

2023-2024 学年第一学期

机器学习期末大作业

一、简述奇异值分解的原理和算法实现步骤,用代码实现对一张三通道彩色图像的压缩。(25 分)

要求:

- (1) 输出图像的维度信息(宽 width、高 height、维度数 channels),给出 6 个不同数量奇异值下图像的压缩效果(奇异值在代码中用 k 表示),要求奇异值数值分布均衡,能够较好体现图像压缩的效果;
- (2) 输出不同 k 值下压缩后的三通道图像,并给出相应图像的大小信息(右键属性查看图片大小并截图);
- (3) 关键行代码需要对算法进行注释;
- (4) 算法末尾需同步输出个人信息以及系统时间,可通过创建函数或语句实现。如(李某某机器学习作业 **模块 完成时间: ****年**月**日);(注:注释和个人信息代码全无,采用图片重复者不得分)

二、简述决策树算法的原理和算法实现步骤,重点描述样本集分裂的过程,用代码实现决策树分类(代码源自实验四)。(35 分)

实验四中的数据集为:

编号	年龄(岁)	身高(cm)	学历	月薪(元)	是否相亲
1	35	176	本科	20000	否
2	28	178	硕士	10000	是
3	26	172	本科	25000	否
4	29	173	博士	20000	是
5	28	174	本科	15000	是

在实验四所给代码中的对应训练集样本表示为:

```
# 表3-1 某人相亲数据
blind_date = [[35, 176, 0, 20000, 0],
               [28, 178, 1, 10000, 1],
               [26, 172, 0, 25000, 0],
               [29, 173, 2, 20000, 1],
               [28, 174, 0, 15000, 1]]
```

代码中数据第一列与“年龄”特征对应(代码中特征编号为0),第二列与“身高”特

征对应（代码中特征编号为 1），第三列与“学历”特征对应（代码中特征编号为 2），其中 0 代表本科，1 代表硕士，2 代表博士，第四列与“月薪”特征对应（代码中特征编号为 3），代码最后一列为决策结果，0 表示不相亲（否），1 表示相亲（是）。

测试集样本表示为：

```
test_sample = [[24, 178, 2, 17000],
               [27, 176, 0, 25000],
               [27, 176, 0, 10000]]
```

决策结果显示（根据红框代码给出输出结果）：

```
print('信息增益二叉树对样本进行预测的结果：')
test_sample = [[24, 178, 2, 17000],
               [27, 176, 0, 25000],
               [27, 176, 0, 10000]]
for x in test_sample:
    print(predict(x, tree))

print("基尼指数二叉树：")
tree = build_biTree(blind_date, splitInfo="gini")
print tree(tree)
print('基尼指数二叉树对样本进行预测的结果：')
test_sample = [[24, 178, 2, 17000],
               [27, 176, 0, 25000],
               [27, 176, 0, 10000]]
```

要求：

- （1）给自己拟定虚拟的“对象”，依据自己的心仪标准修改训练样本集，要求至少 10 个/条样本以上，对决策树模型进行训练；
- （2）输入 1~3 个测试样本，给出相应的测试结果；
- （3）算法末尾需同步输出个人信息以及系统时间。如（李某某机器学习作业 **模块 完成时间：****年**月**日）；

（注：注释和个人信息代码全无、未更改训练集者不得分）

三、简述 BP 神经网络的算法原理和算法实现步骤，重点描述误差反向传播过程，用神经网络模型来拟合多项式，采用 Tensorflow2.0 框架来实现。首先要确定多项式的次数，多项式的次数根据本人学号予以确定，如某位同学学号为：

2005140036，则取非零的数字 2、5、1、4、3、6 为系数拟合多项式，拟合多项式为：

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_n x^n$$

$$h(x) = 2 + 5x + 1.x^2 + 4.x^3 + 3.x^4 + 6.x^5$$

注：多项式次数不超过 6。

(40 分)

要求:

- (1) 分别输出训练 1000、3000、5000、10000 轮次数的拟合结果 (或你认为能较好反应训练效果的几个训练次数), 完整展示出欠拟合、拟合程度较好、过拟合几种训练结果, 给出拟合结果图;
- (2); 截图显示训练轮次和损失的结果;
- (3) 关键行代码需要对算法进行注释;
- (4) 算法末尾需同步输出个人信息以及系统时间。如 (李某某机器学习作业 **模块 完成时间: ****年**月**日);

(注: 需读懂题干内容, 确认输入正确, 注释和个人信息代码全无者不得分)

封面参照下页: