一. 实验名称：K-近邻算法

K-近邻算法是一个分类算法，算法采用不同特征值之间的距离方法进行分类。在一个样本集中每个数据都存在标签，由此我们知道样本集中每一个数据与所属分类的对应关系。输入没有标签的新数据后，将新数据的每个特征与样本集中数据对应的特征进行比较，选择前 K 个最相似的数据里面出现次数最多的类别，作为新数据的分类。

二. 数据描述：

（此处选择的数据样例仅供参考，也可以选择任何测试数据）采用预测隐形眼镜类型数据作为样例数据，Lenses 数据集包含很多患者眼部状况的观察条件以及医生推荐的隐形眼镜类型。 Lenses.txt 数据集一共 24 条数据， 每条数据为一个患者的观察情况(age，prescript，astigmatic，tearRate),隐形眼镜类型包括硬材质、软材质以及不适合佩戴隐形眼镜（hard，soft，no lenses）。

三. 实验内容

1. 分析数据
2. 可视化展示
3. 算法实现

四. 实验环境和编程工具

系统环境： Windows all/Linux
程序语言： Python
编译工具： PyCharm
第三方包： pickle， matplotlib， operator， numpy

五. 实验步骤

1. 读取并分析数据集

   根据数据描述，lenses 数据集一共有 4 个特征：age，prescript，astigmatic，tearRate 和一个类别标签信息。根据 lenses.txt 了解数据在文件中的存储格式。将数据集按比例分为训练集、测试集

2. 数据量化

   将 4 个特征维度的数据量化，如每个特征可能出现 3 中情况，则分别把每种情况编码为 1、2、3，以此把文本属性数据量化

   由于不同特征数值差距很大的数据会对算法的准确性影响较大，所以在使用不同取值范围的特征值时，需要先归一化。一般是将数值转换到 0 到 1 区间内的值：newX = (x - min)/(max-min)

3. 数据可视化

   在不同的两个特征上，用 matplotlib 做散点图，并标记出类别标签，分析特征与类别的分布情况。

4. 测试算法

   构建 KNN 算法，测试训练数据集上，选择不同的 K 值，计算模型准确率、错误率