# 课程项目二 分类、集成模型及其应用

## 【1】项目背景

在回归模型中假设响应变量是定量的, 但很多情况下响应变量却是定性的. 例如, 眼睛的颜色是定性变量, 取值蓝色、棕色或绿色. 定性变量也称为分类变量, 两者的统计含义是一样的. 预测一个观测的定性响应值也指对观测分类, 因为它涉及将观测分配到一个类别中. 一些分类方法先从预测定性变量不同类别的概率开始, 将分类问题作为概率估计的一个结果. 从这个角度上看, 分类与回归方法有许多类似之处.

《数据科学导论》课程中已习得几种分类方法, 如逻辑回归、近邻、决策树、支持向量机、随机森林和AdaBoost等. 现有分类器中, 支持向量机(support vector machine, SVM)是一种起源于机器学习社区的流行方法, 它是使用单一分类函数对二分类问题进行最大间隔分类的典型例子, 在实际应用中性能良好. 在很多问题中, 数据通常包含两个以上的类别. 尽管SVM在二分类上取得了成功, 但如何将其应用于多类别分类问题仍然是一个挑战. 文献中使用SVM进行多分类的方法可分为两大类: 第一类方法先训练一系列的二值支持向量机, 再将结果组合起来进行多类别分类. 例子包括一对一(one-versus-one)和一对多(one-versus-rest)方法. 尽管这种方法在概念和实现上都很简单, 但它也有自身的缺点. 第二类方法是在一个优化问题中同时考虑所有个类. 常用的方法是用个分类函数来表示对应优化中的个类, 并根据哪个分类函数最大给出最终的预测结果. 目前, 在第二类框架下提出了许多同步多分类支持向量机分类器. 本项目重点关注支持向量机与其他方法的比较, 以及如何使用SVM处理多分类问题.

## 【2】方法陈述

**问题一:** 请对现有的种分类方法进行总结, 试搜寻每类方法的优缺点(SVM需要着重介绍).

1. **SVM支持向量机**

SVM优缺点总结分析：

**优点：**

1.高维数据处理能力

**理论基础**：在高维空间中，不同类别的数据点更有可能是线性可分的。SVM 通过构造一个在高维空间中有效工作的决策边界，可以有效地处理高维数据。

**实际应用**：这使得 SVM 在处理具有大量特征的数据集时非常有效，例如在文本分类或生物信息学中，其中每个单词或基因可以视为一个维度。

**特征转换**：即使在原始特征空间中数据不是线性可分的，SVM 的核技巧也允许它通过映射到高维空间来找到可分的边界。

2. 泛化能力强

**最大化边界**：SVM 的核心思想是最大化不同类别之间的边界，这种策略减少了模型对特定样本的依赖，从而提高了对新数据的泛化能力。

**过拟合风险降低**：与仅尝试最小化训练误差的方法相比，SVM 通过其正则化特性（控制边界的大小）减少了过拟合的风险。

3. 灵活性

**核函数的选择**：SVM 通过不同的核函数适应不同类型的数据。核函数（如线性核、多项式核、RBF核）允许 SVM 在不同的数据分布中找到复杂的非线性关系。

**适应不同数据分布**：这种灵活性使得 SVM 可以用于各种实际问题，从简单的线性可分问题到复杂的非线性分类任务。

4. 由支持向量决定的模型

**模型简洁性**：SVM 的决策函数主要由支持向量决定，这些是距离决策边界最近的点。这意味着模型的复杂度不是由数据的维度，而是由支持向量的数量决定。

**效率和稳健性**：这种依赖于边界附近的少数点使得模型在预测时更加高效且稳健，尤其是在处理大型数据集时。

5. 适用于复杂决策边界

**处理复杂模式**：通过将数据映射到高维空间，SVM 能够学习复杂的决策边界，这对于很多现实世界的问题是必要的。

**适用性广泛**：这种能力使得 SVM 在各种数据集上都能找到有效的分类规则，包括那些在原始特征空间中表现出复杂模式的数据集。

**缺点：**

1. 参数选择和调整复杂

**核函数和正则化参数**：正确选择核函数（如线性核、RBF核等）和调整其参数（如C值、γ值）对SVM的性能至关重要。错误的选择可能导致模型过拟合（对训练数据过于敏感）或欠拟合（不能很好地捕捉数据特征）。

**专业知识要求**：理解不同核函数的特性以及如何调整它们的参数通常需要深入的专业知识和经验。

2. 不直接提供概率估计

**默认输出**：SVM默认只提供类别预测（即数据点属于哪个类别），而不提供概率估计（即数据点属于每个类别的概率是多少）。

**概率输出需额外计算**：虽然可以通过一些方法（如 Platt 缩放）来计算概率估计，但这需要额外的计算，并且可能不如直接以概率输出为目标的算法准确。

3. 计算和存储资源要求高

**大规模数据集**：在大规模数据集上，尤其是在使用复杂核函数时，SVM的训练过程可能非常耗时和资源密集。

**内存使用**：SVM在训练过程中可能需要较大的内存，尤其是在核矩阵较大的情况下。

4. 结果解释性差

**黑盒模型**：与决策树等模型相比，SVM的决策过程更难以解释，通常被视为“黑盒”模型。

**实际应用局限**：在需要对模型做出详细解释的应用场景（如医疗诊断）中，这种缺乏透明度可能是一个重要的限制。

5. 对数据预处理和缩放敏感

**数据预处理**：SVM对数据的缩放非常敏感，特别是在使用RBF等核函数时。不同特征的缩放不一致可能导致模型偏向于某些特征。

**标准化的重要性**：因此，通常需要对数据进行归一化或标准化处理，以确保所有特征对模型的贡献是均衡的。

6. 不适合处理大规模数据集

**训练时间**：SVM在处理大规模数据集时，由于计算复杂度和内存需求的原因，其训练时间可能会非常长。

**实用性降低**：这限制了其在大数据时代一些实时或近实时应用的实用性。

总结：支持向量机（SVM）是一种强大的机器学习算法，特别擅长处理高维数据，具有出色的泛化能力和灵活性，适用于各种线性和非线性分类问题。然而，它在参数选择和调整方面具有一定复杂性，对数据预处理和缩放敏感，不直接提供概率估计，且对大规模数据集的处理效率较低，且其结果的解释性相对较差。因此，在选择SVM时，需要根据具体问题和数据特性仔细考虑这些优缺点。

总结：

支持向量机（SVM）是一种在各种数据集上表现卓越的机器学习算法，尤其擅长处理高维数据。它的核心优点包括出色的泛化能力，这意味着它能够很好地处理未见过的数据；对特征空间维度的高容忍度，使其在特征数量多的情况下依然有效；以及通过不同的核函数（如线性核、多项式核、RBF核）处理线性和非线性问题的能力。SVM的模型基于最大化数据集类别之间的边界，因此通常能实现较高的分类准确率，特别是在二分类问题上。此外，SVM提供了一种有效的方法来处理过拟合，尤其是在样本数量较少的情况下。

然而，SVM也有其局限性和缺点。首先，对于大规模数据集，SVM的训练时间可能会非常长，这使得它在处理大型数据集时效率不高。其次，SVM模型对参数的选择（如正则化参数C、核函数参数）非常敏感，不当的参数设置可能会导致模型性能不佳。此外，由于SVM在优化过程中涉及复杂的数学计算，其模型结果的解释性较差，不像决策树那样直观。而且，SVM在原生形式下不提供概率估计，这在某些需要概率输出的应用场景中可能是一个缺陷。此外，SVM对数据预处理（如特征缩放）也非常敏感，需要在应用之前进行适当的数据标准化或归一化处理。

综上所述，SVM是一种功能强大但复杂的模型，适用于特征维度高且数据量适中的问题，但在应用时需要仔细考虑参数选择、数据预处理和模型解释性等因素。

1. **随机森林模型**

**优点：**

**1. 高准确度和抗过拟合能力**

随机森林通过集成多个决策树，可以显著提高预测准确性。每棵树独立训练，通过对它们的预测结果进行平均或多数投票，可以减少模型的偏差和方差。

相比单个决策树，随机森林在处理各种数据集时更不容易过拟合。这是因为它通过引入随机性，降低了模型对特定样本的依赖，增强了模型对新数据的泛化能力。

**2. 适用性广泛**

随机森林可以处理各种类型的数据，包括数值型和类别型变量，无需进行复杂的数据预处理，如特征缩放或归一化。

它既适用于分类问题，也适用于回归问题，使其成为多种应用场景的强大工具。

**3. 特征重要性评估**

随机森林能够评估出哪些特征对模型预测最为重要。这一点对于特征选择和理解数据结构尤其有价值。

**4. 容错性和鲁棒性**

由于其集成的性质，随机森林对单个决策树中的错误更具容忍性。即使其中一些树的表现不佳，整体模型仍然能够维持较高的准确率。

对于数据集中的噪声和异常值，随机森林展现出较好的鲁棒性。

**5. 并行化能力**

随机森林的训练过程可以被高效地并行化，因为每棵树的构建是独立的。这使得随机森林在处理大型数据集时能够有效利用现代计算资源。

**6. 易于使用和参数调整较少**

随机森林通常不需要复杂的参数调整即可达到不错的效果。它的主要参数如树的数量和树的深度相对容易设置，且模型对这些参数的选择不太敏感。

**7. 处理不平衡数据的能力**

在处理类别不平衡的数据时，随机森林通过调整类权重或采样策略可以有效改善模型的表现。

**缺点：**

**1. 模型解释性差**

相比于单个决策树，随机森林由于涉及多个树的集成，其结果通常更难解释。这在需要明确理解模型每一步决策的应用中可能是一个问题。

**2. 性能对参数敏感**

虽然随机森林通常不需要太多的参数调整，但某些参数（如树的数量、树的深度和节点分裂时考虑的特征数量）对最终模型的性能还是有显著影响。

**3. 大规模数据下的计算和内存需求**

在处理非常大的数据集时，随机森林模型可能需要较长的训练时间和较大的内存。尽管可以通过并行处理来缓解，但这在资源有限的环境下仍然是一个挑战。

**4. 对不平衡数据处理不佳**

在处理类别极不平衡的数据时，随机森林可能倾向于偏向多数类。这可能需要额外的技术来平衡数据或调整模型。

**5. 预测过程可能较慢**

由于需要整合所有决策树的预测结果，随机森林在预测阶段可能比某些更简单的模型慢。

**6. 对噪声和异常值敏感**

尽管随机森林通常对异常值具有一定的抵抗力，但在特征包含大量噪声或异常值的情况下，模型的性能可能会受到影响。

**总结：**随机森林是一种在多种数据集上表现出色的机器学习算法，它通过集成多个决策树提高了预测的准确性和鲁棒性，抵抗过拟合，并且能有效处理高维数据和不平衡数据。同时，它提供特征重要性的评估，易于使用，且对参数调整不太敏感。然而，随机森林在处理大规模数据时可能会有性能挑战，且由于其集成性质，其模型解释性较差，对噪声和异常值也相对敏感。

1. **K-近邻算法**

**优点：**

**1. 简单且直观**

KNN是一种非常直观的算法，基本思想是“物以类聚”，即一个样本将与其最近的K个邻居归为同一类别。这种方法易于理解和实现。

**2. 训练成本低**

KNN是一种基于实例的学习，不需要像其他算法那样进行显式的训练过程。它在使用时直接进行分类或回归，因此，算法的“训练”阶段开销几乎为零。

**3. 适应性强**

KNN不假设数据的分布，因此在某些非线性和复杂分布的数据集上表现良好。

**4. 易于添加新数据**

由于KNN不需要训练模型，所以新增数据时不需要重新训练，只需将新数据点添加到数据集中即可。

**5. 功能性多样**

KNN既可以用于分类问题，也可以用于回归问题，只需调整对邻居的投票或平均的方式即可。

**6. 对异常值的鲁棒性**

如果选择的K值适当，KNN可以对异常值和噪声数据具有一定的鲁棒性。

**7. 特征空间的解释性**

KNN在特征空间中的决策边界通常容易解释，特别是在特征维度较低的情况下。

**缺点：**

**1. 效率问题**

计算密集型：在预测阶段，KNN需要对每个测试实例计算与所有训练实例的距离，这在大数据集上尤其耗时。

存储密集型：由于KNN是一种基于实例的学习，需要保留整个训练数据集，这在大数据集上可能会导致存储问题。

**2. 高维效果不理想**

KNN在处理高维数据时性能通常不佳。随着维度的增加，计算每个维度间距离的成本显著增加，而且高维空间中的数据点可能会在距离上变得越来越“稀疏”。

**3. 对不平衡数据敏感**

在处理类别不平衡的数据时，KNN可能倾向于过度代表多数类。如果一个类的样本数量远多于其他类，那么新的样本更有可能被划分到这个多数类。

**4. 对噪声和异常值敏感**

尽管适当选择K值可以提高对噪声的鲁棒性，但KNN仍然对异常值和噪声较为敏感，尤其是当K值较小时。

**5. 调参较复杂**

确定最佳的K值有时是一个挑战，并且K值的选择可能对预测结果有显著影响。

此外，选择适当的距离度量（如欧几里得距离、曼哈顿距离等）也是实现高效KNN模型的关键。

**6. 特征缩放敏感性**

KNN对特征的缩放非常敏感。如果不同的特征具有不同的量级，则需要先进行标准化或归一化，否则较大量级的特征可能会主导距离计算。

**总结**：K-近邻（KNN）算法是一种简单直观且易于实现的机器学习方法，它适用于分类和回归问题，对小型数据集效果良好，无需训练过程，易于添加新数据，且对异常值具有一定的鲁棒性。然而，KNN在处理大型数据集时效率较低，对高维数据和不平衡数据敏感，对噪声和异常值的处理能力有限，且对特征缩放非常敏感。此外，正确选择邻居数量（K值）和距离度量对模型性能影响显著，这些都是使用KNN时需要考虑的关键因素。

1. **朴素贝叶斯**

**优点：**

**1. 简单性和高效性**

朴素贝叶斯模型的原理简单，易于理解和实现。它的算法结构直观，使得在训练和预测时的计算效率都非常高。这种简洁性使得朴素贝叶斯特别适合于大数据集。

**2. 低数据需求**

相对于需要大量数据的复杂算法（如深度学习），朴素贝叶斯能够在较少的数据上仍然获得不错的效果。这使得它在数据受限的情况下仍然是一个有效的选择。

**3. 良好的多类别分类性能**

朴素贝叶斯在处理多类别问题时表现良好，能够提供每个类别的概率估计，这对于需要区分多个可能类别的问题非常有用。

**4. 适合高维数据**

在处理具有大量特征的数据集时，如文本分类，朴素贝叶斯能够有效地工作。它可以处理成千上万的输入变量而不影响性能，这在其他算法中可能是个挑战。

**5. 对缺失数据的容忍度**

朴素贝叶斯模型可以处理缺失数据的情况。由于其独立性假设，每个特征的贡献是分开考虑的，因此缺失某些数据不会对整个模型造成太大影响。

**6. 基于概率的决策框架**

该模型提供了基于概率的决策框架，使得结果的解释和理解变得更加直接。这种明确的概率输出可以帮助更好地理解模型的预测。

**7. 模型更新简单**

当有新数据时，朴素贝叶斯模型可以轻松地进行更新，无需重新训练整个模型。这使得它适用于那些数据频繁更新的应用场景。

**缺点：**

**1. 特征独立性假设**

朴素贝叶斯模型的主要局限在于它的“朴素”假设——即假设所有特征在给定类别的条件下是相互独立的。在现实世界的数据中，这种假设往往过于简化，特征之间常常存在某种程度的依赖。

**2. 对数据分布的假设**

该模型对输入数据的分布类型（如高斯、多项式或伯努利分布）做出了特定的假设。如果实际数据的分布与模型假设不符，可能会影响模型的性能。

**3. 对输入数据的质量敏感**

朴素贝叶斯对数据质量敏感。数据中的噪声和不相关的特征可能会对模型的预测性能产生负面影响。

**4. 概率估计可能不准确**

当特征空间很大时，尤其在文本分类等应用中，概率估计可能因为数据稀疏性而变得不准确。

**5. 对不平衡数据的处理**

在处理类别极不平衡的数据时，朴素贝叶斯可能会对多数类别产生偏见。

**6. 难以捕捉复杂关系**

由于其简单性，朴素贝叶斯通常无法捕捉数据中复杂的关系，这在某些需要捕捉复杂模式的应用中可能是一个限制。

**7. 预测性能的局限性**

尽管朴素贝叶斯在很多情况下都能提供不错的结果，但其性能通常不如更复杂的方法，特别是在数据量大和特征关系复杂的情况下。

**总结：**

朴素贝叶斯（Naive Bayes）模型是一种基于概率的简易分类算法，以其高效性、处理高维数据的能力、对小数据集的适应性、以及对缺失数据的容忍度而受到青睐。它特别适合于文本数据处理，如垃圾邮件识别和情感分析。然而，其性能受限于特征独立性的“朴素”假设，这在现实世界的数据中往往不成立。此外，朴素贝叶斯对数据分布类型的假设可能限制了其在某些情况下的准确性，且在处理类别不平衡的数据和捕捉复杂特征关系方面可能表现不佳。因此，在应用朴素贝叶斯模型时，需要考虑数据的特点和实际问题的复杂性。

**问题二:** 尝试借助课程及网络, 检索SVM及其变体模型, 并对其相关机理进行论述.

1.SVM（支持向量机）

支持向量机（SVM）的核心机理是在特征空间中寻找一个最优超平面，以便最大化分类的间隔。这个超平面的选择基于间隔最大化的原则，即最大化数据点到决策边界的最小距离。以下是对SVM的主要机理的详细论述：

1.1超平面和支持向量

**超平面**：超平面是SVM进行数据分类的决策边界。在二维空间中，这个超平面就是一条线；在更高维度的空间中，它成为一个多维面。SVM的目标是找到最能代表数据分布的超平面。在特征空间中，超平面可以表示为 ：



其中 w 是超平面的法向量，b 是偏置项。

**支持向量**：支持向量是离决策边界最近的那些点。这些点对于定义边界和间隔是至关重要的。SVM的优化过程就是确保这些支持向量离超平面的间隔最大化。

1.2间隔最大化

**间隔的定义**：间隔是数据点到超平面的距离。在SVM中，间隔是分类决策的关键。最优超平面是使得离平面最近的点（支持向量）的间隔最大的那个平面。SVM通过最大化间隔来优化和。间隔计算为



最大化间隔实际上是最小化的问题。

**为何重视间隔**：最大化间隔有助于提高分类器的泛化能力。一个较大的间隔意味着对于未见示例的分类更为稳健，因为它提供了更大的容错空间。

1.3优化方法

**求解优化问题**：找到最优超平面涉及到求解一个优化问题，即最小化与超平面法向量相关的范数的平方。这个优化问题受到每个数据点分类正确的约束。

**拉格朗日乘子法**：SVM通过拉格朗日乘子法将原始的优化问题转化为对偶问题。通过引入拉格朗日乘子，上述问题可以转化为拉格朗日函数：



其中，是每个约束的拉格朗日乘子。

而通过对和的偏导数置零，可以得到优化问题的对偶形式，这通常是更易解的问题。对偶问题的目标是最大化关于 的函数：





这种方法不仅简化了问题的求解，还揭示了只有支持向量对最终解有贡献这一重要性质。

1.4核函数

**处理非线性数据**：现实世界中的很多数据是非线性可分的。SVM通过引入核函数，核函数替换了原始特征的点积，将原始数据映射到更高维的空间，使得在这个新空间中数据变得线性可分。

**常见核函数**：包括线性核、多项式核、径向基函数（RBF）等。选择哪种核函数取决于数据的特性和分布。常见核函数如下：

**线性核**：，

**多项式核：，**其中是多项式的度数，是可调参数。

**径向基函数（RBF）核：**，其中是可调参数。

1.5总结

支持向量机（SVM）的作用机理主要基于在特征空间中寻找最优超平面，以最大化不同类别数据点之间的间隔。通过这个超平面，SVM有效地将数据分类。在处理非线性可分数据时，SVM利用核函数将数据映射到更高维的空间，使得原本在低维空间中线性不可分的数据在高维空间中变得可分。其核心在于只有最靠近超平面的数据点（即支持向量）对决策边界的定义起决定性作用，这提高了模型的泛化能力并降低了计算复杂度。由于这些特性，SVM在许多领域如图像处理、文本分类等都有广泛应用。

2. SVM变体

2.1. 一对一（One-vs-One）方法

作用机理：

**基本概念**：在一对一方法中，对于个类别的数据集，会创建个独立的SVM分类器，每个分类器只在两个类别之间做决策。

**训练过程**：在训练时，每个分类器被训练以区分一对类别。例如，如果有三个类别A、B和C，将会训练三个分类器：一个用于区分A和B，一个用于区分A和C，另一个用于区分B和C。

**分类决策**：在进行分类时，每个分类器对新样本给出一个投票，最终的类别判定是基于哪个类别获得最多票数。这意味着，对于每个测试样本，SVM模型将运行多次，每次在两个类别之间做出决策。

**应用场景**：一对一方法适用于类别数量不是特别多的多分类问题。随着类别数的增加，需要训练的分类器数量会呈二次方增长，这可能导致计算和存储成本显著增加。

2.2一对其余（One-vs-Rest）方法

作用机理：

**基本概念**：一对其余（One-vs-Rest，也称为One-vs-All）方法是另一种多分类策略。在这种方法中，为每个类别训练一个单独的SVM分类器，将这个类别与所有其他类别区分开。

**训练过程**：对于每个类别，训练一个SVM分类器，将该类别的样本作为正类，所有其他类别的样本合并作为负类。这样，如果有个类别，就会有个分类器。

**分类决策**：在分类新样本时，每个分类器都会对样本进行评分，通常是基于距离超平面的距离。样本最终被分类到得分最高的类别。

**应用场景**：一对其余方法适用于类别数量较多的情况。相比一对一方法，它仅需要训练与类别数相等的分类器，因此在有很多类别的情况下更高效。

2.3有向无环图SVM（Directed Acyclic Graph SVM, DAG SVM）

作用机理：

**基本概念**：有向无环图SVM（DAG SVM）是一种特殊的SVM变体，用于多分类问题。它结合了一对一方法的优点，并优化了分类过程中的决策步骤。

**构建DAG**：在DAG SVM中，首先构建一个有向无环图，每个节点代表一个二分类决策。对于个类别，图中会有个节点，与一对一方法中的分类器数量相同。

**决策过程**：在分类新样本时，过程从DAG的顶点开始，根据每个节点的决策指向下一个节点。每一步决策都会排除一个类别，直至最后到达叶节点，其标记为最终分类结果。

**应用场景**：DAG SVM适用于类别数较多的情况。由于它在分类时比一对一方法减少了必要的比较次数，因此在分类效率上更有优势。

2.4 错误校正输出编码（Error-Correcting Output Codes, ECOC）

作用机理：

**基本概念**：错误校正输出编码（ECOC）是一种将多分类问题转换为多个二分类问题的方法。它通过构建一系列的二进制代码（或“编码”）来表示每个类别，每个代码对应于一系列的二分类任务。

**编码过程**：在ECOC中，每个类别被赋予一个唯一的二进制代码。例如，在一个四类问题中，可能为每个类别分配如“001”、“010”、“100”、“111”等代码。

**分类器训练**：对于每个二进制位，训练一个二分类SVM。这些分类器共同工作，以确定测试样本的编码。

**解码过程**：在分类新样本时，通过运行所有的二分类器来为样本生成一个代码，然后选择与生成的代码最接近的类别编码。

**应用场景**：ECOC特别适用于大规模的多分类问题，尤其是当类别之间存在一定的层次结构或关系时。它通过引入冗余和错误校正能力，提高了分类的准确性和鲁棒性。

## 【3】案例实战

**1. 教材案例回顾:** **乳腺癌诊断**

早期的乳腺癌检测主要检查乳腺组织的异常肿块. 如果能利用机器学习算法, 通过乳腺肿块的检测数据自动进行诊断, 将会给医疗系统带来很大的益处: 不仅能够大大提高检测效率, 还可以降低误判的风险.

数据集共有569个样本, 每个样本包含乳腺细胞的30个特征, 这30个特征都是由数字化细胞核的10个基础特征(包括半径、质地、周长、面积、对称性等) 的均值、标准差及最大值构成. 请使用分类方法(包含SVM), 对乳腺癌数据进行分类, 并选择适当的指标对最终的分类结果进行评价.

注: 评价指标可参考教材附录E.

**数据结果**：

数据集没有任何缺失值，并包括各种特征和诊断标签，这是模型建立的一个良好迹象。

良性（B）：357例

恶性（M）：212例

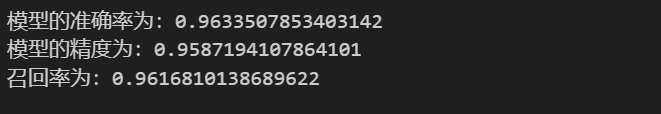
尽管良性病例数量更多，但数据集并不是极度不平衡的。然而，在训练和评估模型时，仍然需要考虑这种不平衡的影响，因为它可能会影响精确度、召回率和准确率等性能指标。在这类医疗诊断任务中，对恶性类别高召回率至关重要，因为未能检测到实际癌症病例可能比误报更有害。

对于乳腺癌数据集的分类结果，我们有四种模型的评估指标：支持向量机（SVM）、朴素贝叶斯（Naive Bayes）、随机森林（Random Forest）和K最近邻（KNN）。每种模型都有其召回率、准确率和精度。

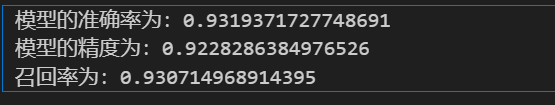
SVM表现最佳，具有最高的准确率（97.37%）和精度（97.56%），召回率也很高（95.24%）。这表明SVM对于乳腺癌数据集的分类是非常有效的，能够以高准确度识别出正类（如诊断为乳腺癌的病例）。



KNN的表现紧随其后，具有最高的召回率（96.17%）和非常高的准确率（96.34%）和精度（95.87%）。这表明KNN在找回所有正类病例方面非常有效，同时保持了高准确性和高精度。



随机森林的表现逊色于SVM和KNN，但仍然有着不错的召回率（93.07%）、准确率（93.19%）和精度（92.28%）。这表明随机森林是一个坚实的选择，虽然在这些指标上略低于SVM和KNN。



朴素贝叶斯的表现是四种模型中最低的，召回率为90.48%，准确率为92.98%，精度也为90.48%。尽管如此，这些数值仍然表示朴素贝叶斯是一个可行的模型，尤其是在计算资源有限或需要快速原型开发时。



**分析模型结果差异的原因：**

支持向量机（SVM）：SVM的优异表现可能是由于其能有效处理高维空间和复杂的决策边界，特别是当使用合适的核函数时。乳腺癌数据集中的特征可能构成了明显的分类边界，使得SVM能够以较高的精度和召回率进行分类。

K最近邻（KNN）：KNN的高召回率可能归因于其直观的分类方式—通过观察最近邻的样本。如果恶性和良性病例在特征空间中形成相对明显的簇，KNN将能够有效地识别出这些簇，尤其是在选择适当的邻居数量（K值）后。

随机森林（Random Forest）：随机森林通过构建多个决策树并综合它们的预测来做出决策，通常对于各种数据都有不错的适应性和稳定性。它在乳腺癌数据集上的表现可能没有SVM和KNN那么出色，可能是因为单个决策树的性能限制或者模型没有完全捕捉到所有复杂的特征关系。

朴素贝叶斯（Naive Bayes）：朴素贝叶斯的表现可能不如其他模型，部分是因为它基于特征间独立的假设。在实际的乳腺癌数据中，很多特征之间可能存在相互依赖或相关性，这使得朴素贝叶斯的简化假设在这里不那么有效。

**总结：**

在乳腺癌诊断中，非常重视不遗漏任何一个阳性病例，那么可能会更偏好召回率更高的模型。而在某些需要非常精确诊断的场合，可能会更看重精度或准确率。总体来说，SVM和KNN在大多数指标上表现优异，可能是更可取的选择。不过，每种模型都有其优势和局限，选择哪种模型还要基于具体的应用需求和数据特性进行综合考量。

**2. 请将问题1中的分类器应用至下述数据集(二选一)**

(1) 教材第四、五章中其余6个数据任选三个

(2) Gas、Optical、Vertebral数据集(数据集见附件二)

强调分析不同分类器在不同数据集下的分类结果.

数据集与评价指标选取：

案例二本小组选取的数据集为光学字符识别数据集、个人信用风险评估数据集、玻璃类型识别数据集。其中个人信用风险评估为二分类问题，光学字符识别与玻璃类型识别为多分类问题。

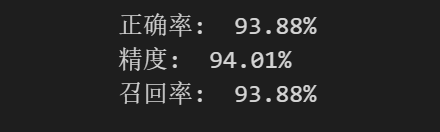
对于不同的分类问题本小组采取的模型效果评估方式如下：对二分类问题，评价指标为准确率（accuracy）、精度（precision）、召回率（recall）。对多分类问题，评价指标为准确率（accuracy）、宏平均精度（marco precision）、宏平均召回率（marco recall）。

模型分类结果说明：

光学字符识别数据集：

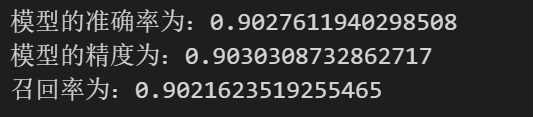
光学字符识别要求根据20000个样本各自包含的16个输入特征构建识别光学字符的分类器。各分类器的表现结果如下：

SVM：



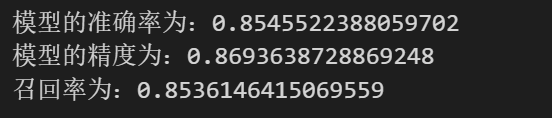
SVM模型在惩罚系数为1、核函数为‘rbf’的参数下对光学字符识别的正确率为93.88%，精度为94.01%，召回率为93.88%,可见svm模型对光学字符识别的效果非常好，能够正确地识别大部分光学字符。

KNN：



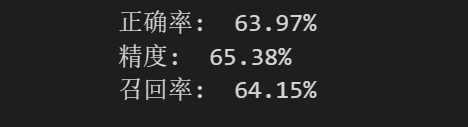
KNN模型在k值为2的参数下对光学字符识别的正确率为90.28%，精度为90.3%，召回率为90.22%，整体的模型分类效果较好。

随机森林：



随机森林模型对光学字符识别的正确率为85.46%，精度为86.94%，召回率为85.36%，整体的表现也不错，但略逊于SVM模型与KNN模型

朴素贝叶斯：

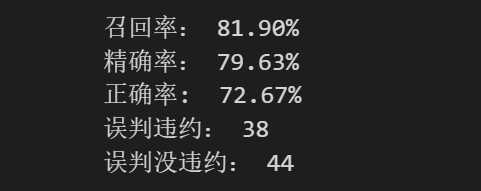


朴素贝叶斯模型对光学字符识别的正确率为63.97%，精度为65.38%,召回率为64.15%，是四种分类模型中表现最差的，整体的指标也较差。

个人信用风险评估数据集：

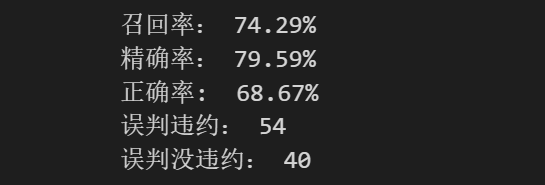
个人信用风险评估要求根据1000个贷款样本及其包含的20个相关特征对贷款是否违约进行分类。各模型的分类效果如下：

SVM：



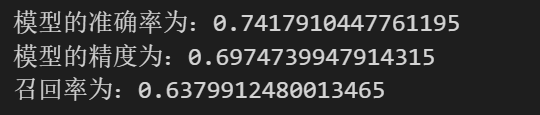
SVM模型在惩罚系数为10，核函数为‘rbf’的参数下对个人信用风险评估分类的正确率为72.67%，精确率为79.63%，召回率为81.9%，SVM模型分类的正确率并不高，但却有着不错的召回率和精确率，对正类样本的分类效果较好。

朴素贝叶斯：



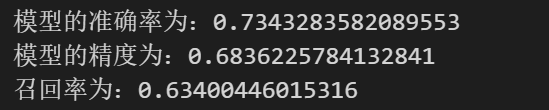
朴素贝叶斯模型对个人信用风险评估分类的正确率为68.67%，精确率为79.59%，召回率为74.29%，各类指标并不高，分类效果略逊色于SVM模型，但也有较不错的召回率和精确率。

随机森林：



随机森林模型对个人信用风险评估分类的正确率为74.18%，精确率为69.75%，召回率为63.8%，随机森林模型的正确率是四种模型中最高的，对个人信用风险评估分类结果较好，但从精度和召回率来看，随机森林模型对正类样本的分类效果较差。

KNN：

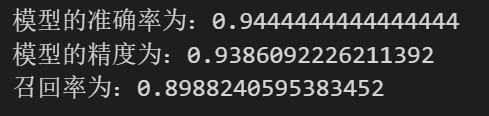


KNN模型在k值为9的参数下对个人信用风险评估分类的正确率为73.43%，精确率为68.36%，召回率为63.4%，整体的指标与随机森林模型非常相近，也有着较高的正确率和偏低的精度和召回率。

玻璃类型识别数据集：

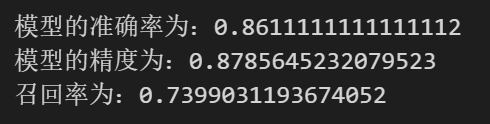
玻璃类型识别要求根据214个样本及其包含的9个特征对玻璃的7个类型进行分类。各模型的分类效果如下：

随机森林：



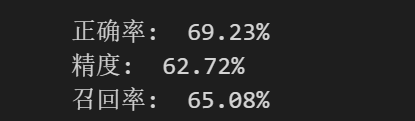
随机森林模型对玻璃类型识别的正确率为94.44%，精确率为93.86%，召回率为89.88%，各项指标均很高，随机森林模型对玻璃类型的分类效果非常好，能够正确分类大部分的样本。

KNN：



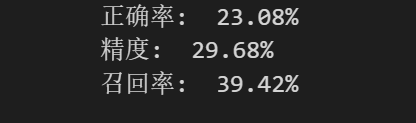
KNN模型在k值为4的参数下对玻璃类型识别的正确率为86.11%，精确率为87.86%，召回率为73.99%，，正确率和精确率也较高，KNN模型对玻璃类型的分类效果略差于随机森林模型。

SVM模型：



SVM模型在惩罚系数为10，核函数为‘rbf’的参数下对玻璃类型识别的正确率为69.23%，精确率为62.72%，召回率为65.08%，整体的指标均偏低，在玻璃类型识别的分类任务下，SVM模型的表现效果并不好。

朴素贝叶斯：



朴素贝叶斯模型对玻璃类型识别的正确率为23.08%，精确率为29.68%，召回率为39.42%，各类指标均是四类模型中最低的，而且朴素贝叶斯对玻璃类型识别的效果非常差。

总结：

在不同数据集中，朴素贝叶斯的表现效果在四类模型中并不突出，或许是由于朴素贝叶斯模型使用了样本属性独立性假设，在样本属性有关联性时的模型效果并不好。

SVM模型在处理大样本的多分类任务，即光学字符识别的效果很好，而对于小样本多分类任务，即玻璃类型识别时的效果并不好。在二分类任务中，SVM模型的精确率和召回率比较高。

随机森林模型构建多个决策树并综合分类结果，使随机森林模型在三个数据集中都有较不错的表现。

KNN模型在二分类任务，即个人信用风险评估任务中的分类效果较差，但在两个多分类任务，即玻璃类型识别，光学字符识别中的表现较好。