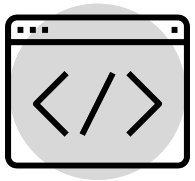
escola britânica de artes criativas & tecnologia

Profissão Cientista de Dados

escola britânica de artes criativas & tecnologia

BOAS PRÁTICAS

PCA



- Explore a redução de dimensionalidade
- Descubra usos
- Aprenda como fazer
- Analise a matéria-prima
- Entenda a Geometria do PCA
- Capture a Informação
- Examine a explicação da variância
- Determine quantos componentes utilizar

Explore a redução de dimensionalidade

- Esteja ciente da "Maldição da Dimensionalidade": Este fenômeno ocorre quando o número de variáveis aumenta, tornando o espaço vetorial mais esparso e a extrapolação mais exagerada. Isso pode levar a um aumento da chance de overfitting e a uma queda na precisão do modelo após um certo ponto.

- Considere duas abordagens para resolver a maldição da dimensionalidade: seleção de variáveis e redução de dimensionalidade. A seleção de variáveis envolve escolher apenas as melhores variáveis para trabalhar, enquanto a redução de dimensionalidade envolve trabalhar com combinações de todas as variáveis.

Explore a redução de dimensionalidade

- Ao reduzir a dimensionalidade, tente perder o mínimo possível de informação: A redução de dimensionalidade visa reduzir substancialmente o número de variáveis ou dimensões e perder o mínimo possível de informação, aproveitando a correlação entre as variáveis.

Descubra usos

- Ao lidar com um grande número de variáveis, considere usar técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA). Isso pode ajudar a tornar o modelo mais eficiente e fácil de interpretar.
- Ao usar PCA, é importante comparar o desempenho do modelo com e sem a redução de dimensionalidade. Isso pode ajudar a determinar se a redução de dimensionalidade está realmente melhorando o desempenho do modelo.

Aprenda como fazer

- Ao usar a Análise de Componentes Principais (PCA), lembre-se de que ela é uma técnica de redução de dimensionalidade que transforma o conjunto de dados original em um novo conjunto de variáveis, chamadas componentes principais. Essas são combinações lineares das variáveis originais e são ordenadas de acordo com a quantidade de informação que agregam ao conjunto de dados.

- Ao implementar o PCA em Python, você pode usar a função PCA do pacote sklearn. Lembre-se de que, ao contrário de outros algoritmos de aprendizado de máquina, o PCA não tem uma variável de resposta; todas as variáveis são tratadas igualmente e o algoritmo se aproveita da correlação entre elas para criar as componentes principais.

Analise a matéria-prima

- A análise de componentes principais trabalha com a matriz de covariância ou com a matriz de correlação. Certifique-se de entender como essas matrizes são construídas e como elas podem ser usadas na análise de dados.
- Lembre-se que a correlação e a covariância são a matéria-prima da análise de componentes principais. Portanto, é fundamental entender esses conceitos para realizar uma boa análise de componentes principais.

Entenda a Geometria do PCA

- Evite redundância de informação: Se houver alta correlação entre duas variáveis, pode-se considerar a remoção de uma delas, pois ela pode não estar fornecendo informações novas ou úteis.

- Utilize transformações lineares para obter novas perspectivas: Transformações lineares podem ajudar a visualizar a média e a evolução do desempenho de um conjunto de dados de uma nova maneira.

Entenda a Geometria do PCA

- Preserve a informação original: Ao manipular dados, é importante tentar preservar a informação original tanto quanto possível. Isso pode ser feito mantendo a escala de variabilidade dos dados.

- Utilize a rotação de dados para encontrar eixos não correlacionados: A rotação de dados pode ajudar a encontrar eixos que não estão correlacionados, o que significa que eles não compartilham informação. Isso pode ser útil para reduzir a dimensão dos dados.

Entenda a Geometria do PCA

- Use auto vetores e auto valores para encontrar os eixos de maior variabilidade: Esses eixos podem ser usados para criar novas variáveis, chamadas componentes principais, que são essencialmente uma rotação dos dados originais.

- Aplique a Análise de Componentes Principais (PCA) para reduzir a dimensão dos dados: A PCA pode ser usada para reduzir a dimensão dos dados, preservando a maior quantidade possível de informação. Isso pode ser especialmente útil quando se trabalha com conjuntos de dados de alta dimensão.

Capture a Informação

- Ao trabalhar com a Análise de Componentes Principais (PCA), é importante entender que os componentes principais são ordenados pela explicação da variabilidade. O primeiro componente principal é o que explica a maior parte da variabilidade.

- A variância explicada é um conceito importante na PCA. Ela é a variância da componente principal dividida pela variância total. É útil visualizar isso de forma automática para entender melhor a distribuição da variância.

Capture a Informação

- A visualização dos autores e auto vetores pode ajudar a interpretar a variância explicada.
- A visualização da matriz de correlação pode ser uma ferramenta útil para entender as relações entre as variáveis.
- Ao aplicar a PCA em uma base de dados geral, lembre-se de que o número de componentes será igual ao número de variáveis.

Capture a Informação

- Ao analisar a correlação entre as variáveis, é importante observar as diferenças entre as correlações intra e inter variáveis. Por exemplo, as provas de uma mesma matéria podem ter maior correlação entre si do que com provas de outras matérias.

- A escolha do número de componentes para a redução de dimensão é uma decisão importante e deve ser feita com base na explicação da variabilidade dos dados. Não é sempre que precisamos usar todas as componentes principais, às vezes, um subconjunto delas já é suficiente para explicar a maior parte da variabilidade dos dados.

Examine a explicação da variância

- Sempre considere a decomposição em auto vetores e autores da matriz de variância. Lembre-se de que os autovetores apontam na direção de maior variabilidade dos dados e as componentes principais são as projeções dos pontos dos dados originais nesses auto vetores.

- Utilize as propriedades das componentes principais a seu favor. Elas trazem toda a informação original dos dados, têm correlação zero entre si e podem ser ordenadas de acordo com a explicação da variância.

Determine Quantos Componentes Utilizar

- Transforme os dados em um formato que seja mais fácil de analisar: No exemplo da aula, as imagens foram transformadas em vetores para facilitar a análise.
- Lembre-se que a redução de dimensionalidade pode levar a uma ligeira diminuição na precisão do modelo, mas pode tornar o modelo mais eficiente e viável.

escola britânica de artes criativas & tecnologia

Bons estudos!