机器学习第二次作业

1.分类问题与聚类问题有什么区别和联系？

**区别：**

**任务目标与学习方式：分类问题的目标是将数据点划分到预定义的类别中，每个类别具有明确的标签,所以分类问题是一种监督学习任务。而聚类问题的目标是将数据点分组到未知的类别中，不需要预先定义类别标签，所以聚类问题是一种无监督学习任务。**

**输出：分类问题的输出是预测的类别标签，可以是离散的类别值或概率分布。而聚类问题的输出是对数据点的分组，通常用簇中心或簇标识来表示。**

**评估指标：分类问题通常使用准确率、精确率、召回率等指标来评估模型性能。而聚类问题的评估通常使用聚类内部距离和聚类间距离等指标来评估聚类的紧密度和分离度。**

**联系：**

**数据探索：聚类可以用于数据探索和发现数据的内在结构，有助于后续的分类任务。例如，可以先对数据进行聚类分析，然后再基于聚类结果进行分类建模。**

**特征工程：聚类可以用于特征工程，将原始特征转换成更具代表性的特征，然后再应用于分类任务。例如，可以使用聚类结果生成新的特征，然后将这些新特征用于分类模型的训练。**

**无监督预处理：聚类可以作为无监督预处理步骤，帮助发现数据中的模式和规律，从而提高分类任务的性能。例如，在分类之前先对数据进行聚类，然后再基于聚类结果进行有监督学习。**

2.解决分类问题可以使用哪些算法，请罗列5种以上。

**解决分类问题可以使用的算法：**

**1.KNN算法**

**2.决策树算法**

**3.随机森林算法，XGBOOST算法**

**4.朴素贝叶斯方法**

**5.支持向量机算法**

**6.逻辑回归算法**

**7.神经网络与深度学习算法**

3.对于人脸识别问题，请问用k近邻（KNN）算法、随机森林算法、支持向量机算法各自的特点有什么？

**KNN算法的特点：**

**通过计算待分类样本与训练样本之间的距离，然后选择最近的K个样本进行投票来确定待分类样本的类别。**

**易于理解和实现，不需要训练过程，适用于小规模数据集，但在处理大规模数据集时计算量较大，并且对数据中的噪声和异常值敏感。**

**随机森林算法的特点：**

**通过组合多个决策树来进行分类或回归任务。**

**具有较高的准确性和泛化能力，对于高维数据和大规模数据集具有良好的适应性，能够处理缺失值和不平衡数据；但容易产生过拟合问题，需要不断调节参数来优化模型性能。**

**支持向量机算法的特点：**

**通过找到能够将不同类别分隔开的最优超平面来进行分类，可以使用不同的核函数来处理非线性分类问题。**

**在高维空间中具有较好的分类性能，对于小样本数据和非线性数据具有较强的泛化能力，能够有效处理特征维度高于样本数量的情况，但对大规模数据集和噪声数据较为敏感，训练时间较长。**

**所以可以根据所要人脸数据的数据集规模、特征维度和具体需要识别的人脸规模来确定选用的算法。**

4.什么是过拟合（过学习）问题？对于决策树算法，解决过拟合有哪些策略？对于神经网络，解决过拟合有哪些策略？

**对于一个假设，当存在其他的假设对训练样例的拟合比它差，但事实上在实例的整个分布上表现得却更好 时，我们说这个假设过度拟合训练样例。**

**对于决策树算法，解决过拟合的策略有：**

**1.阈值方法：通过设置一个阈值来限制树的深度或者叶子节点数，及早停止树生长。**

**2.剪枝方法：通过减去决策树的一部分节点来降低模型复杂度，提高泛化能力。**

**3.验证集法： 将训练样本分成训练集和确认集，使用训练集训练决策树，确认集评估决策树的精度，并在通过确认集的测试来选择最优的决策树。**

**对于神经网络，解决过拟合的策略有：**

**1.正则化：在损失函数中添加正则化项来惩罚模型复杂度，防止权重过大导致过拟 合。**

**2.在训练过程中监控验证集的性能，当验证集性能不再提升时停止训练，避免模型 在训练集上过拟合。**

**3.增加样本数量，或者可以使用数据增强对训练数据进行随机变换或扩增来增加数 据的多样性，从而提升泛化能力**

5.什么是集成学习？集成学习常见的算法有哪些，这些算法的基本思想是什么？

**集成学习是一种通过组合多个基学习器来构建强学习器的机器学习方法。其基本思想是通过结合多个模型的预测结果，从而达到更好的泛化性能和预测能力。**

**集成学习常见的算法有：**

**1.Bagging（自举汇聚法）：**

**基于自助采样的思想，通过随机从训练集中有放回地抽取样本来构建多个子模型，然后通过投票或平均来获得最终预测结果。常见的算法包括随机森林算法等。**

**2.Boosting（提升法）：**

**通过迭代训练多个弱学习器，每次训练都会根据前一次模型的表现调整样本权重或模型参数，使得后续的模型更关注前一次训练中被错误分类的样本。常见的算法包括Adaboost、Gradient Boosting Machine（GBM）、XGBoost和LightGBM等。**

**3.Stacking（堆叠法）：**

**将多个不同模型的预测结果作为新特征，然后将这些新特征作为输入训练一个元模型，从而组合多个模型的预测能力。堆叠法可以结合各种类型的模型，如线性模型、非线性模型等。**