Exercício Reconhecimento de Padrões - PCA e Aplicação

Lucas Ribeiro da Silva - 2022055564
Universidade Federal de Minas Gerais
Belo Horizonte - Minas Gerais - Brasil

lucasrsilvak@ufmg.br

1 Introdução

Neste relatório, implementaremos o método PCA para extrair características mais relevantes e reduzir a dimensionalidade dos dados, mantendo o máximo possível da variância da informação original, com o intuito de acelerar o treinamento dos modelos de classificação.

2 Visualização dos Dados

Os dados utilizados são provenientes da base de dados MNIST e consiste essencialmente de dígitos:

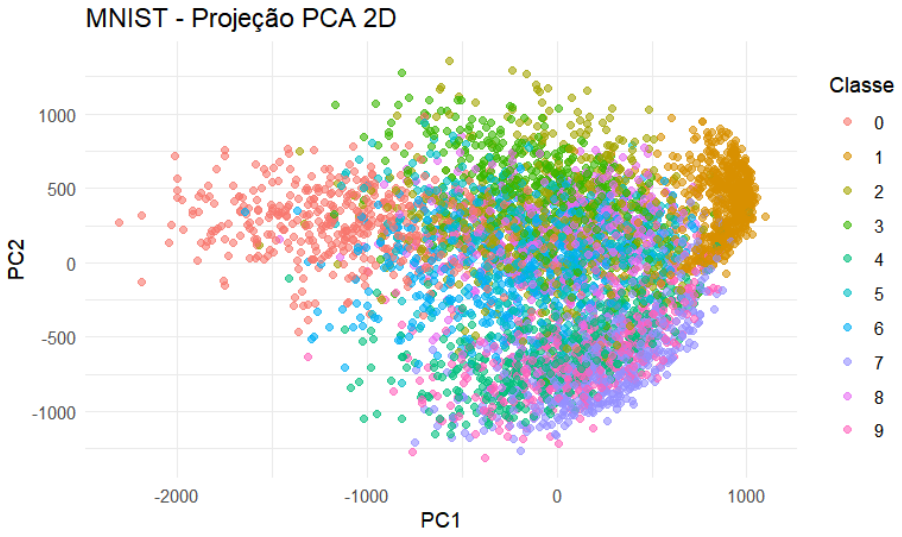

Figura 1: Dataset

3 Variância Explicada

A tabela abaixo explica a % de informação explicada pelas componentes do PCA. Percebe que somente 10 componentes já explicam praticamente metade da variância dos dados, enquanto 100 componentes concentram 90% da informação total.

Tabela 1: Porcentagem de Variância Explicada pelo PCA

Número de Componentes	**Variância Explicada (%)**
10	49,14
30	73,31
50	82,69
100	91,67

4 Classificador SVM

A tabela abaixo mostra a acurácia do classificador SVM após utilização da PCA. Percebe-se que com 10 componentes apenas já se obtém uma grande precisão na classificação. Esse valor atinge o máximo para 50 componentes no teste e posteriormente começa a cair, chegando a 12% para todas as componentes. Isto indica um mal da dimensionalidade e possivelmente overfitting.

Tabela 2: Acurácia de Classificação usando SVM após Redução por PCA

Número de Componentes	**Acurácia (%)**
10	88,72
30	94,93
50	94,99
100	94,19
784 (original)	12,28

5 Visualização das duas componentes principais

A imagem abaixo mostra a informação retida pelas duas componentes principais num gráfico 2D.

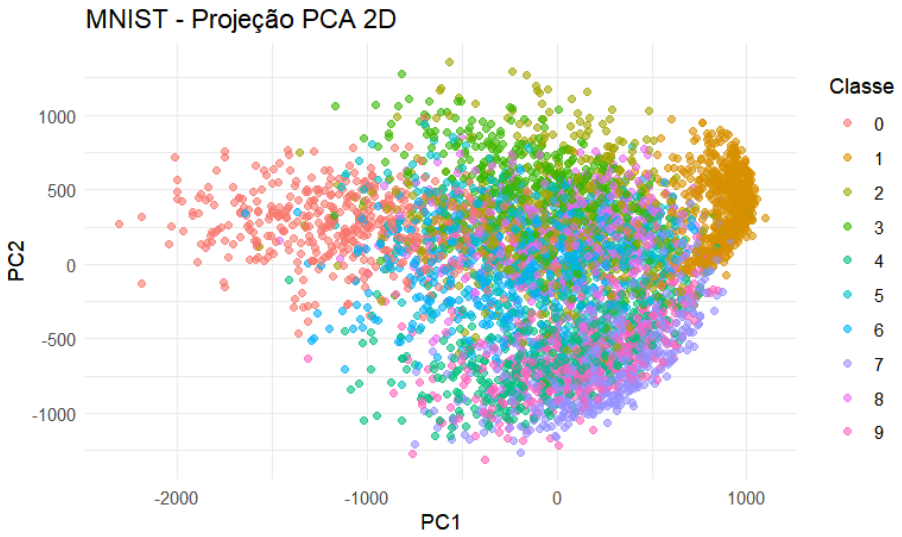

Figura 2: Enter Caption

Neste caso, é perceptível a inseparabilidade das classes, sendo necessário um número maior de dimensões para diferenciá-las. Entretanto pode se perceber a formação de alguns clusters.

6 Conclusões

A utilização do PCA, um método não-supervisionado obteve resultados excelentes em reduzir a dimensionalidade do problema, com o valor de 50 componentes sendo o ótimo aproximado do número de componentes, preservando a interpretabilidade e possivelmente melhorando a acurácia, devido ao contorno do mal da dimensionalidade.