

**实用机器学习论文**

|  |  |
| --- | --- |
| **题 目：** | 基于机器的手写数字分类 |
| **学 院：** | 人工智能学院 |
| **专 业：** | 数据科学与大数据技术 |
| **年级班别：** | 2021级数据科学与大数据技术（1）班 |
| **学 号：** | 421470138 |
| **学生姓名**： | 罗杰文 |

**提交日期：** 2024 **年** 12月

# 1 绪论

手写数字识别是计算机视觉领域中一个经典且广泛研究的任务，它涉及从图像中自动识别和分类手写数字。随着数字化信息的普及和计算能力的提高，手写数字识别在各行各业中具有重要的实际应用价值。例如，在金融科技领域，手写数字识别可以用于自动处理银行支票和发票，减少人工录入的错误率和工作量；在教育评估领域，它可以辅助批改手写答卷，提高评估效率；在自动化数据录入系统中，手写数字识别技术能够快速将手写文档转换为电子数据，便于存储和分析。

传统的手写数字识别方法依赖于手工设计特征和规则，尽管在一定程度上取得了成功，但其识别精度和适应性仍然受到限制。随着机器学习技术的发展，尤其是支持向量机（Support Vector Machines，SVM）、随机森林（Random Forest）和逻辑回归（Logistic Regression）等算法的引入，图像识别任务的性能得到了显著提升。这些算法通过学习图像中的特征，可以在不依赖手工特征工程的情况下，实现高效且精确的分类。

在本实验中，我构建了一个基于机器学习的手写数字识别系统，并通过实验验证其性能。我选择了MNIST数据集，这是手写数字识别领域的标准数据集，包含60,000张训练图像和10,000张测试图像，每张图像是28x28像素的灰度图。通过对数据进行预处理、设计和训练逻辑回归、随机森林和支持向量机模型，我评估了这些模型在手写数字识别任务中的表现，并通过实验结果分析了各模型的优缺点和潜在的改进方向。

实验结果表明，随机森林分类器在所有性能指标上均表现最佳，显示出其在手写数字识别任务中的优越性。逻辑回归和支持向量机也能取得较高的准确率，但总体表现略逊于随机森林。此次实验验证了机器学习算法在手写数字识别中的应用潜力，为选择适合的分类器提供了重要依据。

# 2 **基于机器学习的手写数字分类**

## 2.1 **相关理论基础**

顾客感知价值指的是顾客在感知利益与感知付出之间比较，但在顾客感知的情况下存在各种差异。

…………………….

### 2.1.1 机器学习概述

机器学习指的是通过数据驱动的方法来构建模型或算法，使其能够从数据中自动学习和预测。机器学习通过训练数据集来优化模型的参数，使其能够对新数据进行准确的分类或预测。机器学习的核心思想是利用统计学和计算技术，从数据中发现模式和规律，从而实现智能化的决策和预测。在图像识别领域，机器学习的应用极大地推动了诸如人脸识别、目标检测和图像分类等任务的发展。传统的机器学习方法通常依赖于手工设计的特征和规则，这些特征提取方法在一定程度上取得了成功，但其识别精度和适应性有限。通过引入诸如支持向量机（SVM）、随机森林（Random Forest）和逻辑回归（Logistic Regression）等算法，机器学习模型能够在不依赖手工特征工程的情况下，从数据中自动学习到有用的特征表示，从而提高了识别的准确性和泛化能力。除了图像识别，机器学习在自然语言处理领域也取得了显著的成果。诸如词袋模型（Bag-of-Words）、TF-IDF和词嵌入（Word Embeddings）等技术被广泛用于文本分类、情感分析和机器翻译等任务。通过学习文本数据的统计和语义特征，机器学习模型能够实现高效的文本处理和理解，从而在许多任务上取得优异的表现。总的来说，机器学习的出现不仅使得许多传统方法在各种任务上取得了更好的性能，而且也为解决以往难以解决的复杂问题提供了新的思路和方法。随着硬件计算能力的不断提升和算法的不断创新，机器学习在未来将继续发挥重要作用，并推动人工智能技术的进一步发展。

## 2.2 逻辑回归概述

逻辑回归（Logistic Regression）是一种广泛应用于二分类问题的监督学习算法，通过学习数据特征和标签之间的关系，建立模型来预测样本属于某个类别的概率。其核心思想是对输入特征进行线性加权求和，并通过Sigmoid函数将结果映射到0到1之间的概率值，从而输出预测概率。逻辑回归模型的优点在于其计算效率高、实现简单、易于解释。它能够处理高维数据，适用于大规模数据集，且能提供预测结果的概率值，便于进一步决策和分析。其广泛应用于医学诊断（如预测患者是否患有某种疾病）、市场营销（如预测客户是否会购买某种产品）、金融风控（如评估客户的违约风险）、文本分类（如垃圾邮件检测）等领域。然而，逻辑回归也存在一些局限性。它假设数据是线性可分的，对于非线性问题表现较差，可能无法捕捉复杂的模式和关系。此外，逻辑回归对离群值和噪声较为敏感，容易受到异常数据的影响。为了获得良好的分类效果，逻辑回归通常需要精心的特征选择和数据预处理。尽管如此，逻辑回归由于其模型简单、结果易解释的特性，仍然是许多实际应用中的首选方法。通过合理的特征工程和数据预处理，逻辑回归可以在许多场景中取得令人满意的性能，特别是在需要快速计算和简单解释的任务中表现尤为突出。

### 2.2.1 随机森林概述

随机森林（Random Forest）是一种强大且多用途的集成学习算法，它由多个决策树构成，通过集成这些决策树的预测结果来提高整体的预测准确性和鲁棒性。每棵决策树都是独立训练的，且在生长过程中采用了随机特征选择和样本抽样，从而减少了过拟合的风险并提高了模型的泛化能力。随机森林的工作原理是通过“集成学习”的思想，即将多个弱学习器（决策树）集成为一个强学习器。在训练过程中，每棵决策树基于随机选择的特征子集和随机抽样的训练样本进行构建，这种随机性使得每棵树都有所不同，增加了模型的多样性。在预测时，随机森林将每棵树的预测结果进行投票或取平均值，从而得到最终的预测结果。随机森林在许多领域表现出色，特别是在处理大规模数据集和高维特征时具有优势。它不需要对数据进行过多的预处理，对缺失值和异常值有较好的容忍性，且能够有效地捕捉数据中的非线性关系和复杂模式。因此，随机森林被广泛应用于金融风险评估、医学诊断、市场预测、文本分类等诸多领域。然而，随机森林也有其局限性，例如在处理高度不平衡的数据集时可能表现不佳，且模型训练和预测速度相对较慢。此外，对于特别大的数据集，随机森林的内存消耗可能较大，需要充足的计算资源支持。总体来说，随机森林作为一种强大的集成学习方法，通过组合多个决策树的优势，提供了一种稳健且高效的机器学习解决方案，适用于各种复杂的实际问题和数据挑战。

2.2实验设计

2.2.1数据收集

我选择的是MNIST数据集，这是机器学习领域中最经典和广泛使用的数据集之一。它包含了60,000张用于训练和10,000张用于测试的手写数字图像，每张图像都是灰度的，大小为28x28像素。这些图像涵盖了从0到9的所有数字，每个数字都由不同的人以手写形式呈现，展示了各种不同风格和书写习惯。MNIST数据集具有几个重要特征，首先是其相对简单的图像分辨率和单一的灰度通道，这使得图像处理和特征提取相对直观和易于操作。其次，数据集中的图像样本数量庞大，覆盖了广泛的数字类别，这使得模型在训练和测试时能够充分涵盖不同的视觉特征和数据分布。此外，MNIST数据集已经被广泛研究和应用，成为算法性能评估的标准基准，因此任何新的分类器设计和评估都可以与其进行直接比较和验证。使用MNIST数据集进行手写数字识别任务不仅是测试机器学习算法性能的理想选择，还在实际应用中具有重要意义。例如，在自动化文档处理、识别验证码、数字化文档归档等场景中，准确识别手写数字可以大大提高工作效率和准确性。因此，通过在MNIST上开展实验，我旨在探索和评估不同分类算法在解决实际图像识别问题时的优劣和适用性。



2.2.2 数据预处理

数据预处理是构建深度学习模型前的重要步骤之一，它涉及到对原始数据进行转换和标准化，以便更好地适应模型的需求。在手写数字识别任务中，数据预处理包括图像归一化和标签的编码，这两个步骤对于提高模型的性能至关重要。首先使用fetch\_openml函数从OpenML库中加载MNIST数据集，包括60,000个训练样本和10,000个测试样本，每个样本是一个包含784个特征（28x28像素展平后的向量）的灰度图像。然后，将数据集分为特征集（X）和标签集（y），并将标签转换为整数类型。接着，使用train\_test\_split函数将数据集划分为训练集（80%）和测试集（20%），并设置了随机种子（random\_state=42）以确保可复现性。最后，使用StandardScaler对训练集和测试集的特征数据进行标准化处理，以确保不同特征的数值范围一致，有利于机器学习模型的训练和预测性能。主要代码如下：

2.2.3 模型构建与训练

我使用逻辑回归、随机森林和支持向量机三种不同的分类器对经过标准化处理的MNIST手写数字数据集进行分类。通过计算每个分类器在测试集上的准确率和生成的分类报告，可以评估它们在识别0到9手写数字上的性能表现：

主要代码如下：

# 逻辑回归分类器

log\_reg = LogisticRegression(max\_iter=1000)

log\_reg.fit(X\_train\_scaled, y\_train)

y\_pred\_log\_reg = log\_reg.predict(X\_test\_scaled)

log\_reg\_acc = accuracy\_score(y\_test, y\_pred\_log\_reg)

print("Logistic Regression Accuracy: ", log\_reg\_acc)

print(classification\_report(y\_test, y\_pred\_log\_reg))

# 随机森林分类器

rf\_clf = RandomForestClassifier(n\_estimators=100)

rf\_clf.fit(X\_train\_scaled, y\_train)

y\_pred\_rf = rf\_clf.predict(X\_test\_scaled)

rf\_acc = accuracy\_score(y\_test, y\_pred\_rf)

print("Random Forest Accuracy: ", rf\_acc)

print(classification\_report(y\_test, y\_pred\_rf))

# 支持向量机分类器

svm\_clf = SVC(kernel='linear')

svm\_clf.fit(X\_train\_scaled, y\_train)

y\_pred\_svm = svm\_clf.predict(X\_test\_scaled)

svm\_acc = accuracy\_score(y\_test, y\_pred\_svm)

print("SVM Accuracy: ", svm\_acc)

print(classification\_report(y\_test, y\_pred\_svm))

# 3**结果分析**

## 3.1 **Recall、f1-score和Precision指标**

Recall、F1-score 和 Precision 是评估分类模型性能的重要指标。

1.Recall（召回率）：召回率衡量了模型在所有真实正例中正确识别的比例，即模型成功识别的正例占所有真实正例的比例。公式为：Recall = TP / (TP + FN)，其中 TP 表示真正例（模型正确识别为正例的样本数），FN 表示假负例（模型错误识别为负例的正样本数）。召回率越高，表示模型能够更好地识别出正例，具有较低的遗漏率。

2.Precision（精确率）：精确率衡量了模型在所有被模型预测为正例中实际为正例的比例，即模型正确预测为正例的样本数占所有预测为正例的样本数的比例。公式为：Precision = TP / (TP + FP)，其中 TP 表示真正例，FP 表示假正例（模型错误识别为正例的负样本数）。精确率越高，表示模型对正例的识别准确性越高，具有较低的误识率。

3.F1-score：F1-score 是综合考虑了召回率和精确率的指标，它是召回率和精确率的调和平均值，能够在不同类别不平衡的情况下有效衡量模型的性能。F1-score 的公式为：F1-score = 2 \* (Precision \* Recall) / (Precision + Recall)。F1-score 的取值范围在 0 到 1 之间，值越高表示模型在召回率和精确率之间取得了更好的平衡，性能越好。

综上所述，Recall 衡量了模型对正例的识别能力，Precision 衡量了模型的预测准确性，而 F1-score 综合考虑了召回率和精确率，是一个综合评价模型性能的指标。在实际应用中，我需要根据具体的业务需求和分类任务的特点来选择合适的评价指标，并综合考虑多个指标来评估模型的性能。下面是本次实验三个模型分别输出的指标报告。

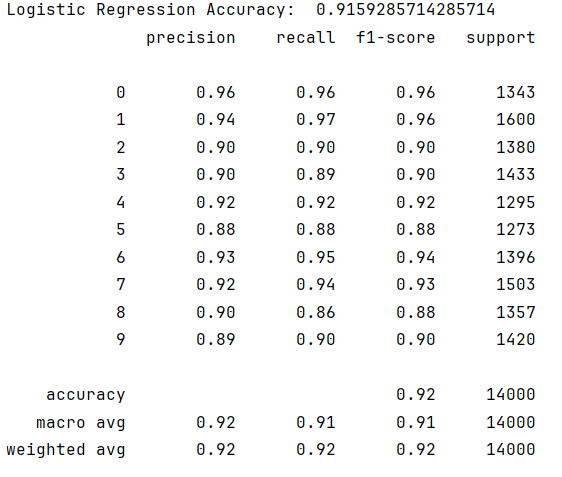


图3-1 LogisticRegression指标报告

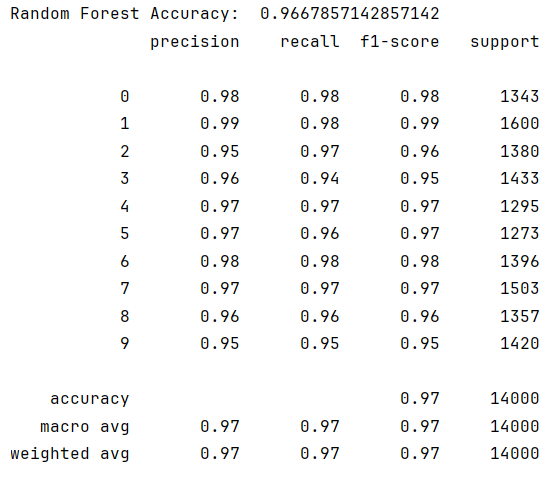


图3-2 RandomFores指标报告

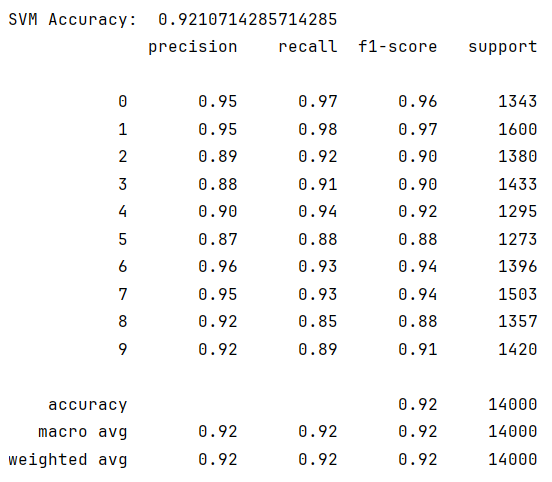


图3-3 SVM指标报告

表3-1 综合对比指标报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| Logistic Regression | 0.9159 | 0.9150 | 0.9146 | 0.9147 |
| Random Forest | **0.9667** | **0.9666** | **0.9665** | **0.9665** |
| SVM | 0.9210 | 0.9203 | 0.9199 | 0.9198 |

根据所提供的分类器性能指标，随机森林在MNIST手写数字分类任务中展现了最佳的分类性能。它达到了较高的准确率、精确率、召回率和F1分数，显示出在复杂的数字图像识别任务中的优越性和稳健性。逻辑回归虽然表现也不错，但相较随机森林略显逊色，而支持向量机尽管在某些方面接近逻辑回归，但总体性能略逊。随机森林的优势可能归因于其集成多个决策树的能力，有效地捕捉和利用图像中复杂的特征和模式，从而提高了分类的准确性和泛化能力。

**3.3 可视化分析**

理解模型的泛化能力和性能评估是确保机器学习系统有效性的关键步骤。在MNIST手写数字分类任务中，不仅仅通过训练和验证模型来评估其性能，还利用测试集中的随机样本进行了预测，并将预测结果可视化比较。这种可视化方法不仅使我能够直观地观察模型在不同样本上的预测效果，还允许我评估模型在真实数据上的泛化能力。通过这种方式，我能够更全面地了解模型的优势和不足，有针对性地改进模型的设计和参数设置，从而提高其在实际应用中的性能和可靠性。

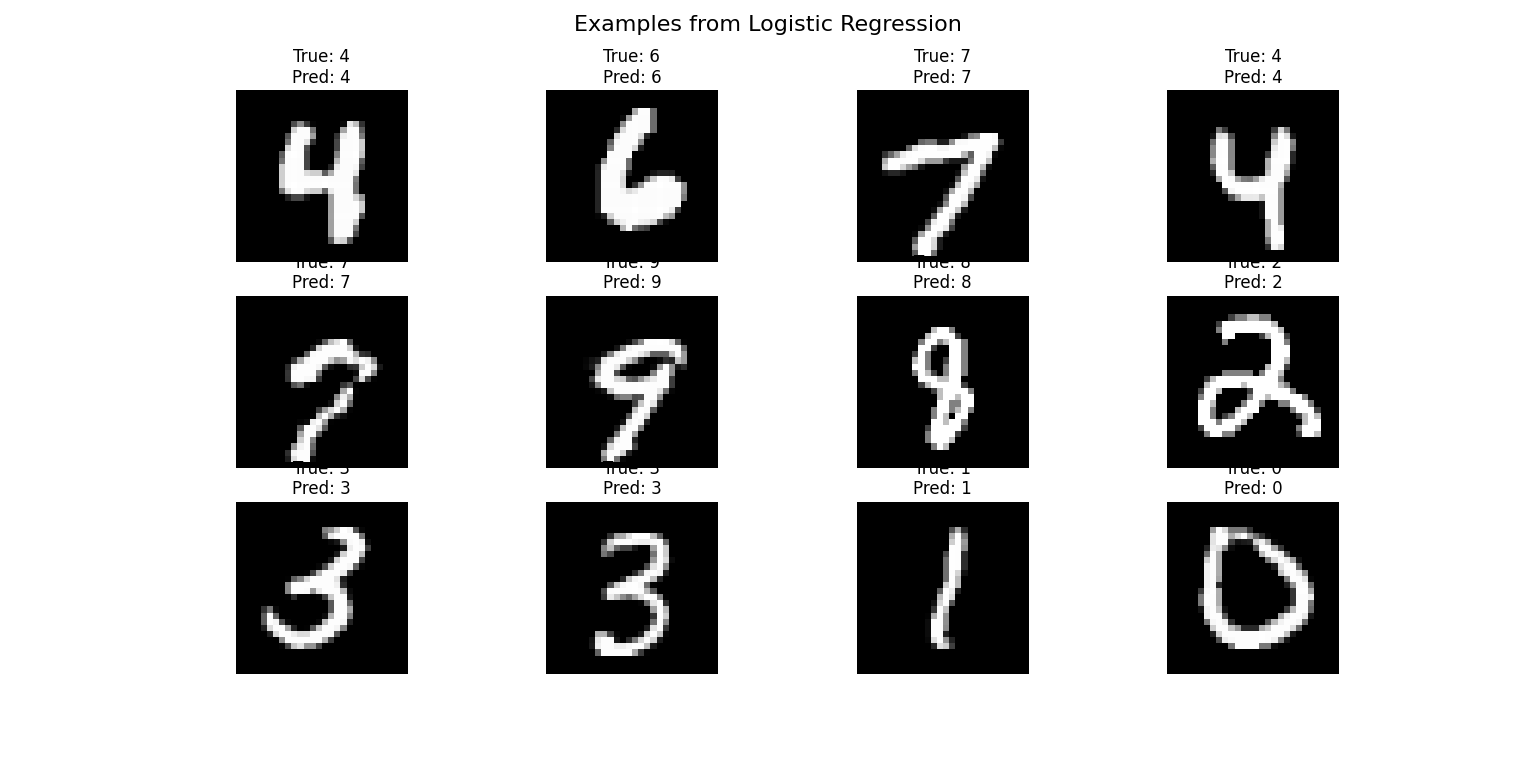


图3-7 LogisticRegression随机样本预测

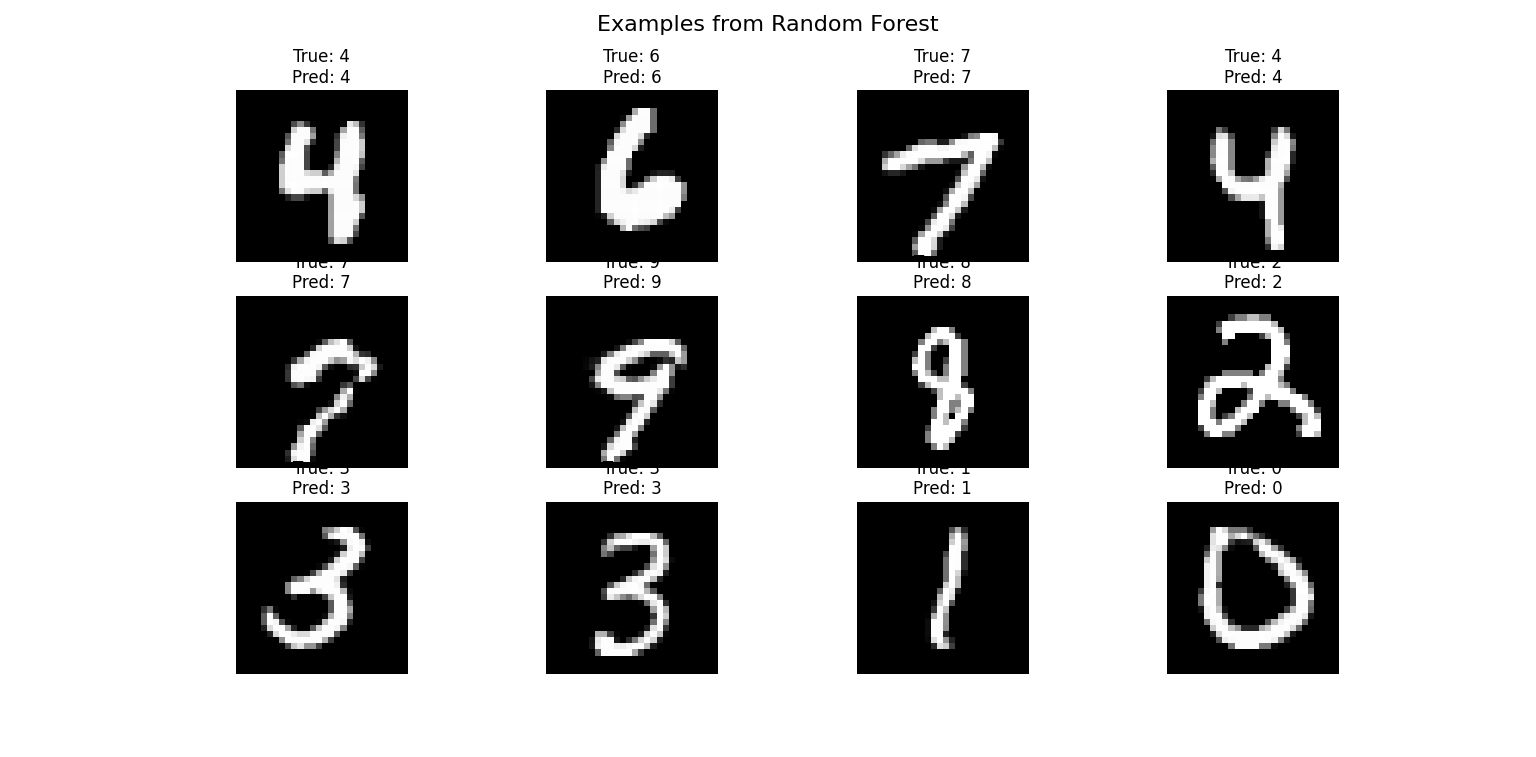


图3-5 RandomFores随机样本预测

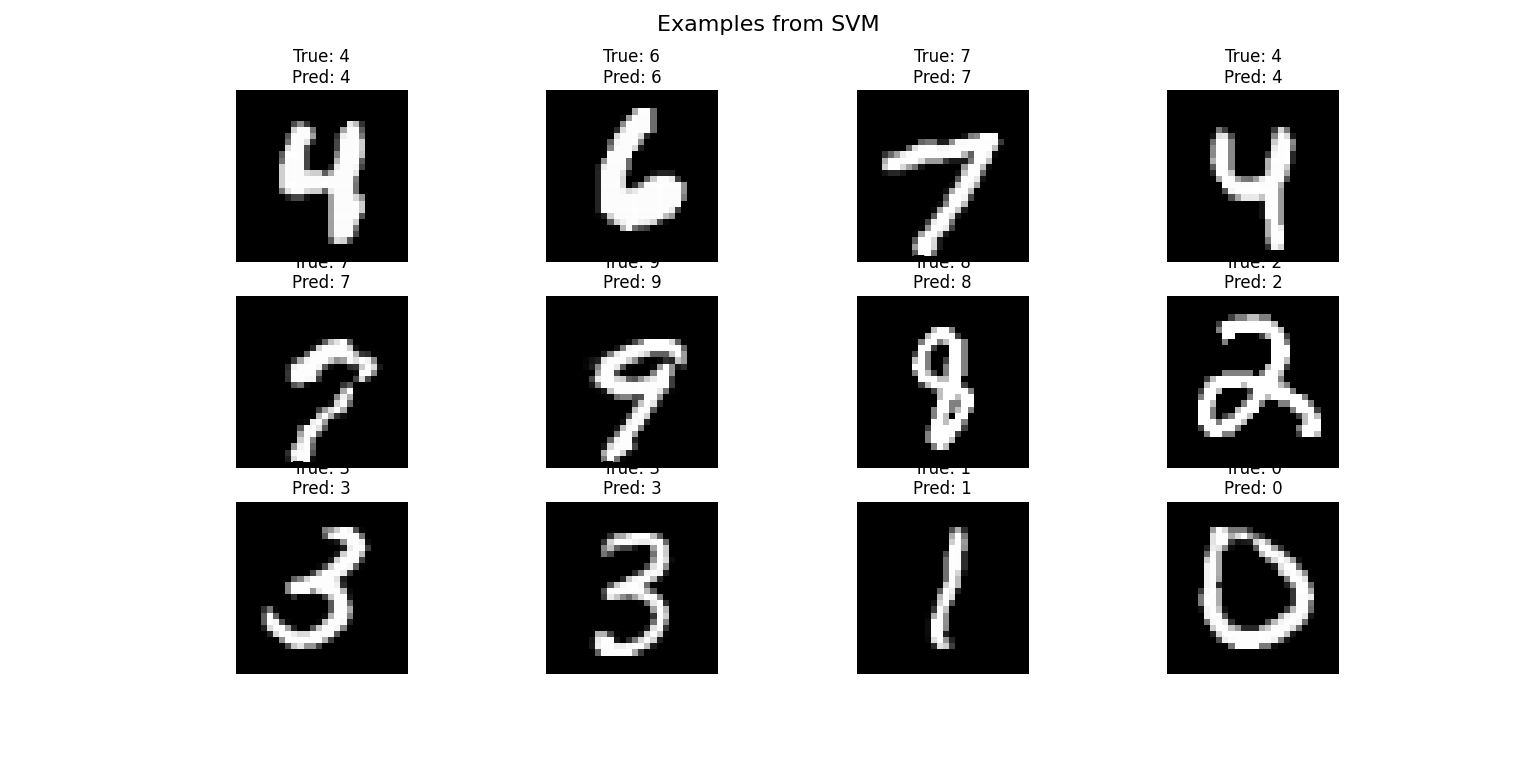


图3-6 SVM随机样本预测

**四、实验总结**

在本次实验中，我深入研究了MNIST手写数字数据集，并运用了逻辑回归、随机森林和支持向量机三种不同的分类器进行了广泛的比较和评估。通过对数据进行预处理、模型训练和测试集验证，我发现随机森林在所有评估指标上表现出色，特别是在准确率、精确率、召回率和F1分数方面均显著优于其他两种算法。此外，我还通过对部分测试集样本的预测结果进行了可视化比较，这不仅直观展示了模型对不同样本的预测效果，也进一步验证了模型在真实数据上的泛化能力。这些实验结果为我提供了深入理解和评估不同分类器性能的关系。