数据挖掘2024考试回忆

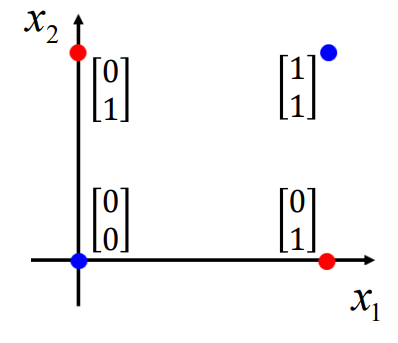
十道选择+六个简答+一个计算

英文试题+中文作答

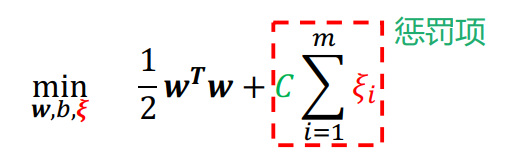
## 选择题，10个，3分/个

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

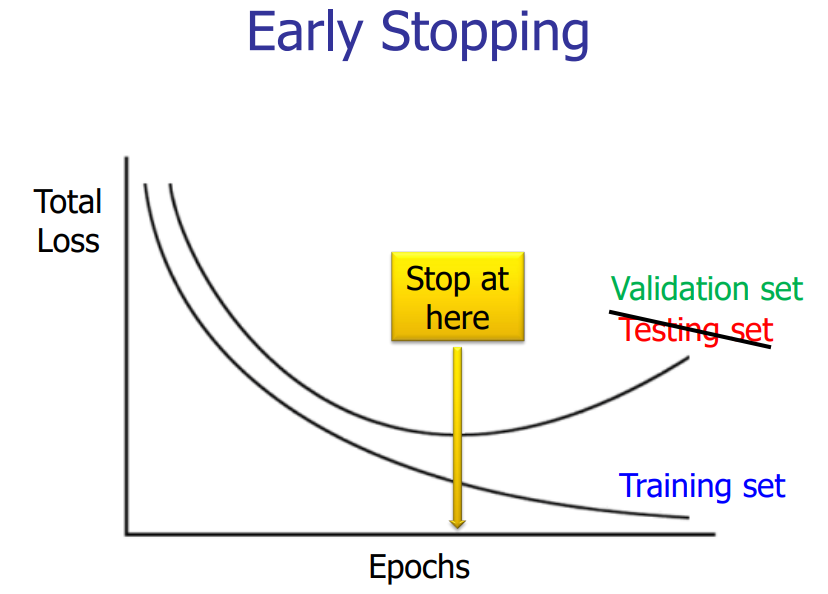
1. 哪个分类器能实现100%正确率（C）
2. SVM B.逻辑回归 C.决策树



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



1. 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()
2. 1 B. 4 C.
3. 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

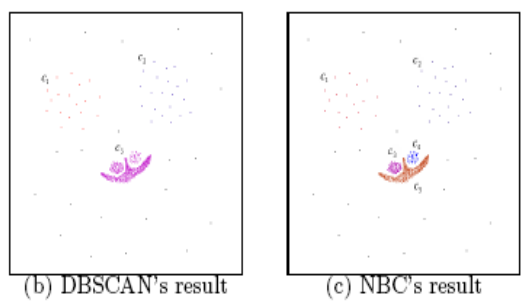


1. 在RNN中，输入神经网络和激活函数的值为”Nan”，最可能的原因()
2. 梯度消失 B. 梯度爆炸 C. 神经网络层数太多
3. 造成梯度消失的激活函数为()
4. Sigmoid B. ReLu C. none
5. 随机森林的构造，不放回采样等
6. K-means

## 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
3. DBSCAN的实现
4. 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
3. Self-attention相比RNN的优点
4. Transformer中encoder和decoder的区别
5. 解释Dropout和集成学习的关联
6. 给定一个batch data，使用drop来模拟训练和测试过程
7. sigmoid函数在RNN中为什么会造成梯度消失
8. 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