**一、课题综述**

* 1. **课题说明**

**包括小组成员任务划分，每个成员完成哪块工作**

* 1. **课题目标（示例）**

本课题使用的数据集来自于数据分析与数据挖掘竞赛Kaggle，该竞赛为数据科学领域著名的国际性赛事之一。课题使用的数据集为带标签的图像数据集，包含带有裂痕和不带有裂痕的桥梁、墙和人行道图片。课题的目标为对于目标数据集，搭建相应的传统机器学习模型和深度神经网络模型，完成基于机器学习的图片分类任务，并能取得较优的分类性能。

* 1. **课题数据集（示例）**

本课题使用的数据集为来自Kaggle竞赛的“Structural Defects Network (SDNET) 2018”(<https://www.kaggle.com/aniruddhsharma/structural-defects-network-concrete-crack-images>)，该数据集包含有桥梁表面、步行道表面和墙面三种不同的图片，每种图片都包含有带裂痕的图片和不带裂痕的图片共两类图片。数据集共包含56000 张图片，其中图片上的建筑裂痕最窄为0.06 毫米，最宽为25毫米。数据集中的部分图片可能存在着一定的遮挡干扰，例如阴影、建筑表面粗糙、建筑表面脱落、拍摄角度变换、建筑表面的孔洞和背景噪声，目标数据集的分类任务即为区分带裂痕和不带裂痕的图片。

**二、实验报告设计**

* 1. **数据准备**

说明实验数据集的获取过程

* 1. **数据预处理**

本实验采用的数据预处理包括：1）特征归一化Feature Scaling，2）主成分分析Principal Components Analysis (PCA)，3）多项式回归Polynomial regression

1. 特征归一化

Feature Scaling可以将特征的取值范围缩放到相似的尺度上，以便更好的拟合模型，即消除不同特征之间的尺度差异，使得每个特征对模型的贡献相对均衡。本次实验使用的是标准化，其公式如下：

1. 主成分分析

主成分分析可以将原始的高维特征空间降维到一个低维的特征空间，以更好的拟合模型和提高模型的性能。通过降维，可以减少特征数量，降低计算复杂度和减少存储空间，同时可以提高模型的鲁棒性和泛化能力。在进行PCA之前，需要对数据进行标准化处理，避免特征之间的数量级差异对PCA结果的影响。

PCA主要是通过线性变换将原始特征空间中的数据映射到一个新的地位空间中，使得新空间中的数据保留原始数据中的主要信息。

在本实验中，PCA的主要步骤是：

1. 计算协方差矩阵：协方差矩阵可以反映不同特征之间的相关性。
2. 计算协方差矩阵的特征值和特征向量：特征向量可以表示协方差矩阵的主要方向，而特征值表示在该方向上的方差大小。可以通过特征值分解的方法，得到协方差矩阵的特征值和特征向量。
3. 选择主成分：选择在新的地位空间中保留的特征数量，通常可以选择保留协方差矩阵的前k个特征向量。这些向量构成了一个转换矩阵W，可以将原始数据集X通过矩阵乘法映射到新的低维空间中去。
4. 映射：公式如下，其中，W为m x k的矩阵，X为 n x m的矩阵（n为样本数量，m为特征数量）
5. 多项式回归

数据升维是一种处理非线性关系的方法，通过将原始数据进行转换，将地位特征映射到更高维空间中，从而使得模型可以更好的拟合数据。

在本实验中，数据升维可以通过多项式回归来实现。多项式回归是一种特殊的线性回归算法，将特征向量转化成高位向量的形式，来处理非线性关系的问题。即用更高阶的函数来拟合数据，从而提高模型的拟合度和预测性能。

但是其具有更复杂的模型，可能导致过拟合等问题

* 1. **模型搭建**

本次实验主要是搭建线性回归模型，封装在项目的regression\_model.py中的LR类中，主要包含1）初始化，2）训练，3）预测，4）评测功能

1. 初始化：将线性回归所需要的数据和参数进行初始化。将数据的数目和特征数分别存入num\_item和num\_feature中。包含的参数：权重矩阵theta（维数：num\_feature + 1），训练轮数epoch，学习率lr，正则化参数alpha，优化器参数gamma和beta，batch个数，优化器方法optimizer。需要特别注意的是：类中存储的训练数据x比实际的数据多一个特征，恒为1，使得x与theta的乘积包含bias。
2. 训练：
   1. 根据参数batch计算batch\_size
   2. 为优化器Momentum，RMSprop，Adam初始化初始化向量m, v, M
   3. 循环epoch
      1. 随机化x，y
      2. 根据batch\_size从x，y中取当前batch
      3. 计算梯度grad
      4. 根据不同优化器的方法对权重矩阵进行更新
      5. 对优化器向量进行更新
      6. 计算cost函数，输出或存储
3. 预测：使用权重矩阵与输入参数x点乘，得到预测值y\_pred
4. 评测：包含MSE，RMSE，MAE，R\_Squared四种评测函数
   1. **模型训练测试**

展示并分析机器学习模型的训练和测试过程

* 1. **结果可视化**

展示实验结果并可视化的分析实验结果

* 1. **分析和优化**

本实验在基础的线性回归模型上，主要做了如下优化：1）Feature Scaling，2）Batch，3）Normalization，4）PCA，5）Optimizer。并分别对其进行了实验进行分析。

1. Feature Scaling

原理参考2.2第一部分

在进行实验过程，出现了梯度爆炸（cost不断增加或者cost巨大）的情况。可能是由于学习率过大，导致模型参数更新过大，导致模型在优化过程中跳过最优点，从而使损失函数的值不断增大；或者模型过拟合，即当模型过于复杂或者训练集中样本数量不足，可能会导致模型过拟合训练数据，从而使得模型在测试数据上的表现不佳，损失函数的值不断增大。

Feature Scaling可以帮助解决上述问题的出现，通过避免特征值对模型训练的影响不均和加速模型收敛（如果特征值的范围差异较大，会导致梯度下降算法在更新模型参数时需要更多的迭代次数才能收敛。通过将特征值缩放到相似的范围，可以加速梯度下降算法的收敛速度）来解决上述问题。

根据实验，可以发现当合适的学习率（不会发生梯度爆炸）为1e-6的数量级。对学习率1e-5和1e-6，是否使用Feature Scaling进行了实验，可以发现，其可以有效的避免在较大学习率出现的梯度爆炸的问题，同时具有更快的收敛速度。

1. Batch

相比较于BGD，SGD有更快的收敛速度：SGD每次只更新一个样本的参数，因此每次迭代计算的速度比BGD要快的多；更少的内存消耗：由于SGD只需要考虑一个样本的梯度，因此只需要存储一个样本的参数和梯度，而BGD需要存储整个训练集的参数和梯度；更好的泛化能力：由于SGD更新参数时只考虑一个样本的梯度，因此SGD可以避免陷入局部最优解，并且可以更好的泛化到未见过的数据。

根据如图的实验结果，cost值下降到稳定值所用的时间SGD比mini-BGD和BGD要快，证明了SGD的收敛速度更快。同时根据验证数据test上MSE，RMSE，MAR等数据的比较，SGD的效果比mini-BGD和BGD更好。但是，由于单个样本的梯度时不稳定的，因此可以看到SGD的优化路径比较嘈杂。

1. Normalization

正则化可以用来减少模型的过拟合现象，提高模型的泛化能力。正则化通过添加一个惩罚项来限制模型的复杂度，从而降低模型对训练数据的过拟合程度。本实验主要采用的时L2正则化：惩罚项是模型参数的平方和，达到缩小模型参数，改善模型的鲁棒性的效果。

根据如图实验结果，正则化的效果不明显。可以是因为数据集比较小，所含的信息优先，过拟合的可能性相对较小；或者因为特征之间的相关性较小，过拟合的可能性也比较小，正则化的作用也相对较小。

1. PCA

原理参考2.2第二部分

根据如图实验结果，经过PCA降维后，在测试集上有较小的MSE，RMSE，MAE等，说明PCA降维可以有效的提高实验效果。主要原因可能为PCA提高了数据的线性可分性，即PCA将数据投影到了主成分上，使得数据在坐标系下更加线性可分；去除了数据噪声，PCA可以识别和去除数据中的噪声，从而提高模型的鲁棒性和泛化能力。

1. Optimizer

本次实验所采用的优化器为Momentum，RMSprop和Adam。

对于学习率的优化是必要的：如果learning rate选的过小，收敛速度会很慢，如果过大，loss function就会在极小值处不停地震荡甚至偏离。对于非凸函数，还要避免陷于局部极小值处。

若采取措施：先设定较大的学习率，当两次迭代之间的变化低于某个阈值后，就减小learning rate。其可以在一定程度上缓解上面问题，但是需要提前设定阈值，不能够适应数据集的特点。

因之需要引入学习率优化器：

（1）Momentum通过下面公式：

使得在梯度不变的维度上速度变快，梯度方向有所改变的维度上的更新速度变慢，达到加快收敛并减少震荡的效果。

（2）RMSprop

利用了二阶导数信息，解决学习率的设置导致的梯度爆炸问题。它的更新规则是基于当前梯度的平方的指数加权移动平均数来跟新权重，同时除以平方根。这个操作可以让权重在更新时的缩放更为平滑，因此更新的步长会更加稳定，即减少随机梯度下降算法中的震荡。

1. **总结**

总结：

要求说明：在提交实验报告时请**将代码以附件形式随报告一起提交**

**实验报告的格式无限制，页数限制在10页内**