5，分别计算分类准确率（accuracy）、查准率（precision）、查全率（recall，又称真正例率，true positive rate，TPR）、假正例率（false positive rate，FPR）和F-measure；
2. 对于分类模型，取阈值为0.5，分别计算分类准确率（accuracy）、查准率（precision）、查全率（recall，又称真正例率，true positive rate，TPR）、假正例率（false positive rate，FPR）和F-measure；并与分类模型比较，分析哪个分类模型在这个测试集上表现更好；
3. 对于分类模型，取阈值为0.2，分别计算分类准确率（accuracy）、查准率（precision）、查全率（recall，又称真正例率，true positive rate，TPR）、假正例率（false positive rate，FPR）和F-measure；并讨论当阈值为0.2或0.5时，哪个分类模型的分类结果哪个更好；
4. 试讨论是否存在更好的阈值；若存在，请求出最优阈值并说明原因。
5. **神经网络**

考虑以下的二类训练样本集

|  |  |  |
| --- | --- | --- |
| Instance | Feature vector | Output label |
| 1 | (0, 0) | + |
| 2 | (1, 0) | + |
| 3 | (0, 1) | - |
| 4 | (-1, 0) | - |
| 5 | (1, -1) | - |

对此训练样本集，我们需要训练一个三层神经网络（输入层、单隐层、输出层），其中单隐层的单元（神经元）数目设为2，激活函数（activation function）为Sigmoid函数：

1. 在二维坐标系中画出这5个训练样本点，并讨论此训练样本集是否线性可分；
2. 试分析将Sigmoid激活函数换成线性函数的缺陷；
3. 令初始化参数全部为0，试运用前馈（feedforward）算法计算在初始化参数下此三层神经网络的输出；然后运用反向传播（backpropagation）算法，计算代价函数对所有参数的偏导数，并讨论将初始化参数全部设为0所带来的问题；
4. 试给出一个神经网络（画出架构图，并写出激活函数及其对应的参数），使此训练样本集的5个训练样本点都可以被正确分类。
5. **决策树**

考虑以下的二类训练样本集

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | A | B | Class Label |
| 1 | T | F | + |
| 2 | T | T | + |
| 3 | T | T | + |
| 4 | T | F | - |
| 5 | T | T | + |
| 6 | F | F | - |
| 7 | F | F | - |
| 8 | F | F | - |
| 9 | T | T | - |
| 10 | T | F | - |

1. 计算以属性A或B为划分的信息熵（Entropy）增益，并说明决策树学习算法选择哪个属性进行划分；
2. 计算以属性A或B为划分的Gini增益，并说明决策树学习算法选择哪个属性进行划分；
3. 计算以属性A或B为划分的分类误差（Classification Error）增益，并说明决策树学习算法选择哪个属性进行划分；
4. 说明信息熵增益、Gini增益和分类误差增益对属性选择有不一样的偏好。