Comparação entre SVM e Regressão Logística no Dataset Iris

Objetivo

O objetivo deste código é comparar dois algoritmos de aprendizado de máquina — **Máquina de Vetores de Suporte (SVM)** e **Regressão Logística** — para classificar o conjunto de dados de flores Iris. Ao final do processo, são apresentadas as métricas de desempenho de ambos os modelos, com foco na **acurácia** de cada um.

Sobre o Dataset Iris

O **Iris dataset** é um conjunto de dados amplamente utilizado para teste de algoritmos de aprendizado de máquina. Ele contém 150 observações sobre flores Iris, distribuídas em três espécies (setosa, versicolor e virginica), com 4 características: comprimento e largura da sépala e da pétala.

Passos do Processo

1. **Carregamento dos Dados**: O código começa carregando o dataset Iris a partir de um arquivo CSV.
2. **Pré-processamento**: O código realiza a codificação da variável categórica (`Especie`) para valores numéricos, o que é necessário para que os algoritmos de aprendizado de máquina possam operar.
3. **Divisão de Dados**: Os dados são divididos em conjuntos de treino (80%) e teste (20%).
4. **Treinamento**: O modelo é treinado utilizando o algoritmo escolhido, sendo possível escolher entre **SVM** ou **Regressão Logística**.
5. **Avaliação**: Após o treinamento, o modelo é avaliado utilizando o conjunto de teste, e são calculadas as métricas de **acurácia**, **matriz de confusão** e o **relatório de classificação**. Além disso, é realizada uma **validação cruzada** para avaliar a consistência do modelo.

Modelos Utilizados

Máquina de Vetores de Suporte (SVM)

- O **SVM** é um modelo baseado em uma abordagem geométrica, que busca encontrar o **hiperplano ótimo** que separa as classes de dados com a maior margem possível.
- O SVM é particularmente eficaz quando há uma clara margem de separação entre as classes. Em problemas de classificação como o Iris dataset, o SVM costuma oferecer boa performance.
- SVM pode ser mais robusto em relação a conjuntos de dados com distribuições não lineares, especialmente quando se utiliza kernels não lineares (no nosso caso, o kernel linear foi usado).

Regressão Logística

- A **Regressão Logística** é um modelo estatístico que estima a probabilidade de um evento (classe) ocorrer. Ele trabalha melhor com dados lineares e pode não ter um desempenho tão bom em conjuntos de dados mais complexos ou com classes que não são linearmente separáveis.
- Apesar de ser simples e rápido de treinar, a regressão logística pode ser limitada em casos onde a separação entre as classes é não linear.

Resultados e Observações

Após rodar os dois modelos com o **Iris dataset**, os resultados obtidos são os seguintes:

SVM

- **Acurácia no Conjunto de Teste**: Aproximadamente 1.00 (100%)
- **Acurácia na Validação Cruzada**: Aproximadamente 0.98 (98%)

Regressão Logística

- **Acurácia no Conjunto de Teste**: Aproximadamente 0.95 (95%)
- **Acurácia na Validação Cruzada**: Aproximadamente 0.94 (94%)

Justificativa: Por que o SVM é melhor

- **Melhor Desempenho**: Como podemos ver, o **SVM** apresentou uma **acurácia superior**, tanto no conjunto de teste quanto na validação cruzada. Isso sugere que o SVM foi mais eficaz em encontrar a fronteira de decisão ótima para classificar as espécies de Iris.
- **Acurácia Consistente**: A validação cruzada, que avalia a performance do modelo em diferentes divisões dos dados, mostrou uma média de acurácia mais alta para o SVM, indicando que o modelo tem um desempenho mais consistente em diferentes subconjuntos dos dados.
- **Natureza dos Dados**: O conjunto de dados Iris tem algumas classes que podem ser separadas de forma linear (com uma boa margem), o que favorece o SVM, que tem uma boa capacidade de maximizar a margem entre as classes.

Conclusão

O **SVM** superou a **Regressão Logística** em termos de acurácia e consistência no caso do **Iris dataset**. Isso pode ser atribuído à sua capacidade de encontrar uma margem de separação ótima, especialmente em conjuntos de dados com diferentes distribuições. Embora a **Regressão Logística** também tenha apresentado bons resultados, ela não foi tão eficaz quanto o **SVM** para este problema específico.