Relatório III - Tópicos Especiais em Computação

Marcos Antonio e Vandirleya Barbosa

[1]Universidade Federal de Piauí (UFPI)

{marcos.brito, vandirleya.barbosa}@ufpi.edu.br

Resumo. *Este trabalho apresenta uma análise detalhada do conjunto de dados Iris, um dos conjuntos de dados mais amplamente utilizados no campo do aprendizado de máquina. Originário do UCI Machine Learning Repository, o conjunto de dados Iris é composto por 150 amostras de três espécies de íris (Iris setosa, Iris virginica e Iris versicolor), cada uma com quatro características: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. O objetivo principal desta análise foi conduzir uma análise descritiva do conjunto de dados Iris e identificar padrões ou tendências que possam indicar características importantes relacionadas à qualidade das espécies de íris. Para isso, foram realizadas várias análises estatísticas, incluindo a Análise de Componentes Principais (PCA), uma técnica de redução de dimensionalidade que permite visualizar a estrutura de alta dimensão dos dados em um espaço de dimensão menor. Através do PCA, foi possível identificar a quantidade de variância explicada por cada componente principal, o que ajudou a determinar o número adequado de componentes a serem considerados. Além disso, a visualização dos dois primeiros componentes principais permitiu uma clara separação entre as diferentes espécies de íris, destacando a eficácia do PCA como uma ferramenta de análise de dados. Este estudo também propôs um resumo exploratório dos resultados obtidos, com o objetivo de compreender melhor as relações entre as características das espécies de íris. Os resultados indicam que as medidas das pétalas e sépalas são características distintivas que podem ser usadas para diferenciar efetivamente entre as espécies de íris.*

Abstract. *This work presents a detailed analysis of the Iris dataset, one of the most widely used datasets in the field of machine learning. Originating from the UCI Machine Learning Repository, the Iris dataset consists of 150 samples of three species of iris (Iris setosa, Iris virginica, and Iris versicolor), each with four features: sepal length, sepal width, petal length, and petal width. The main objective of this analysis was to conduct a descriptive analysis of the Iris dataset and identify patterns or trends that may indicate important characteristics related to the quality of iris species. For this purpose, various statistical analyses were performed, including Principal Component Analysis (PCA), a dimensionality reduction technique that allows visualizing the high-dimensional structure of data in a lower-dimensional space. Through PCA, it was possible to identify the amount of variance explained by each principal component, which helped determine the appropriate number of components to consider. Furthermore, the visualization of the first two principal components allowed a clear separation between the different iris species, highlighting the effectiveness of PCA as a data analysis tool. This study also proposed an exploratory summary of the results*

obtained, aiming to better understand the relationships between the characteristics of iris species. The results indicate that the measurements of petals and sepals are distinctive characteristics that can be effectively used to differentiate between iris species.

1. Introdução

As flores desempenham um papel crucial em diversos aspectos da vida na Terra. Elas são mais do que apenas uma fonte de beleza estética; são fundamentais para a reprodução de uma grande variedade de plantas, contribuindo para a biodiversidade e a sustentabilidade dos ecossistemas. Além disso, as flores têm um papel significativo na agricultura, horticultura e na indústria farmacêutica, onde são usadas para a produção de alimentos, medicamentos e outros produtos de consumo. A classificação das flores é uma tarefa importante na botânica e na horticultura. Ao classificar as flores, os cientistas podem entender melhor suas características, comportamentos e relações evolutivas. Isso pode ajudar na conservação de espécies ameaçadas, no desenvolvimento de novas variedades de plantas e na melhoria das práticas agrícolas. Além disso, a classificação precisa das flores é essencial para garantir a segurança e a eficácia dos produtos farmacêuticos derivados de plantas.

Neste contexto, a análise de dados se torna uma ferramenta poderosa para a classificação das flores. Técnicas avançadas de aprendizado de máquina e estatística, como a Análise de Componentes Principais (PCA), permitem que os pesquisadores identifiquem padrões nos dados que podem não ser imediatamente aparentes. Esses padrões podem revelar relações importantes entre as diferentes características das flores, como o comprimento e a largura das sépalas e pétalas. Ao explorar essas relações, podemos obter uma compreensão mais profunda das diferenças e semelhanças entre as espécies de flores, o que pode ter implicações significativas para a botânica, a horticultura e outras áreas relacionadas. Portanto, a análise de dados não é apenas uma ferramenta para a classificação das flores, mas também uma janela para a compreensão da complexa beleza e diversidade do mundo natural.

Este trabalho propõe a aplicação da Análise de Componentes Principais (PCA) para a classificação do conjunto de dados Iris, que inclui três espécies de íris. A PCA é uma técnica estatística poderosa que pode reduzir a dimensionalidade dos dados, tornando mais fácil visualizar e interpretar as relações entre as características das flores. Ao aplicar a PCA ao conjunto de dados Iris, esperamos não apenas classificar com precisão as diferentes espécies de íris, mas também ganhar insights sobre as características que distinguem essas espécies. Esses insights podem ser úteis para botânicos, horticultores e outros profissionais que trabalham com flores, ajudando-os a tomar decisões informadas em seu trabalho.

2. Metodologia

Esta seção descreve a metodologia adotada para a elaboração deste relatório. Serão detalhados o uso do conjunto de dados e os passos seguidos para a análise e obtenção dos resultados.Isso inclui a seleção das variáveis de interesse, a aplicação de técnicas estatísticas específicas e a seleção dos gráficos para interpretação dos padrões identificados.

2.1. Códigos

Neste estudo, utilizamos a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade do conjunto de dados Iris, um conhecido conjunto de dados de aprendizado de máquina que contém 150 amostras de íris, divididas em três espécies: Setosa, Versicolor e Virginica. Cada amostra possui quatro características (comprimento e largura das sépalas e pétalas) que são usadas para prever a espécie da íris. A seguir estão detalhadas as etapas implementadas para carregar, pré-processar e analisar esses dados.

Primeiramente, carregamos o conjunto de dados Iris utilizando a função `load_iris` da biblioteca `sklearn.datasets`. Este passo é ilustrado na Figura 1. Em seguida, transformamos os dados em um DataFrame do Pandas para facilitar a manipulação e análise. Também adicionamos uma nova coluna para a variável alvo, que representa as espécies das flores.

```python
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import load_iris
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np

# Carregando o conjunto de dados Iris
iris = load_iris()
df = pd.DataFrame(data=iris.data, columns=iris.feature_names)
df['target'] = iris.target

#verificar os tipos de dados, valores nulos e estatísticas descritivas básicas
print(df.info())

#estatisitcas descritivas
print(df.describe())
```

Figura 1. Carregamento do conjunto de dados Iris.

Após o carregamento dos dados, verificamos os tipos de dados, a presença de valores nulos e calculamos estatísticas descritivas básicas, como média, desvio padrão, valores mínimos e máximos para cada característica. Esta etapa é crucial para entender a estrutura e a qualidade dos dados antes de proceder com o pré-processamento e análise. Para preparar os dados para a análise PCA, realizamos a normalização das características numéricas usando a biblioteca `StandardScaler` do `sklearn.preprocessing`. Esta etapa de normalização é essencial para que todas as variáveis contribuam igualmente para a análise, evitando que variáveis com diferentes escalas influenciem desproporcionalmente os componentes principais. A Figura 2 ilustra este processo.

Em seguida, implementamos o PCA para determinar a variância explicada por cada componente principal. O PCA é uma técnica estatística que transforma as características originais em um novo conjunto de variáveis não correlacionadas, chamadas componentes principais, que explicam a maior parte da variância presente nos dados originais. A Figura 3 mostra a implementação do PCA e a variância explicada por cada

```
# Pré-processamento
scaler = StandardScaler()
scaled_df = scaler.fit_transform(df.iloc[:, :-1])
```

Figura 2. Normalização dos dados utilizando StandardScaler.

componente principal. Um gráfico de cotovelo é utilizado para visualizar a variância explicada cumulativa, ajudando na escolha do número ideal de componentes principais.

```
# Implementação do PCA
pca = PCA()
pca.fit(scaled_df)
explained_variance = pca.explained_variance_ratio_


#grafico de cotovelo
plt.plot(np.cumsum(explained_variance))
plt.xlabel('Número de Componentes')
plt.ylabel('Variância Explicada Cumulativa')
plt.show()
```

Figura 3. Implementação do PCA e variância explicada.

Para a visualização dos resultados, reduzimos os dados para dois componentes principais usando PCA. Esta redução de dimensionalidade facilita a visualização dos dados em um gráfico bidimensional, onde podemos observar as relações e a separação entre as diferentes classes. Criamos um DataFrame contendo os componentes principais e a variável alvo. A Figura 4 mostra o gráfico de dispersão dos dois primeiros componentes principais, coloridos de acordo com a espécie. Esta visualização ajuda a entender a separabilidade das classes após a aplicação do PCA.

```
# Aplicando PCA com os dois primeiros componentes
pca = PCA(n_components=2)
principalComponents = pca.fit_transform(scaled_df)
principalDf = pd.DataFrame(data=principalComponents, columns=['PC1', 'PC2'])
finalDf = pd.concat([principalDf, df[['target']]], axis=1)

# Visualização dos Resultados
fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(1, 1, 1)
ax.set_xlabel('Componente Principal 1', fontsize=15)
ax.set_ylabel('Componente Principal 2', fontsize=15)
targets = [0, 1, 2]
colors = ['r', 'g', 'b']
iris_species = {0: 'Setosa', 1: 'Versicolor', 2: 'Virginica'}
for target, color in zip(targets, colors):
    indicesToKeep = finalDf['target'] == target
    ax.scatter(finalDf.loc[indicesToKeep, 'PC1'], finalDf.loc[indicesToKeep, 'PC2'], c=color, s=50)
ax.legend([iris_species[target] for target in targets])
ax.grid()
plt.show()
```

Figura 4. Visualização dos dados após aplicação do PCA.

3. Resultados

A Tabela 1 descreve os componentes da estrutura do conjunto de dados. Para cada coluna, a tabela indica que existem 150 valores não nulos, o que significa que o conjunto de dados não contém valores ausentes. Além disso, a tabela mostra o tipo de dados para cada coluna. As quatro primeiras colunas são do tipo float64, indicando que são números de ponto flutuante, enquanto a coluna Target é do tipo int32, indicando que é composta por números inteiros.

Tabela 1. Estrutura do DataFrame Iris.

Coluna	Contagem Não Nula	Tipo
Comprimento da sépala (cm)	150	float64
Largura da sépala (cm)	150	float64
Comprimento da pétala (cm)	150	float64
Largura da pétala (cm)	150	float64
Target	150	int32

A Tabela 2 fornece um resumo estatístico das quatro colunas numéricas do conjunto de dados. Para cada coluna, a tabela lista várias estatísticas descritivas, incluindo a contagem (o número de observações), a média, o desvio padrão (uma medida da dispersão dos dados), o mínimo, o primeiro quartil (25%), a mediana (50%), o terceiro quartil (75%) e o máximo. Essas estatísticas fornecem uma visão geral da distribuição dos dados em cada coluna.

Tabela 2. Estatísticas Descritivas do DataFrame Iris.

	Comprimento da sépala (cm)	Largura da sépala (cm)	Comprimento da pétala (cm)	Largura da pétala (cm)
Contagem	150.000000	150.000000	150.000000	150.000000
Média	5.843333	3.057333	3.758000	1.199333
Desvio padrão	0.828066	0.435866	1.765298	0.762238
Mínimo	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
Máximo	7.900000	4.400000	6.900000	2.500000

A Figura 5 apresenta o gráfico de linha que representa a variância explicada cumulativa em função do número de componentes. No eixo horizontal, temos o Número de Componentes"], que varia de 0 a 3. Isso representa o número de componentes principais considerados na análise. No eixo vertical, temos a Variância Explicada Cumulativa, que varia de 0,75 a 1,00. Isso representa a proporção total da variância nos dados que é explicada pelos componentes principais. A linha no gráfico mostra como a variância explicada cumulativa aumenta à medida que incluímos mais componentes na análise. Você pode ver que a linha sobe rapidamente no início e depois se nivela, o que indica que a adição de mais componentes após um certo ponto não contribui muito para explicar a variância adicional. Neste caso, parece que 2 ou 3 componentes são suficientes para capturar a maior parte da variância nos dados.

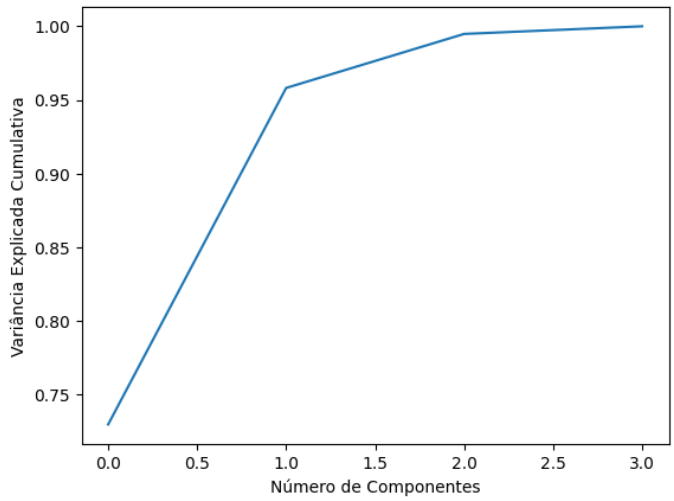

Figura 5. Variância acumulada dos componentes.

A Figura 6 apresenta o gráfico de dispersão que representa os dois primeiros componentes principais de um conjunto de dados. Cada ponto representa uma observação do conjunto de dados e a posição de cada ponto é determinada pelos valores dos dois primeiros componentes principais dessa observação. Os componentes principais são combinações lineares das variáveis originais que são calculadas de tal forma que capturam a maior parte da variância nos dados. As cores dos pontos representam diferentes categorias ou classes dentro do conjunto de dados. Podemos ver que as três espécies de íris formam três grupos distintos quando plotadas nos dois primeiros componentes principais. Isso sugere que essas espécies têm diferenças significativas em termos das variáveis medidas (neste caso, provavelmente o comprimento e a largura das sépalas e pétalas), e que essas diferenças são capturadas pelos componentes principais.

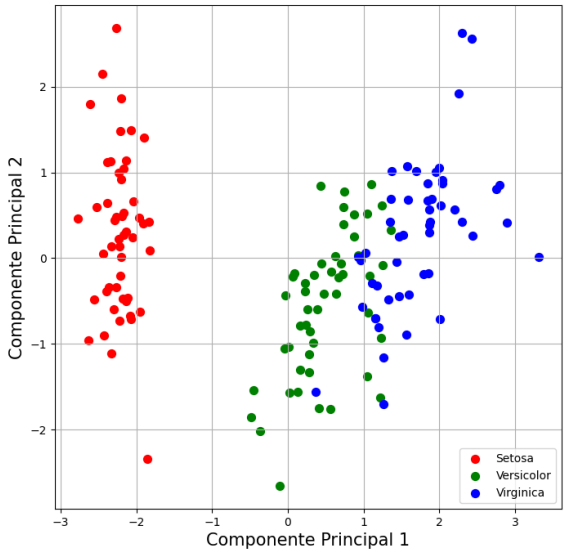

Figura 6. Dispersão de dados dos componentes principais após o PCA.

4. Conclusão

Neste estudo, aplicamos a Análise de Componentes Principais (PCA) ao conjunto de dados Iris para reduzir a dimensionalidade dos dados e identificar as principais características que distinguem as três espécies de íris: Setosa, Versicolor e Virginica. A PCA provou ser uma técnica eficaz, permitindo a visualização das relações entre as características das flores em um espaço bidimensional. A análise descritiva inicial mostrou que o conjunto de dados é bem equilibrado e não possui valores ausentes, facilitando as etapas subsequentes de normalização e aplicação do PCA. Os dois primeiros componentes principais explicaram a maior parte da variância, e a visualização resultante destacou uma clara separação entre as espécies, indicando que as características medidas são efetivas para a distinção entre elas. Os resultados obtidos confirmam a utilidade do conjunto de dados Iris para estudos de classificação e demonstram como a PCA pode simplificar e interpretar conjuntos de dados complexos. Esta técnica revelou padrões e relações importantes que podem ser úteis para botânicos, horticultores e cientistas de dados. Futuramente, a análise pode ser expandida para incluir outras técnicas de aprendizado de máquina, visando melhorar a precisão da classificação das espécies de íris. Este estudo reforça a importância das técnicas de redução de dimensionalidade na análise de dados, sugerindo que métodos como a PCA são valiosos para extrair informações significativas em diversas áreas biológicas.

Referências