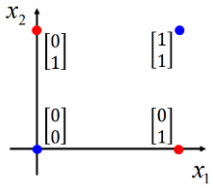
数据挖掘 2024 考试回忆

十道选择+六个简答+一个计算
英文试题+中文作答

选择题，10 个，3 分/个

记不太全，每道题三个选项，感觉算试卷中比较考察对内容的理解程度的，难度稍微大一点，但每道题分值不多

1. 哪个分类器能实现 100%正确率 （C）
   A. SVM　B.逻辑回归　C.决策树

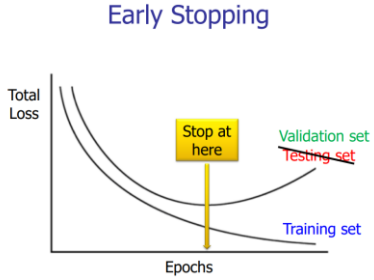


2. 增大软间隔 SVM 的 C 会导致什么？
   A．会导致训练误差增大
   B．会导致对离群点的检测更加敏感
   C．会导致 margin 增大

$$\min_{w,b,\xi} \quad \frac{1}{2}\boldsymbol{w}^{T}\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i$$

惩罚项

3. 给定一个 3*64*64 的图片，一个 1*1 的卷积，输出通道为 1，每个卷积核有一个 bias，询问参数量()
   A. 1　B. 4　C.
4. 对早停概念的考察，训练到模型在测试集的 err 最低时，终止训练



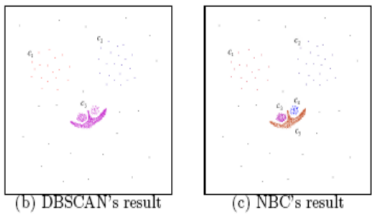
5. 在 RNN 中，输入神经网络和激活函数的值为"Nan"，最可能的原因()
   A. 梯度消失　B. 梯度爆炸　C. 神经网络层数太多
6. 造成梯度消失的激活函数为()
   A. Sigmoid　B. ReLu　C. none
7. 随机森林的构造，不放回采样等
8. K-means

简答题

共有 6 个简单题，每个题目两小问

1. PCA 的概念
2. PCA 算法的核心步骤

1. DBSCAN 的实现
2. DBSCAN 的主要缺点，举例说明



1. CNN 的 motivation：稀疏连接、参数共享、Translation-invariant
2. CNN 和全连接前馈神经网络的联系

1. Self-attention 相比 RNN 的优点
2. Transformer 中 encoder 和 decoder 的区别

1. 解释 Dropout 和集成学习的关联
2. 给定一个 batch data，使用 drop 来模拟训练和测试过程

1. sigmoid 函数在 RNN 中为什么会造成梯度消失
2. 提出一种解决方案，ReLu 激活

计算题

一个 CNN 网络，卷积-池化-卷积-池化
输入 32*32 图片

Conv1: 8 个卷积核 3*3，padding=1，每个卷积核有一个 bias
Maxpool1：最大池化，卷积核 2*2，stride=2
Conv2：16 个卷积核 5*5，padding=2，每个卷积核有一个 bias
Maxpool2：最大池化，卷积核 2*3，stride=2

求出每个步骤操作后的输出维度和参数量