Relatório - 2ª Atividade de Aprendizado de Máquina Redução de Dimensionalidade

Erick Santos do Nascimento - 1440434

Curso de Ciência da Computação - UECE

erick.nascimento@aluno.uece.br

Resumo

Relatório da atividade sobre redução de dimensionalidade da disciplina de Aprendizado de Máquina (2022.1). O objetivo é explorar as técnicas de redução de dimensionalidade apresentadas em aula.

1 Introdução

Utilizando o dataset *Water Quality Dataset* e esta imagem, as seguintes atividades foram feitas:

- Seleção de Atributos usando Chi-square
- Redução de dimensionalidade usando PCA
    - Calcular matriz de covariância
    - Calcular os Autovalores e Autovetores da matriz de covariância
    - Calcular a Variância explicada de acordo com a quantidade de autovetores
    - Aplicar a matriz de autovetores à matriz dos dados originais
    - Exibir a matriz de dados projetada no novo espaço gerado pelas componentes principais
- Aplicação de imagem em PCA
    - Calcular matriz de covariância
    - Calcular os Autovalores e Autovetores da matriz de covariância
    - Calcular a Variância explicada de acordo com a quantidade

de autovetores
    - Aplicar a matriz de autovetores à imagem original
    - Exibir a imagem original e algumas versões com diferentes quantidades de autovetores

Para saber mais sobre o dataset:

https://www.kaggle.com/datasets/adityakadiwal/water-potability

2 Redução de Dimensionalidade

No aprendizado de máquina, para capturar indicadores úteis e obter um resultado mais preciso, tendemos a adicionar o máximo de features possível logo de início. Porém, a partir de um certo ponto, o desempenho do modelo começa a diminuir com o aumento do número de features. Este fenômeno é frequentemente referido como “The Curse of Dimensionality” (A Maldição da Dimensionalidade).

A maldição da dimensionalidade ocorre porque a densidade do conjunto de dados diminui exponencialmente com o aumento da dimensionalidade. Quando continuamos adicionando features sem aumentar o número de amostras do dataset de treinamento, a dimensionalidade do espaço de features aumenta e se torna cada vez mais esparsa. Devido a essa dispersão, torna-se muito mais fácil encontrar uma solução “perfeita” para o modelo de aprendizado de máquina, o que muito provavelmente leva ao problema do overfitting.

Técnicas de redução de dimensionalidade, tais como a seleção e a extração de atributos, são usadas para diminuir a dimensão desses dados, removendo atributos irrelevantes ou informação redundante e que podem atrapalhar o processo de aprendizagem, esse processo aumenta a densidade das amostras, facilita a visualização e interpretação e reduz o esforço computacional do treinamento.

A finalidade aqui é aplicar técnicas de redução de dimensionalidade no dataset *Water Quality Dataset* e nesta imagem.

3 Seleção de Atributos

A redução de dimensionalidade é o processo de reduzir a dimensionalidade do espaço de features considerando a obtenção de um conjunto de features principais. A redução da dimensionalidade pode ser dividida em **feature selection** e **feature extraction**.

Feature selection é o processo de seleção de um subconjunto de features relevantes para uso na construção do modelo.

Há muitas maneiras de realizar esse processo, mas a maioria dos métodos pode ser dividida em três grupos principais

- Baseados em filtros: Nós especificamos alguma métrica e baseado nela filtramos features. Alguns exemplos dessa métrica são o teste de correlação e o **chi-square (chi-quadrado)**
- Wrapper-based: esses métodos consideram a seleção de um conjunto de features como um problema de busca. Um exemplo é o RFE, Recursive Feature Elimination (Eliminação Recursiva de Variáveis)
- Embutidos: métodos embutidos usam algoritmos que já têm naturalmente um processo de feature selection. Um exemplo é o decision tree.

Uma das técnicas de feature selection baseadas em filtros é conhecida como filtro chi-square (chi-quadrado).

Chi Quadrado mede a relação de dependência entre duas variáveis, verificando como os valores esperados desviam dos valores observados.

Quando temos um alto valor de Chi-quadrado (nosso p-value será baixo), significa que temos evidência estatística para inferir que os valores observados e esperados não são os mesmos, portanto possuem dependência entre si. Quanto mais alto o Chi-quadrado, maior a dependência entre as variáveis.

$$X^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$

Freqüência observada = Nº de observações da classe

Frequência esperada = Nº de observações esperadas da classe se não houver relação entre o recurso e o alvo.

Pares de colunas com dependência superior a um limite são reduzidos a apenas um. É importante observar que esse processo é sensível à escala das variáveis, portanto, é recomendado a normalização das colunas.

4 PCA (Principal Component Analysis)

Feature extraction visa reduzir o número de features em um conjunto de dados criando novos recursos a partir dos existentes (e descartando as features originais). Esse novo conjunto reduzido de recursos deve ser capaz de resumir a maioria das informações contidas no conjunto original de recursos. Dessa forma, uma versão resumida dos recursos originais pode ser criada a partir de uma combinação do conjunto original.

Uma das técnicas mais conhecidas de feature extraction é o PCA. Essa realiza um mapeamento linear dos dados para um espaço de menor dimensão de forma que a variância dos dados na representação de baixa dimensão seja maximizada. Essa representação é feita buscando novos eixos onde a projeção dos dados nos novos eixos, maximiza a variância.

Os novos eixos reúnem a variância dos dados de forma decrescente: variância no 1° eixo > variância no 2° eixo > variância no 3°

eixo …

A variância dos dados é acumulada nos eixos de forma decrescente e a variância calculada das projeções no eixo dividido pela variância total diz a contribuição do eixo para a variância total.

5 Metodologia

Descreva o procedimento utilizado para a aplicação da seleção de atributos.

1. O primeiro passo é encontrar quais valores esperaríamos ver em cada casa assumindo independência entre as features.
2. Para cada uma das features se calcula a estatística de chi-quadrado(se a variável explicativa influencia a variável resposta, então a diferença entre o número de observações real e o número de observações esperado deve ser alta, o que aumenta o valor da estatística).
3. Para escolher as features podemos fixar o número de features que queremos no final ou o valor mínimo calculado para dizer se a feature será selecionada ou não.

Descreva o procedimento utilizado para a aplicação da técnica PCA.

1. Normalização e dos dados (média zero e variância 1)
2. Cálculo da matriz de covariância dos dados
3. Cálculo dos autovalores e autovetores da matriz
4. Ordenação dos autovetores de acordo com a contribuição na variância dos dados
5. A aplicação da matriz nos dados originais
6. *Para imagem o retorno às dimensões originais para exibição**

6 Resultados

1. Seleção de atributos

    • Apresente os atributos selecionados pelo método filtro

['Solids' 'Sulfate' 'Conductivity' 'Trihalomethanes']

- Relacione/compare os atributos selecionados nessa atividade com os atributos selecionados na Atividade 1

Atributos selecionados na atividade 1 = ph, Hardness, Solids, Chloramines

Atributos selecionados na atividade 2 usando o método chi_quadrado = Solids, Sulfate, Conductivity,Trihalomethanes

Solids foi o único que se manteve

2. PCA (para os experimentos realizados tanto com a base de dados, quanto com a imagem escolhida)

1- Base de dados

- Apresente os valores encontrados (p.ex: autovalores, autovetores e variância explicada) nos experimentos realizados.

Autovalores

[0.76548182,1.20697306,1.17047771,0.87351307,1.04591793,0.95149163,0.9705474,1.01055863,1.00503875]

Autovetores (ordenados)

[-0.47404071 -0.21606914 0.66225649 -0.06779952 -0.45863379 -0.01664511 -0.16382524 -0.04620069 0.21406511]

[-0.38863879 -0.62587007 -0.16821895 0.2597866 0.54146414 -0.16084655 -0.08868279 -0.04434396 0.18048214]

[-0.02855533 -0.17244337 0.11815196 -0.59443501 0.2424053 0.37054766 0.51549467 -0.36757885 0.07225053]

[0.06350484 0.28462248 -0.15566146 0.17142755
  -0.03075806 0.03163354 -0.25103235 -0.69650358
  0.55790507]

[-0.0458843   0.07133336 0.12673165 -0.1188623   0.04773887
  -0.7479588 0.10830838 -0.45206732 -0.42988381]

[0.02559969 -0.07436301 -0.05769798 -0.50058809
  0.15568366 0.15423477 -0.78419574 -0.07045015
  -0.26822632]

[-0.26977197 0.02846038 0.0030525   0.45328296
  -0.05988011 0.49747192 0.01743162 -0.35339691
  -0.58709553]

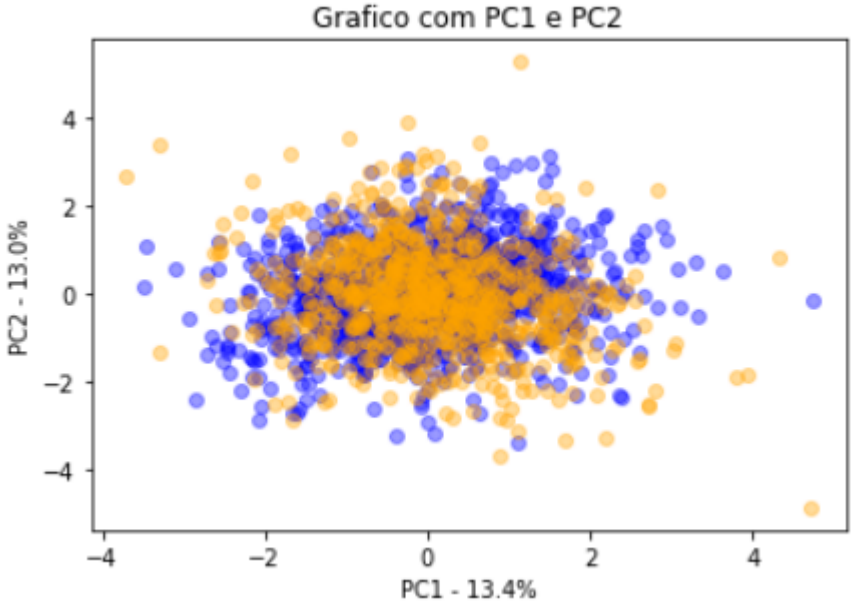
[0.73741317 -0.47688068 0.37314318 0.2331144
  -0.01397518 0.04687941 -0.07095278 -0.16076458
  -0.04422602]

[0.00901164 -0.46044611 -0.57990796 -0.14369872
  -0.63754835 -0.05136489 0.06717762 -0.12491976
  -0.04173214]

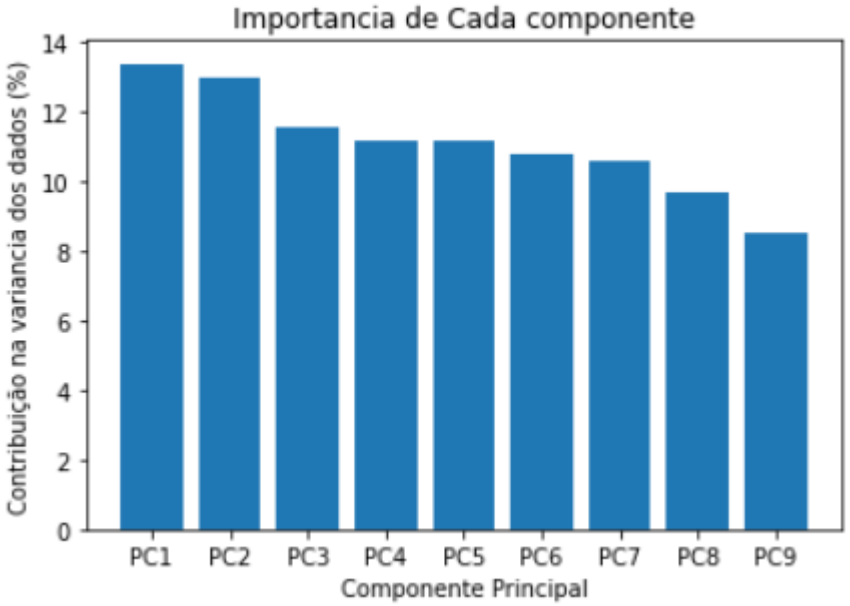
Dados plotados com PC-1 e PC-2

Em azul a classe 0

Em azul a classe 1



• Apresenta o gráfico da variância explicada vs quantidade de autovalores para cada experimento.



• Analise os resultados apresentados.

Com as matrizes de covariancia a correlação fica clara a baixa relação entre as variáveis, o que pode ser visto também na concentração de informação presente nas PCs que está distribuída de forma equilibrada em todas as PCs e não concentrada nas primeiras, com isso acho que a redução de dimensionalidade teria poucos benefícios neste caso, mais para uma melhor conclusão seriam necessários testes para ter mais segurança nesta conclusão

2 - Imagem

• Apresente os valores encontrados (p.ex: autovalores, autovetores e variância explicada) nos experimentos realizados.

Autovalores

[1173529.33000878 509873.73686527 280691.35683635
200294.04346139 135088.4876888   126507.50591156
91523.4856776   63974.53853976   54243.06155901
52371.08570209   45911.50702772   37112.70486291
35653.71739898   33762.20036408   28783.9828935
26669.63015107   24021.00855646   22872.86741848
20681.2557403   18208.39056033]

Autovetores

Não adicionei por ser uma matriz muito grande

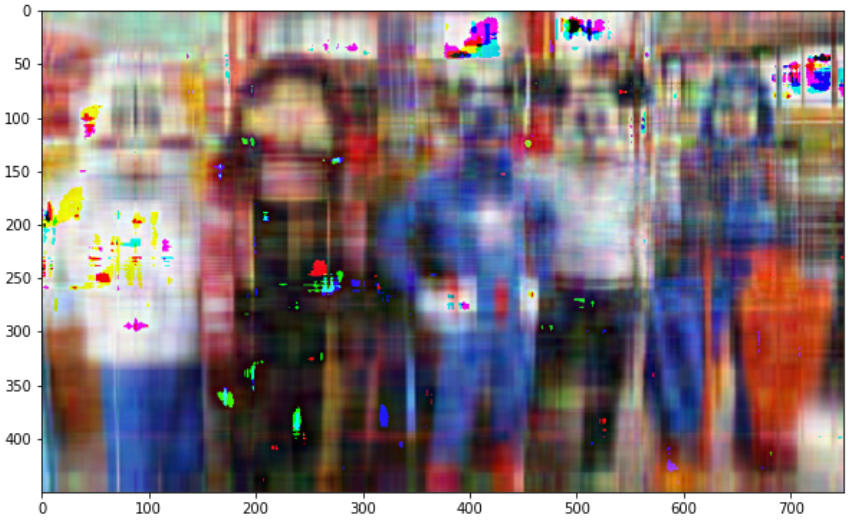
Imagem original

Com 200 componentes

Com 100 componentes

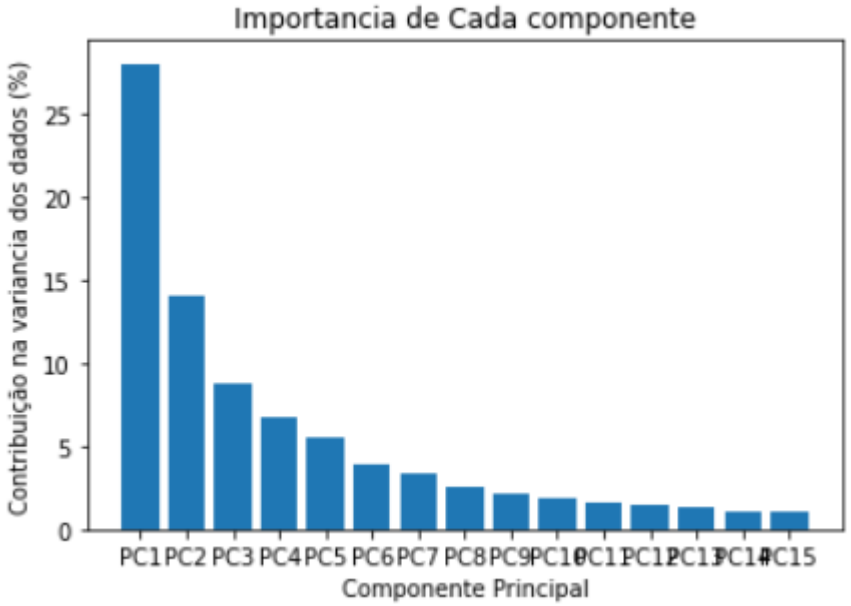
Com 50 componentes

Com 25 componentes



• Apresente o gráfico da variância explicada vs quantidade de autovalores

Gráfico das primeiras 15 componentes



Vendo o gráfico pode-se ver que boa parte da informação está concentrada nas primeiras componentes, mas mesmo assim, mesmo com um grande número de componentes pode-se ver claramente ruído na imagem logo para compressão para posterior uso da imagem seria necessário um tratamento para mascarar o ruído. Outra aplicação seria para aplicação em reconhecimento/classificação de imagem, neste caso seriam necessários testes para verificar a validade dessa premissa, mas como boa parte da informação está concentrada nas primeiras componentes pode gerar bons resultados.

Referências

[1]Fodor, I. (2002). A survey of dimension reduction techniques (Technical report). UCRL-ID-148494.

[2]Documentação do sklearn , dispnivel em: https://scikit-learn.org/stable/modules/classes.html

[3]https://analyticsindiamag.com/a-beginners-guide-to-chi-square-test-in-python-from-scratch/

[4]https://www.statisticshowto.com/probability-and-statistics/chi-square/#chisquareqtest

[5] PCA in studying coordination and variability: a tutorial, Andreas Daffertshofer a,*, Claudine J.C. Lamoth a,b, Onno G. Meijer a, Peter J. Beek a · (2004) - https://www.researchgate.net/publication/8594933_PCA_in_Studying_Coordination_and_Variability

[6] Principal component analysis Rasmus Bro a and Age K. Smilde ab a Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958, Frederiksberg C, Denmark b Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands - https://pubs.rsc.org/en/content/articlehtml/2014/ay/c3ay41907j

*Além das aulas ministradas durante a disciplina