第五次作业答案

1、区别：分类就是按照某种标准给对象贴标签，再根据标签来区分归类，是有监督的学习；聚类是指事先没有“标签”而通过某种成团分析找出事物之间存在聚集性原因的过程，属于无监督的学习。

2、

（1）$\text{info}(D) = 0.991076$

（2）

属性 A 的信息增益

$$\text{info}_A(D) = \frac{4}{9} * \text{info}_A(D_T) + \frac{5}{9} * \text{info}_A(D_F)$$

$\text{info}_A(D_T) = 0.811278$
$\text{info}_A(D_F) = 0.721928$
$\text{info}_A(D) = 0.761639$
$Gain(A) = \text{info}(D) - \text{info}_A(D) = 0.229437$

属性 B 的信息增益

$$\text{info}_B(D) = \frac{5}{9} * \text{info}_B(D_T) + \frac{4}{9} * \text{info}_B(D_F)$$

$\text{info}_B(D_T) = 0.97095$
$\text{info}_B(D_F) = 1$
$\text{info}_B(D) = 0.983861$
$Gain(B) = \text{info}(D) - \text{info}_B(D) = 0.007215$

（3）

$Gain(A) = \text{info}(D) - \text{info}_A(D) = 0.229437$
$Gain(A) = \text{info}(D) - \text{info}_B(D) = 0.007215$
$Gain(C) = \text{info}(D) - \text{info}_C(D) = 0.143$

根据信息增益，选择 A 属性作为最佳划分。

（4）

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} = 0.231$$

$$GainRatio(B) = \frac{Gain(B)}{SplitInfo_A(D)} = 0.007$$

根据信息增益率，选择 A 属性

（5）

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 = 0.494$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) = 0.344$$

$$Gini_B(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) = 0.488$$

根据 gini 系数最小的属性，因此选择 A 属性。

3、

（1）

$$P(A = 1|T) = \frac{3}{5}$$

$$P(B = 1|T) = \frac{2}{5}$$

$$P(C = 1|T) = \frac{4}{5}$$

$$P(A = 1|F) = \frac{2}{5}$$

$$P(B = 1|F) = \frac{2}{5}$$

$$P(C = 1|F) = \frac{1}{5}$$

（2）

$$P(A = 1, B = 1, C = 1|T) = P(A = 1|T) * P(B = 1|T) * P(C = 1|T) = 24/125$$

$$P(A = 1, B = 1, C = 1|F) = P(A = 1|F) * P(B = 1|F) * P(C = 1|F) = 4/125$$

$$P(T) * P(A = 1, B = 1, C = 1|T) = 12/125 > P(F) * P(A = 1, B = 1, C = 1|F) = 2/125$$

因此预测样本的类标号为 T。

（3）

$$P(A = 1|T) = \frac{3}{5}$$

$$P(B = 1|T) = \frac{2}{5}$$

$P(A = 1|T) * P(B = 1|T)$ 子 $P(A = 1, B = 1|T)$

因此说明给定类 T，变量 A 和 B 条件不独立。

4、

基本原理

支持向量机的基本模型是定义在特征空间上的间隔最大的线性分类器， 支持向量机还包括核技巧， 这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化， 可形式化为一个求解凸二次规划的问题， 也等价于正则化的合页损失丽数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

常用的核函数

线性核函数:主要用于线性可分的情况。

多项式核函数:一种非稳态核函数，适合于正交归一化后的数据。

径向基核函数:具有很强的灵活性， 应用广泛。大多数情况下有较好的性能。

Sigmoid 核:来源于 MLP 中的激活函数， SVM 使用 Sigmoid 相当于一个两层的感知机网络。

5、

神经网络是由具有适应性的简单单元组成的广泛并行互联的网络， 它的组织能够模拟生物神经系统对真实世界物体做出的交互反应。神经网络的训练采用 BP 算法，给定训练集和学习率，随机初始化网络中的连接权重和阈值。不断地根据当前的参数计算当前样本的输出， 并和理想的输出计算均方误差， 并且根据链式法则计算输出层和隐层神经元的梯度项， 来更新网络中的连接权重和阈值，直到达到停止条件。

sigmoid 函数将任意大小的输人都压缩到[0,1]之间， 输入值越小， 压缩后越趋近于 0。它在神经网络中常常用作二分类器最后一层的激活函数， 可以将任意实数值转换为概率。 sigmoid 函数的导数是其本身的函数， 计算方便。其最明显的缺点是容易饱和， 出现梯度消失等问题， 导致层数较多的深度神经网络难以有效训练;另外， 它的输出均大于 0，使得输出不是零均值， 所以 sigmoid 函数现在很少在深度神经网络的中间层作为激活函数使用。

tanh 函数将任意大小的输入都压缩到[-1.1]之间， 其输出均值是 0，使得收敛速度比 sigmoid 要快， 但是其仍然具有饱和性， 会造成梯度消失， 同时还有更复

杂的幂运算。

ReLU将负数部分置零，保留正数部分不变，在计算上非常高效，且避免了梯度消失的问题。 ReLU函数的优点包括：计算简单，不会导致梯度消失问题，收敛速度较快。然而， ReLU函数在负数部分输出为0，可能导致神经元死亡；不是零均值激活函数，可能导致训练时的震荡问题。

6、

过拟合是指模型对训练集的学习程度过高， 学习到了训练集的数据特性， 但没有理解数据背后的规律， 泛化能力差， 导致模型在训练集上表现很好， 但在测试集上却表很差。发生过拟合一般是由于模型复杂度过高， 或者数据集样本不足。

解决过拟合问题，常用的有以下几个方案:

①提前停止训练;

②设置 Dropout;

③获取和使用更多的数据;

④控制模型的复杂度;

⑤删除冗余特征。